Metrology and Fundamental Constants, Course CLXVI

T. W. HANSCH
S. LESCHIUTTA
A. J. WALLARD,
Editors

IOS Press

This page intentionally left blank

SOCIETÀ ITALIANA DI FISICA

RENDICONTI

DELLA

SCUOLA INTERNAZIONALE DI FISICA

“ENRICO FERMI”

CLXVI Corso

a cura di T. W. Hänsch, S. Leschiutta and A. J. Wallard

Direttori del Corso

e di

M. L. Rastello

VARENNA SUL LAGO DI COMO

VILLA MONASTERO

18 – 28 Luglio 2006

Metrologia e costanti fondamentali

2007

SOCIETÀ ITALIANA DI FISICA
BOLOGNA-ITALY

ITALIAN PHYSICAL SOCIETY

PROCEEDINGS

OF THE

INTERNATIONAL SCHOOL OF PHYSICS

"ENRICO FERMI"

Course CLXVI

edited by T. W. Hänsch, S. Leschiutta and A. J. Wallard

Directors of the Course

and

M. L. Rastello

VARENNA ON LAKE COMO

VILLA MONASTERO

18 – 28 July 2006

Metrology and Fundamental Constants

2007

IOS
Press

AMSTERDAM, OXFORD, TOKIO, WASHINGTON DC

ISBN 978-1-58603-784-0 (IOS)
ISBN 978-88-7438-037-4 (SIF)
Library of Congress Control Number: 2007934139

Production Manager
A. Oleandri

Copy Editor
M. Missiroli

jointly published and distributed by:

IOS PRESS
Nieuwe Hemweg 6B
1013 BG Amsterdam
The Netherlands
fax: +31 20 620 34 19
order@iopress.nl

SOCIETÀ ITALIANA DI FISICA
Via Saragozza 12
40123 Bologna
Italy
fax: +39 051 581340
order@sif.it

Distributor in the UK and Ireland
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
sales@gazellebooks.co.uk

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
iosbooks@iospress.com

Proprietà Letteraria Riservata
Printed in Italy

Supported by UNESCO-ROSTE, Contract No.: 875.767.5

Supported by Consiglio Nazionale delle Ricerche (CNR)

Supported by Istituto Nazionale di Fisica Nucleare (INFN)

Supported by INFN, Sezione di Pisa

This page intentionally left blank

INDICE

This page intentionally left blank

Preface

This Course on *Metrology and Fundamental Constants* was held in Varenna in July 2006 and was organised by the Italian Physical Society, the Istituto Nazionale di Ricerca Metrologica of Italy (INRIM), and the Bureau International des Poids et Mesures. Coincidentally 2006 marks the first year of the existence of INRIM, the new Metrological Institution of Italy , resulting from the merging of the Istituto di Metrologia "Gustavo Colonnetti" (IMGC) and the Istituto Elettrotecnico Nazionale "Galileo Ferraris" (IEN).

Besides this particular event, the School in Varenna, as well as the Summer School on Metrology organized by the BIPM in Paris, is justified by three facts, the need to provide, from time to time, a co-ordinated set of lectures which present the relevant progress in Metrology, the increasing intertwining between Fundamental Physics and the practice of Metrological Measurements, and, third, the flurry of new and unexpected discoveries in this field, with a correlated series of Nobel Prizes bestowed to individuals working in Fundamental Constants research and novel experimental methods.

This is the fourth of the Enrico Fermi Schools on Metrology and Fundamental Constants organized by the Italian Physical Society. The first was held in 1976, the second and the third respectively in 1989 and 2000, all of which were supported by the direct presence of BIPM via the Director *pro tempore* and the strong presence of the National Metrological Laboratories. This presence was felt in two ways, first by sending many of their experts as lecturers and, secondly, by supporting the attendance of a large number of their researchers at the School.

One of the most fascinating and exciting characteristics of metrology is its intimate relationship between fundamental physics and the leading edge of technology which is needed to perform advanced and challenging experiments and measurements, as well as the determination of the values and interrelations between the Fundamental Constants.

In some cases, such as the caesium fountains clocks or the optical frequency standards, the definition of the value of a quantity is, in the laboratory, in the region of 10^{-16} and experiments are under way to reach 10^{-18}.

Many of these results and the avenues leading to further advances were discussed during the School, along a major step in metrology, expected in the near future, which could change the "old" definition of the kilogram, still based on a mechanical artefact, toward a new definition resting on a fixed value of a fundamental constant. The current possibilities include a fixed value of the Planck Constants or the Avogadro Number. Several National Metrological Institutes (NMIs) and other organizations are collaborating, worldwide, in the International Avogadro Consortium, and a number of NMIs are involved in the so-called "Watt balance" in the mechanical watt (measurement of a force and of a speed) is directly compared with the electrical watt *in the same device* (measured via Josephson and von Klitzing effects) and which would led to a measurement of the Planck Constant.

The success of this fourth Course was made possible by the close co-operation and the dedication of many Institutions and individuals.

The Directors wish to thank the Italian Physical Society and the INRIM for having provided the financial support for the organization of the School and for the attendance of several student. Of the some seventy or so students attending the School, two thirds were supported by their parent Institution, and the others directly by the School.

The Directors also wish to express their warm thanks to all the lecturers and seminar speakers who offered their expertise to the students, not only during the scheduled lectures, but also by being available for discussions and seminars; their enthusiasm and competence were crucial elements for ths success of the school and were duly appreciated by the students.

A particular debt of gratitude must be expressed to Maria Luisa Rastello who acted as Scientific Secretary of the School, mainly in the three-year period needed to prepare and to organize the school. During the same period the Directors met on a number of times to optimise the program and to identify the Speakers to be invited.

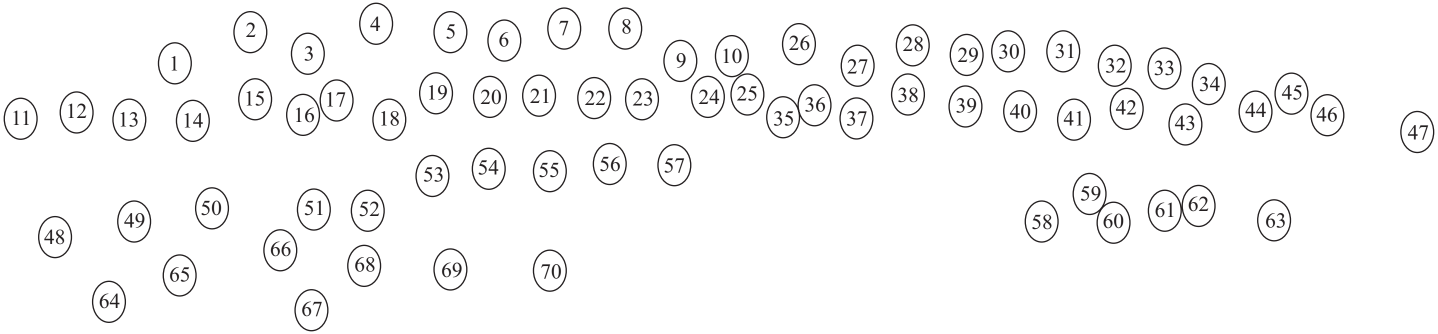
Finally the extremely valuable help and the friendly co-operation of Mrs. Barbara Alzani and Carmen Vasini acting on behalf of SIF must be acknowledged.

The Directors hope that the friendship created, as well as the information shared, will be valuable elements in building the students' future careers. For many metrologists, attendance at a the Varenna school is a seminal point in their training and development. The Directors look forward to seeing many of the 2006 students appear in the future as leaders and specialists in their fields. If the school has played even a small part in this, we shall have achieved our aim.

T. Hänsch, S. Leschiutta and A. J. Wallard

This page intentionally left blank

Società Italiana di Fisica
SCUOLA INTERNAZIONALE DI FISICA «E. FERMI»
CLXVI CORSO - VARENNA SUL LAGO DI COMO
VILLA MONASTERO 18 - 28 Luglio 2006



1) G. Cerretto
2) J. Pearce
3) N. DeLeo
4) Z. Barber
5) A. Krmpot
6) F. Schopfer
7) M. Bart
8) A. Kravchenko
9) D. Baines
10) F. Green
11) A. Piccato
12) F. Pennecchi
13) A. Meda
14) C. Boveri
15) V. Cacciatore
16) I. Sesia
17) S. Giunta
18) V. Schettini
19) B. Jeckelmann
20) R. Rusby
21) T. Quinn
22) R. Davis
23) J. Fischer
24) W. Bich
25) A. Malengo
26) G. Casa
27) A. Meristoudi
28) L. Oberto
29) P. Miglietta
30) M. Sellone
31) R. Winkler
32) S. Djordjevic
33) I. M. Iordache
34) M. Salazar
35) A. Tonina
36) G. D’Agostino
37) M. Siciliani de Cumis
38) C. Zumsteg
39) A. Mura
40) B. Trinchera
41) P. Amerio
42) R. Introzzi
43) F. Garoi
44) R. D. Geckeler
45) S. Vörös
46) F. Pythoud
47) C. Consejo
48) D. Quagliotti
49) D. Mari
50) N. Malossi
51) D. Gatti
52) N. Coluccelli
53) L. Cacciapuoti
54) A. Wallard
55) M. L. Rastello
56) T. W. Hänsch
57) S. Leschiutta
58) E. J. Salumbides
59) X. Baillard
60) A. L. Wolf
61) F. Chapelet
62) P. Wooliams
63) E. Puddu
64) S. Perero
65) A. Bernardi
66) B. Walton
67) L. Corengia
68) C. Hof
69) R. Brigatti
70) B. Alzani

This page intentionally left blank

Metrology and society

A. J. Wallard

Bureau International des Poids et Mesures - Sèvres, F 92312 France

1. – Introduction

Metrology has always had a major impact on society as, in a real sense, it grew from the needs of the market place and for standard measurements to ensure fair trade and consumer protection. Later, the needs of science began to play a role and so also served man's needs for an understanding of nature and the world in which they lived.

This aspect of metrology is as real today as it was many thousands of years ago. Of course, our needs are more sophisticated and place more challenges on our ability to measure and to ensure equivalence of measurements throughout the world. In general, and for good reasons, the lectures in the 2006 Varenna Summer School concentrated on the science of metrology as we practice it in our laboratories. "Varenna" aims to stimulate the understanding and appreciation of metrology for students and lecturers alike. However, a general appreciation of the driving forces behind our laboratory work is important so as to put it into context. An understanding of *why* governments and others are prepared to pay metrologists to carry out their research is important in today's world of formal programme management and so is the need to justify projects to those who pay for them.

This lecture therefore provides a general overview of some of the major applications of metrology to society.

2. – Metrology and trade

2˙1. *The general case.* – The origins of the importance of metrology and trade go back to the great treading nations of the early world. The Greeks, for example, kept copies of the standards held by the countries with which they traded.

In today's world, we are acutely aware of the growth of trade —about 15% per annum— between all nations and the need for products to conform to written specifications. The World Trade Organization, amongst others, points out that the growth in world trade helps to improve the standards of living in the developed as well as in the developing world. Measurements inevitably play a major role and some 80% of global trade is affected, one way or another, by the need to make measurements and to comply with written standards. A number of studies show that about 10% of the cost of production of products is the cost of making measurements on them and ensuring they meet customer's requirements.

The fact that measurement plays such a large part in the cost of production and in trade means that the need for measurements can, unfortunately, be used as a technical barrier to trade.

For example:

- Governments may establish local regulations so as to protect certain industries. Importing companies therefore may have to meet specifications which do not apply elsewhere and so either choose not to compete in that market, or have to pay more to do so.

- There may be specific requirements, such as the need to demonstrate traceability to a named National Metrology Institute (NMIs) may be built into national legislation or regulations. This can also apply to the major trading blocks, which were initially set up to protect free trade between their members. Either real, or perceived, barriers to trade commonly exist.

- Adoption of local specification standards or the need for tests to be carried out in designated laboratories, rather than those specification standards drawn up by the International Standardization (ISO), the International Electrotechnical Commission (IEC), or many other global standards-making bodies also mean higher costs for the company which chooses to compete in these markets.

- Even technology itself may create barriers. For example, in the case of a specified level of, say, residues in food, the authorities in the importing country and the company in the exporting country must be able to measure them with similar uncertainties. This may be a question of whether the exporter and importer can afford the same testing equipment.

Metrology therefore must ensure that measurements made in NMIs or in accredited laboratories, which are traceable to them, are equivalent.

2˙2. *Traceability in trade.* – The draft of the third *International Vocabulary of Metrology* (VIM) defines traceability as "*The property of a measurement result whereby the result can be related to a stated reference through a documented* unbroken chain of calibrations *each contributing to the measurement* uncertainty."

The important emphasis is on uncertainty and the need for the continuous, unbroken chain of measurement. Comparisons of standards or references are also a common way of demonstrating confidence in the measurement processes and in the reference standards held either in NMIs or in accredited laboratories. The National Accreditation Body usually takes care of these comparisons at working levels, sometimes called interlaboratory comparisons (ILCs) or proficiency testing.

At the NMI level, the increased relevance of traceable measurement to trade, and the need for demonstrable equivalence of the national standards held at NMIs, and to which national measurements were traceable, took a major turn in the mid-1990s. This event was stimulated by the need, from the accreditation community as much as from regulators and trade bodies, to know just how well the standards realized at all NMIs agreed with each other. The impossibility of comparing each and every one standard meant that a novel approach had to be adopted. In addition, it became increasingly clear that the important concept was one of measurements which were traceable *to the SI* through the standards maintained at NMIs, rather than to NMIs themselves.

2˙3. *Mutual recognition of NMI standards: the CIPM MRA.* – The result of the global concern about measurement traceability and equivalence at the global level led to the creation, by the International Committee for Weights and Measures (CIPM), of a Mutual Recognition Arrangement (MRA) for the recognition and acceptance of NMI calibration and test certificates. The CIPM MRA is one of the key events of the last few years, and one which may be as significant as the Convention du Mètre itself. The CIPM MRA has a direct impact on the reduction of technical business to trade and to the globalization of world business.

The CIPM-MRA was launched at a meeting of NMIs from Member States of the Metre Convention held in Paris on 14 October 1999, at which the directors of the NMIs of thirty-eight Member States of the Convention and representatives of two international organizations became the first signatories.

2˙4. *The essential points of the CIPM MRA.* – The *objectives of the CIPM MRA are*:

- to establish the degree of equivalence of national measurement standards maintained by NMIs;
- to provide for the mutual recognition of calibration and measurement certificates issued by NMIs; and
- thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.

The procedure through which an NMI —or any other recognized signatory— joins the MRA is based on the need to demonstrate their technical competence, and to convince other signatories of their performance claims. In essence, these performance claims are the uncertainties associated with the routine calibration services which are offered to customers and which are traceable to the SI. The laboratory concerned makes initial claims —called "calibration and measurement capabilities" (CMCs). They are first reviewed by technical experts from the local Regional Metrology Organization (RMO) and, subsequently, by other RMOs. The technical evidence for the CMC claims is generally based on the Institute's performance in a number of comparisons carried out and managed by the relevant CIPM Consultative Committees (CCs) or by the RMO. This apparently complex arrangement is needed because it would be technically, financially or organizationally impossible for each participant to compare its own SI standards with all others. The CIPM places particular importance on two types of comparisons:

- International comparisons of measurements, known as CIPM key comparisons and organized by the CCs and which generally involve only those laboratories which perform at the highest scientific level. The subject of a key comparison is chosen carefully by the CC to be representative of the ability of the laboratory to make a range of related measurements.
- Key or supplementary international comparisons of measurements, usually organized by the RMOs, and which include some of the laboratories which took part in the CIPM comparisons as well as other laboratories from the RMO. RMO Key Comparisons are in the same technical area as the CIPM comparison whereas supplementary comparisons are usually carried out to meet a special regional need.

Using this arrangement, we can establish links between all participants and so as to provide the technical basis for the comparability of the SI standards at each NMI. Reports of all the comparisons are published in the "Key Comparison Data Base" maintained by the BIPM on its web site.

These comparisons differ from those traditionally carried out by the CCs, which were largely for scientific reasons and which established the dependence of the SI realizations on the effects which contributed to the uncertainty of the realization. In CIPM and RMO key or supplementary comparisons, however, each participant carries out the measurements without knowing the performance of others until the comparison has been completed. They provide, therefore, an independent assessment of performance. The CIPM however took the view that comparisons are made at a specific moment in time and so required participating NMIs to install a quality system which could help demonstrate confidence in the continued competence of participants in between comparisons. All participants have chosen to use the ISO/IEC 17025 standard or ISO Guide 34 for some chemical measurement and have the option of a third party accreditation by an ILAC member or a self-declaration together with appropriate peer reviews.

The outcome of this process is that it gives NMIs the confidence to recognize the results of key and supplementary comparisons as stated in the database and therefore to accept the calibration and measurement capabilities of other participating NMIs. NMIs that are signatories to the CIPM MRA are entitled to use a logo (fig. 1) on their calibration certificates.

When drawing up its MRA, the CIPM was acutely aware that its very existence —and the mutual acceptance of test and calibration certificates between its members— might be seen as a technical barrier to trade in itself. The concept of "Associates" of the

Fig. 1. – The logo that can be used to identify calibration certificates issued by signatories to the CIPM MRA.

CGPM was therefore developed. An Associate has, in general, the right to take part in the CIPM MRA but not benefit from the full range of BIPM services and activities which are restricted to Convention Members. The Associate status is increasingly popular with developing countries as it helps them gain recognition world-wide and does not commit them to the additional expense of Convention Membership which may be less appropriate for them at their stage of development.

2˙5. *The key comparison database.* – The key comparison database, referred to in the CIPM MRA is available on the BIPM web pages (`www.bipm.org`). The content of the database is already evolving rapidly. Appendix A lists signatories, and Appendix B the set of key comparisons together with the results of from those that have been completed. It will also contain a list of those old comparisons that are to be used on a provisional basis. Appendix C contains the calibration and measurement capabilities of the NMIs that have already been declared and reviewed within their own Regional Metrology Organization (RMO) as well as those other RMOs that support the MRA.

2˙6. *Take up of the CIPM MRA.* – Whilst the KCDB data is, at the moment, largely of interest to metrologists, it is clear that a number of NMIs are keen to see it taken up more widely by regulators and others. This campaign is at an early stage but at the moment:

- an EU-US trade agreement cites the CIPM MRA as providing an appropriate technical basis for acceptance of measurements and tests; and
- the USA's NIST has opened up a discussion with the Federal Aviation Agency, FAA, and other regulators as to the way in which they can use the KCDB data to help the FAA accept the results of tests and of certificates which have been issues outside the USA.

There is a range of additional benefits and consequences of the CIPM MRA. For example, anyone can use the KCDB to look by themselves at the validated technical capability of any NMI within the CIPM MRA. As a result, they can, with full confidence, choose to use *its* calibration services rather than those of their national laboratory *and* have the results of these services accepted worldwide. They can also use the CIPM MRA database to search for NMIs that can satisfy their needs if they are not available nationally. This easy access and the widespread and cheap availability of information may well drive a globalization of the calibration service market under the CIPM MRA. This will be a real test of market economics.

NMIs, particularly those in EUROMET, the European Regional Metrology Organization, have used the KCDB data to benchmark their measurement capabilities against each other. This was an important element on the "MERA" project which led to a commitment by NMIs in EUROMET to explore a future in which they will be dependent on each other for the provision of services. In this arrangement, an NMI may give up a particular service and direct nationally based calibration service customers to another NMI.

2˙7. *The importance of Mutual Recognition Arrangements.* – The BIPM, the International Organization of Legal Metrology and the International Laboratory Accreditation Cooperation all maintain Mutual Recognition Arrangements which complement each other. All are also important, we believe, for use in trade and regulation. The three bodies therefore made a joint declaration, in January 2006, which drew the attention of the international user community to the importance of the use of the three MRAs in trade and regulation. The text of the tripartite declaration can be found on the web sites of the three organizations.

3. – Metrology and innovation

Whitworth's motto of "you can only make as well as you can measure" suggests that better measurement capabilities can lead to new concepts for new products or enhancements to current ones. Similarly, demands from user industries or groups for better measurement helps stimulate innovation and change in the metrology community.

It is clear that without the most demanding metrology, the microelectronics industry would not be able to continue to shrink the size of chips. Similarly, the ability to navigate better and to make accurate pressure measurements led to the decision by the world's airlines to reduce the separation between high-flying aircraft, so helping engines operate more efficiently thereby saving fuel and improving the cost effectiveness of the industry.

There are many case studies of how metrology drives innovation, amongst which are:

- The use of the world time scale, through GPS and other time systems, to navigate ferries with precision accuracy in crowded harbours.

- The efforts by metrologists to measure "appearance" (and semi-objective quantities such as how "shiny" something is) has led to improvements in the way in which the motor car industry can use different low-energy materials and metals in the same car —and can replace parts so as to maintain colour matching.

- The nanotechnology industry which is requiring better surface measurement techniques and even precision distance measurement at the nanoscale.

Many industries are now collaborating with metrologists in national laboratories to draw up “roadmaps” which set the direction in which an industry hope to improve its products and which show the metrology challenges which must be solved if they are to be taken to market.

Finally, metrology also stimulates innovation in science. Metrolgists have been prolific winners of Nobel Prizes because they operate at the limit of what is possible given the current state of science. As a recent Nobel laureate, Steve Chu said:

> “Accurate measurement is at the heart of physics, and in my experience new physics begins at the next decimal place.”

It is therefore no surprise that the 2006 Varenna School was honoured and stimulated by the presence of two recent Nobel Prize winners, Professors Hänsch and Phillips.

4. – Metrology and medicine

4˙1. *Uses.* – It may be relatively obvious that medicine has a great dependence on measurement. All doctors rely on tests —measurements of what is happening in our blood, for example. Similarly clinicians use tools to treat patients and need to rely on accurate measurement. The use of X-rays, or radioactivity, for example, to treat tumours is commonplace and ultrasound is used to treat muscular injuries or to check the progress of unborn children. All require the most accurate measurements possible of their energy because, in the case of radiation treatments, too large a dose can kill healthy tissue and too low a dose can require more treatments of the malignant tissue which is not destroyed. Ultrasound creates heat and so to avoid damage to an infant, the energies must be kept low, but high enough to provide adequate resolution.

Better measurement and control of therapeutic treatments using X-ray sources or electron beams are at the heart of the improvements in survival rate for cancer patients. The cost benefits of even apparently small reductions in measurement uncertainty lead to increased confidence by radiotherapists and others and have, no doubt, saved many thousands of lives.

These are metrology activities which have become relatively common over the last 30 years or so. However, it is only been in the last decade or so that we have started to make significant inroads into many of the other measurements required by clinicians as chemical metrology community has improved its ability to make them. This need has led to collaboration between the BIPM, the International Laboratory Accreditation Community and the International Federation for Clinical Chemistry and Laboratory Medicine. This collaboration has tackled the need to make comparisons of laboratories’ abilities to measure many of the common diagnostic chemicals used by the clinical community to reduce uncertainty and improve traceability of measurement. Similar collaborations are being opened up in the provision of reference samples of known purity for the calibration of diagnostic equipment in hospitals and in manufacturing industry. The initial success of this work has attracted the attention of the organizations responsible for the control and measurement of drugs used to enhance performance in sport and, because many of the techniques are similar, of the industries concerned with the measurement of chemical residues in food. Much of this work has been co-ordinated through the BIPM and its Chemistry section which serves the CIPM's Consultative Committee for Quantity of Material (metrology in chemistry).

For the industries concerned, improved abilities to make measurements bring substantial benefits. For example, the uncertainty associated with measurements of cholesterol in blood has been reduced from about 6% twenty years ago to about 3% with the best modern techniques. The level of cholesterol in blood is, of course, an important parameter because a clinician will prescribe expensive treatments if the level is above a certain "trigger" level but not so if the measurement is below it. Even with today's best practice, about 10% of patients are wrongly treated but the reduction, brought about by the halving of the measurement uncertainty, in the cost of prescribing drugs unnecessarily has saved the US health authorities over €100 million a year!

4˙2. *Why is chemical metrology "difficult"?* – Physical measurements are usually made on a property of a product, an artefact, or in an experiment where that measurement is well defined. Of course, good metrology practice is to explore the dependence of the results of a measurement on the other system parameters —the so-called "systematic or type-B uncertainties." It is a well-established technique with decades of physics to help understand the complexity of such measurements.

Chemical metrology is, of course, a much younger discipline but many of the current problems are associated with the fact that one tries to measure quite small quantities of a substance in cases where there is a much larger background or "matrix" and which can interfere with the measurement result. This interference can often result from the measurement technique used and can conflict with a metrologist's wish to increase the resolution of his measurement. Similarly, there are areas such as biological standards which are hard to deal with under the SI, although rapid progress is being made.

There are problems with increasing the awareness amongst practitioners as well as the regulators or others who legislate in this area. There is, unfortunately, little appreciation of concepts of uncertainty or error where the judicial system, for example, may require a metrologist to give an unambiguous "yes/no" answer rather than one which introduces the concept of measurement error or uncertainty. Similarly in areas of public concern, such as the proportion of GMOs in food, there was an initial tendency amongst some legislators to demand "zero GMO" levels. This is, metrologically speaking, a relatively meaningless statement. More education of, and collaboration between, regulators and measurement specialists at an early stage in legislation would do much to improve the current situation.

5. – And there is more!

This *résumé* of the lecture delivered at Varenna has, like the lecture itself, only scratched the surface of the impact which metrology makes on society. It has not addressed measurements related to climate change, for example, as these were dealt with in other lectures.

As metrologists, though, we can take pleasure and derive great satisfaction from the fact that our work has a direct benefit to society. Studies show the enormous rate of economic return from national investments in metrology and it is frequently easy to derive economically robust benefit/cost ratios which dwarf many other public investments. The reasons are often discussed in summer schools like Varenna either in the lectures or in the lively social life which goes on in the restaurants and bars of the delightful Italian town by the lake which was our host for a fortnight.

The answer to our dilemma is, though, with you . . . the practicing metrologists in the NMIs and universities or institutes where our subject is studied or taught. It is clearly important to talk about the excitement of our science, but it may, in the long run, be more important to write about and promote the impact and benefits of what we do. It is a frequently stated remark that metrology is an undervalued and under-recognized pursuit, and that we do too good a job. Things do not, in general, go wrong because of a failure in metrological science itself and, as more than one commentator has, somewhat cynically, said; "we need a major problem to draw attention to our subject and to ask for more resources". Whilst this may be true, the associated fact is that, because of the direct, and huge, benefits we bring to society, there would be damage to people, to health, or to trade. We must therefore, fall back on our powers of persuasion or advocacy!

This page intentionally left blank

The evolution of metrology: Past times to the present day

A. J. Wallard

Bureau International des Poids et Mesures - Sèvres, F 92312 France

1. – The roots and evolution of metrology

1˙1. There is nothing much that is new about the concept of making measurements. From the earliest days it has been important to compare measurements made in one place or another. Of course, this initially had much to do with fair exchange, barter, or trade between local communities. The early weights were simple stones, and parts of the body, like hands and arms (the "cubit") were adequate for most needs in length measurement. But as the need to trade or to exchange goods grew, so did people's needs for greater accuracy or for reference standards that did not change too much and were, in some way, equivalent. Initially, wooden length bars were easy to compare and carefully formed weights could be weighed against each other. Indeed, various forms of balance were commonplace in early history and even in religion. Early tomb paintings show the Egyptian god Anubis weighing, with a balance which many of us today would easily recognize, the soul of the dead against an ostrich feather —the sign of purity. Measurement was well established in society and the recording of dimensions was seen as of interest and as important.

1˙2. A steady progression from basic artefacts to naturally occurring reference standards has been part of the entire history of metrology. An early example was found in nature and many metrologists are familiar with the use of the carob seed in the early Mediterranean civilizations as a natural reference for length and for weight and hence volume. Several studies have shown the consistency of the dimensions of the seed and it is, of course, the predecessor of the modern use of the carat for precious stones.

1˙3. Fairness in the marketplace and the needs of the trading nations were major drivers in the development of early metrology. The Greeks were almost certainly one of the best examples of early traders who paid attention to metrology and they were known to keep copies of the weights and measures of the countries with which they traded. Fraud was the other driving force and probably has a good claim to be the second oldest profession!

1˙4. But it was soon clear that there was substantial metrological confusion as different systems evolved in different continents and parts of the world and differed by rather a lot. Kings, Queens and Governments soon realized that they could make money by levying taxes on market transactions and we saw efforts to move towards a framework within which there was some consistency. The English "Magna Carta" of 1215 did many things to set out a framework for citizens' rights and wanted to establish "one measure throughout the land". Nevertheless, imposing consistency was a hugely over-ambitious target and a number of efforts during the early Middle Ages in Europe and elsewhere saw little progress. The problems, though, were clearly recognized and nearly every market place in the world contained replicas of local standards that had to be used for day to day measurements.

1˙5. Traceable measurement was born in the Middle Ages and saw accuracies of a few percent becoming commonplace. Just as in modern metrology, though, the reference standards themselves had to be improved as people wanted better accuracy. The problem with many length measurements was that the standard was made of the commonly available brass or bronze with a fairly large coefficient of expansion. Iron or steel were not developed for a couple of centuries or so. Brass and bronze therefore dominated the length reference business until the early 19th Century by when metallurgy had developed well enough —though still was something of a black art— for new lower-expansion metals to be used and reference conditions to be quoted.

1˙6. Science, fortunately, was taking a parallel path and we have to go back a little to the mid-18th Century, when Britain and France compared their national measurement standards and realized that they differed by a few percent for the same unit. Although the English system was reasonably consistent throughout the country, the French found that differences of up to 50% were common in length measurement. The ensuing technical debate was the start of what we now accept as the metric system. France led the way, and even in the middle of the French Revolution, the Academy of Science was asked to "deduce an invariable standard for all the measures and all the weights". The important word was "invariable". There were two options available —the "seconds pendulum" and the length of an Earth's meridian. The obvious weakness of the pendulum approach was that its period depended on the local acceleration due to gravity. The Academy therefore chose to measure a portion of the Earth's circumference and relate it to the "official" French metre. This led to the famous adventures of two remarkable men —Delambre and Mechain— who were entrusted with a survey of the meridian between Dunkirk and Barcelona. Despite, or more probably because of, having a royal warrant, issued by Louis

XVI in the year before he was executed, that was supposed to protect them, they were regularly thrown into prison and their instruments impounded. Progress was slow. To make matters worse, the Reign of Terror closed down the Academy and removed several famous names of French science —Laplace, Coulomb, and Delambre himself. Lavoisier was guillotined. Nevertheless, the work of Delambre and Mechain led to the famous "Metre des Archives"— a platinum end standard.

1˙7. The rest is, as they say, history and the eventual resolution of the scientific discussion saw the emergence of a recognition —certainly at the scientific, but also at the Government level— of a consistent and increasingly international system for scientific and industrial measurements. This period saw the birth of the metric system and the precursor what we now know as the SI or International System of Units.

1˙8. The international nature of measurement was driven, just like the Greek traders of nearly 2000 years before, by the need for interoperability in trade and advances in engineering measurement. The displays of brilliant engineering at the Great Exhibitions in London and Paris in the mid-19th Century largely rested on the ability to measure well. The giants of engineering met and compared notes on what limited precision engineering and why different lathes out-performed others. This boiled down to huge, detailed debates about how to make flat surfaces and the merits of end and line gauges for dimensional measurement references. The British Victorian engineer Joseph Whitworth —who coined the famous phrase "you can only make as well as you can measure" and who pioneered accurate screw threads— was deeply impressed by the advances. He immediately saw the potential of end standard gauges rather than line standards where the reference length was defined by a scratch on the surface of a bar as optical microscopes were not then good enough to compare measurements of line standards well enough for the best precision engineering. He was determined to make the world's best measuring machine, taking up the challenge to make a quite remarkable instrument —called the "Millionth Machine", based on the principle that touch was better than sight for precision measurement. It appears that the machine had a "feel" of about 1/10 of a thousandth of an inch.

1˙9. The other interesting metrological trend at the 1851 Great Exhibition was military. The ordinary rifle led the way and is linked with the great American engineer Eli Whitney who, in 1798, promised the US government that he could make 12 000 identical muskets. To some extent this was rising to the challenge offered by a French gunsmith, le Blanc, who persuaded the American President Thomas Jefferson that he could mass-produce musket locks. Sadly, both men failed in their ambition because tooling was not good enough and a US Congressional report on "interchangability" in 1827 admitted that even the national armoury at Harpers Ferry could not easily do what was needed. Guns again played a role when the Sam Colt armoury at Hartford supported the work of two young toolmakers —Francis Pratt and Amos Whitney, who adopted the English gauge system based on Whitworth's contribution. After the end of the American Civil war, the Pratt and Whitney company manufactured the best end gauges and Brown and Sharpe developed precision grinding machines. Slip gauges were also developed by Whitworth and improved by Johansson in Sweden. The Swedes had better material technology and therefore dominated the world market until the First World War during which sources of high-quality gauge blocks were not available in the US and UK. They therefore had rapidly to be developed in Pratt and Whitney (USA) and in the BSA company in Enfield (UK).

2. – 20th Century metrology —the National Metrology Institutes

2˙1. At the turn of the 19th Century, many of the industrialized countries had set up National Metrology Institutes —generally based on the model established in Germany of the Physikalisch Technische Reichanstalt which had been founded in 1887. The economic benefits of such a national institute were immediately recognized and as a result, scientific and industrial organizations in a number of industrialized countries began pressing their Governments to make similar investments. In the UK, the British Association for the Advancement of Science reported that, without a national laboratory to act as a focus for metrology, Great Britain's industrial competitiveness would be weakened. The cause was taken up more widely and the UK set up the National Physical Laboratory in 1900. The US created the National Bureau of Standards in 1901 as a result of similar industrial pressure. A number of other countries already had national laboratories, although they had largely been established for verification and testing of measuring instruments. The major "National Metrology Institutes" (NMIs), however, had a dual role. In general they were also the main focus for national research programmes on applied physics and engineering. Their scientific role in the development of the International System of Units, the SI, began, however, to challenge, and even take over, the role of the universities which then were much more concerned with the measurement of the fundamental constants. This was especially true after the development of "quantum physics" in the 1920s and 1930s. Most early NMIs, therefore, began with two major elements to their mission:

- a requirement to satisfy industrial needs for accurate measurements, through standardizing and verifying instruments; and
- determining the physical constants so as to improve and develop what was to become the SI system.

In many ways, their missions now are very similar although the scope and range of measurement responsibilities continues to grow.

2˙2. The new industries which emerged after the First World War made huge demands on metrology and, together with mass production and the beginnings of multinational production sites, raised new challenges which brought NMIs into direct contact with companies and which saw a close link develop between them. At that time —and even up until the mid 1960s, nearly all the calibrations and measurements which were necessary for industrial use were made in the NMIs and in the "gauge rooms" of the major companies, as most measurements were in engineering metrology. The industries

Fig. 1. – The seven base units of the SI.

of the 1920s, however, developed a need for electrical and optical measurements so NMIs expanded their coverage and their technical abilities to meet the need.

2˙3. The story since then was one of steady technical expansion until after the Second World War. In the 1950s, though, there was a renewed interest in a broader applied focus for many NMIs so as to develop civilian applications for much of the de-classified military technology. Many Governments also used them as the focus for science-based research into a whole range of new techniques, and industries, not all of which had a direct bearing on metrology: it was the era of the "National Scientific Champion". The squeeze on public budgets in the 1970s and 80s saw a return to "core metrology" and many other institutions —public and private— took on responsibility for developing many of the technologies which had been initially fostered at NMIs. As an example, NPL's early aeronautics and radar work was transferred to other laboratories and its early expertise in computers and "basic physics" migrated to industry and to the expanding university infrastructure. NMIs adjusted to their new roles. Many restructured and formed new, often improved, ways of serving industrial needs. This "recreation" of the NMI role was also shared by most Governments which increasingly saw them as tools of industrial policy with a mission to stimulate industrial competitiveness and, at the end of the century, to reduce technical barriers to world trade.

3. – The international System of Units, the SI

3˙1. In introducing the international System of Units, (fig. 1) we need first to pause and ask what attributes we want from a universal system of units.

Essentially we need a system which:

– creates a system or hierarchy of measurement units and quantities;

– defines the units; and

– states how multiples and submultiples are formed.

The realization of the units on which the system is based must put the definition into practice. This means that it:

– ensures that the units do not change with time outside their accepted reproducibility;

– is reproducible throughout the world; and

– can be transferred to users without significant loss of accuracy.

3·2. There have been many systems of units which predated the SI which was launched in 1960. All the previous systems emerged from the needs of specific areas of science or engineering. In the 19th century, scientists, for example, dealt with the "centimetre, gram, second" system of units which was convenient for theoretical studies as many of the fundamental constants were simply "set to one."

On the other hand, the electrical community dealt with "practical" units based on the volt, ohm and ampere.

The two systems were merged in the 1920s into a system based on the metre, kilogram, second and ampere —a system which satisfied many needs until the 1950s. The SI was based initially on six "base" quantities (mass, length, time, electric current, thermodynamic temperature and luminous intensity.) The "quantity of matter" —the mole— was added in 1971. The essential hierarchical advantage of the SI is that virtually all physical and chemical quantities can be expressed in a combination of the base units of the SI. A more detailed description of the SI and its current definitions can be found in the "SI brochure" which can be downloaded from the BIPM web site (`www.bipm.org`).

4. – Physics, engineering ... and then chemistry

4·1. Chemical metrology began to appear as a serious discipline in the 1960s and in 1971, a seventh base unit, the mole, was added to the SI.

4·2. Chemistry, biosciences and pharmaceuticals are, for many of us, the "new metrology". We are used to the practices of physical and engineering metrology and so the new technologies are challenging our understanding of familiar concepts like traceability, uncertainty and "primary" standards. Much depends here on the interaction between a particular chemical or species and the solution or matrix in which it is to be found, as well as the processes or methods used to make the measurement. We have much to learn, and at the same time, much to contribute. We are only beginning to tackle and respond to the world of medicine and pharmacy and have created a partnership with the International Federation of Clinical Chemistry (IFCC) and the International Laboratory Accreditation Cooperation (ILAC) to address these needs in a Joint Committee for Traceability in Laboratory Medicine, the JCTLM. This is directed initially at a database of reference materials which meet certain common criteria of performance and of laboratories with worldwide-accepted measurement capabilities. Recognition of the data in the JCTLM data base will in particular help demonstrate compliance of the products of the *In Vitro diagnostic* industry with the requirements of a recent European Union Directive.

5. – Metrology in the 21st Century

5˙1. *Industrial challenges.* – In concluding, it seems appropriate to take a short glance at what the future may have in store for world metrology and, in particular, at the new industries and technologies which require new measurements.

The first challenge is the focus by manufacturers on the design or appearance of a product which differentiates it in the eyes of the consumer, from those of their competitors. These rather subjective attributes of a product are starting to demand an objective basis for comparison. “Appearance” measurement of quantities like gloss, or the need to measure the best conditions in which to display products under different lighting or presentation media (such as a TV tube, flat panel display, or a printed photograph), combine “hard” physical or chemical measurements with the subjective and varying responses of, say, the human eye or ear. Yet these are precisely the quantities that a consumer uses to judge textiles, combinations of coloured products, or the relative sound reproduction of hi-fi systems. They are therefore also the selling points of the marketer and innovator. How can the consumer choose and differentiate? How can they compare different claims (washes whiter than our competitors’ products, gives a shine to your pet’s fur, ...)? Semi-subjective measurements are moving away from the carefully controlled conditions of the laboratory into the supermarket and are presenting exciting new challenges.

NMIs have already become familiar with the needs of users of their colour or acoustical measurement services which require a degree of modelling of the consumer response and the differing reactions depending on environmental conditions such as ambient lighting or noise backgrounds. The fascination of this area is that it combines objective metrology with physiological measurements and the inherent variability of the human eye or ear, or the ways in which our brains process optical or auditory stimuli.

The second area of challenge is familiar in the sense that it tackles the needs of new industries, exploits new technologies and further enhances our ability to measure the large, the small, the fast and the slow. Microelectronics, telecommunications, the study and characterization of surfaces and thin films will benefit. Many of these trends are regularly analyzed by NMIs as they formulate their technical programmes and a summary can be found in the recent report by Dr. Robert Kaarls entitled “Evolving Needs for Metrology in Trade, Industry, and Society and the Role of the BIPM”. The report can be found on the BIPM web site.

Metrology has more recently been applied to climate change and to our environment and basing a number of environmental measurements on the SI may help us extend our current knowledge of the complex interactions of weather, sea currents and the various layers of our atmosphere. In order to do this, the metrologist is beginning to make a recognized contribution by insisting that these slow or small changes should be measured traceably and against the unchanging reference standards offered through the units of the SI system. Similar inroads are being made into the space community where, for example, international and national space agencies are starting to appreciate that solar radiance measurements can be unreliable unless related to absolute measurements. We await the satellite launch of a cryogenic radiometer which will do much to validate solar physics.

The relevance of these activities is far from esoteric. Governments spend huge sums of money in their efforts to tackle environmental issues and it is only by putting measurements on a sound basis that we can begin to make sure that these investments are justified and are making a real difference in the long term.

A third area is the trend towards "quantum based" standards in industry. This is a result of the work of innovative instrument companies which now produce, for example, stabilized lasers, Josephson junction voltage standards and atomic clocks for the "mass market". The availability of such highly accurate standards in industry is itself testimony to companies' relentless quest for improved product performance and quality. But it is all too easy to get the "wrong" answer. Users are advised to undertake comparisons and to co-operate closely with their NMIs to make sure that these instruments are operated with all the proper checks and with attention to best practice so that they may, reliably, bring increased accuracy closer to the end user. Industry presses NMIs —rightly so— for better performance and in some areas of real practical need NMI measurement capabilities are still rather close to what industry requires. It is, perhaps, in these highly competitive and market driven areas that the equivalence of reference standards will prove their worth. Companies specify the performance of their products carefully in these highly competitive markets and any significant differences in the way in which NMIs realize scales and quantities will have a direct bearing on competitiveness, market share and profitability.

The industries of today and tomorrow are starting to nibble away at one of the century-old metrology practices within which the user must bring their own instruments and standards to the NMI for calibration. Some of this relates to the optimization of industrial processes where far more accurate, "real time", "in-process" measurements are made. The economics of huge production processes demand "just in-time" manufacture active data management and clever process modelling. By identifying, reliably, where sub-elements of a process are behaving poorly, plant engineers can take rapid remedial action and so identify trouble spots quickly. But real "real time" measurements are difficult and it is only recently that some NMIs have begun to address the concept of an industrial measurement "system". New business areas such as this will require NMIs to work differently if for no other reason that their *customers* work differently and they need to meet their requirements. Remote telemetry, data fusion and new sensor techniques are becoming linked with process modelling, numerical algorithms and approximations so that accurate measurement can be put —precisely— at the point of measurement. These users are already adopting the systems approach and some NMIs are starting to respond to this challenge.

A few NMIs are also starting to ask how the Internet can change the way they traditionally do things. Wide bandwidth offers a combination of fast data transfer, complete with video. A whole new world opens up and —just as NMIs have broadcast accurate global time signals for decades— they can now test and calibrate a range of electronic devices over the "net".

5'2. *New science—changes to the SI.* – Science never stands still. There are a number of trends, already evident, which may have a direct bearing on the definitions of the SI itself. Much of this is linked with progress in measuring the fundamental constants, their coherence with each other and the SI, and the continued belief that they are time and space invariant. In the next few years we can expect, perhaps, a redefinition of the kilogram based on the results of several experiments in a number of NMIs to relate the mechanical kilogram to the electrical quantities with uncertainties which make it possible to monitor any changes in the prototype itself. Length and time units have been related by a fixed value for the speed of light for a number of years, and length is defined in relation to the time unit. New femtosecond laser techniques and the high performance of ion or atom trap standards may soon enable us to turn this definition on its head and define time in terms of the optical transitions. Temperature and light intensity measurement all relate to the Boltzmann constant and if several efforts are to improve our knowledge of it and the Planck constant are successful, then they might open up the possibility of a different approach to thermodynamic temperature and light intensity measurement. These are exciting times for metrologists. Trends in measurement are taking us into new regimes; chemistry is introducing us to the world of reference materials and processes and to application areas which would have amazed our predecessors.

This page intentionally left blank

The organization of metrology

A. J. Wallard

Bureau International des Poids et Mesures - Sèvres, F 92312 France

1. – Structures

The formal structures of metrology and how it operates at the national and international levels are not usually at the forefront of a young metrologist's mind. Nevertheless, an understanding of them and how they operate becomes more important as the young metrologist progresses up the hierarchy and starts to represent his or her laboratory or country at an international level. Some of the activities which are launched by these organizational bodies also shape the daily work of our young metrologist and an understanding of why they are there can often help to put the experimental work into context. The aim of this Varenna lecture is, then, to give an outline of how structures emerged and to describe today's system functions.

2. – Metre Convention

2˙1. *The Metre Convention.* – The Metre Convention sets many of the structures in the international system. Its roots lie in the 1851 Great Exhibition and the 1860 meeting of the British Association for the Advancement of Science (BAAS) in which a number of scientists and engineers met to develop the case for a single system of units based on the metric system. This built on the early initiative of Gauss to use the 1799 metre and kilogram in the Archives de la République in Paris as a well as the second defined in astronomy as a coherent set of units for the physical sciences. In 1874, the three-dimensional CGS system, based on the centimetre, gram and second, was launched by the BAAS. The sizes of the electrical units which related to the CGS system were not particularly convenient and, in the 1880s, the BAAS and the International

Electrotechnical Commission (IEC) approved a set of practical electrical units based on the ohm, the ampere, and the volt. In parallel with this attention to the units, a number of Governments set up what was then called the "Committee for Weights and Money" which in turn led to the 1870 meeting of the "Commission Internationale du Metre." Twenty-six countries accepted the invitation of the French Government to attend. However, only 16 were able to come as the Franco-Prussian War intervened so the full committee did not meet until 1872. The result was the Convention du Metre and the creation of the Bureau International des Poids et Mesures, the BIPM (in English the International Bureau of Weights and Measures), in the old Pavillon de Breteuil at Sévres as a permanent scientific agency supported by the signatories to the Convention. As this required the support of Governments at the highest level, the Metre Convention was not signed until 20 May 1875.

2˙2. *The BIPM*. – The BIPM's role was to "establish new metric standards, conserve the international prototypes" (then the metre and the kilogram) and "to carry out the comparisons necessary to assure the uniformity of measures throughout the world." As an intergovernmental, diplomatic Treaty organization, the BIPM was placed under the authority of the "General Conference on Weights and Measures" (CGPM). A committee of scientific experts —the International Committee for Weights and Measures (CIPM), supervises the running of the BIPM. The aim of the CGPM and the CIPM was to assure the "international unification and development of the Metric System." The CGPM now meets every four years to review progress, receive reports from the CIPM on the running of the BIPM and to establish the operating budget of the BIPM, whereas the CIPM meets annually to supervise BIPM's work. When it was set up, the staff of BIPM was "a Director, two assistants and the necessary number of employees." In essence, then, a handful of people began then to prepare and disseminate copies of the international prototypes of the metre and the kilogram to Member States. About 30 copies of the metre and 40 copies of the prototype kilogram (fig. 1) were distributed to Member States by ballot. Once this was done, some thought that the job of the BIPM would simply be that of periodically checking (in the jargon, verifying) the national copies of these standards. This was a short lived vision as the early investigations immediately showed the importance of measuring, reliably, a range of quantities which influenced the performance of the international prototypes and their copies. As a result, a number of studies and projects were launched, which dealt with the measurement of temperature, density, pressure and a number of related quantities. BIPM immediately became a research body although this was not recognized formally until 1921!

2˙3. *The MKS system.* – Returning to the development of the SI, however, one of the early decisions of the CIPM was to modify the CGS system to base measurements on the metre, kilogram and second —the MKS system. In 1901, Giorgi showed that it was possible to combine the MKS system with practical electrical units to form a coherent four-dimensional system by adding an electrical unit and rewriting some of the equations of electromagnetism in the so-called "rationalized" form. In 1946, the

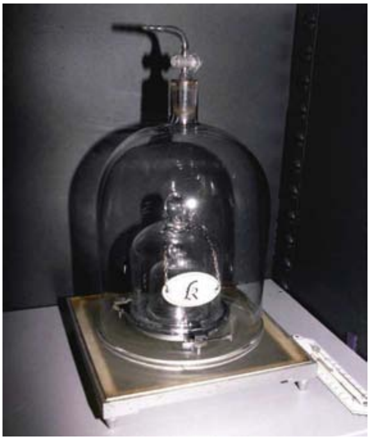


Fig. 1. – The international prototype of the kilogram.

CIPM approved a system based on the metre, kilogram, second and ampere —the MKSA system. Recognizing the ampere as a base unit of the metric system in 1948, and adding, in 1954, units for thermodynamic temperature (the kelvin) and luminous intensity (the candela), the 11th CGPM in 1960 coined the name Système international d'Unités, the SI. At the 14th CGPM in 1971 the present day SI system was completed by adding the mole as the base unit for the amount of substance, bringing the total number of base units to seven. Using these, a hierarchy of derived units and quantities of the SI have been developed for most, if not all, measurements needed in today's society.

2˙4. *The SI Brochure*. – A substantial treatment of the SI is to be found in the 8th edition of the "SI Brochure" published by the BIPM in 2006 and available on the BIPM web site.

3. – BIPM: the first 75 years

3˙1. After its intervention in the initial development of the SI, BIPM continued to develop fundamental metrological techniques in mass and length measurement but soon had to react to the metrological implications of major developments in atomic physics and interferometry. In the early 1920s, Albert Michaelson came to work at BIPM and built an eponymous interferometer to measure the metre in terms of light from the cadmium red line —an instrument which, although modified, did sterling service until the 1980s! In temperature measurement, the old hydrogen thermometer scale was replaced with a thermodynamic-based scale and a number of fixed points. After a great debate, electrical standards were added to the work of the BIPM in the 1920s with the first international comparisons of resistance and voltage. In 1929 an electrical laboratory was added and photometry arrived in 1939.

3˙2. Even in these very early days, it was clear that the BIPM needed to find a way of consulting and collaborating with the world's NMIs. The solution adopted is one which still exists and flourishes today. It was clear that the best way of working was in face-to-face meetings, so the concept of a "Consultative Committee" to the CIPM was born. Members of the Committee were drawn from experts active in the world's NMIs and met to deal with matters concerning the definitions of units and the techniques of comparison and calibration. The Committee is usually chaired by a member of the CIPM. Much information was shared although, for obvious logistical reasons, the meetings were not too frequent. Over the years, the need for new Consultative Committees grew in reaction to the expansion of metrology and now ten Consultative Committees exist, with over twenty-five working groups.

3˙3. This has proved to be a successful approach and provides a clear unifying focus for the people concerned. In recent years there has been an increase in the frequency of meetings, and most Consultative Committees meet every two or three years. The CIPM is rightly cautious about establishing new Consultative Committee but proposals for new ones are considered from time to time, usually after an initial survey through a Working Group. In the last ten years, Joint Committees have been created to tackle cross-cutting issues such as the international approach to the estimation of measurement uncertainties or to the establishment of a common "vocabulary" for metrology. Most of these Joint Committees bring BIPM together with international bodies or intergovernmental organizations such as ISO, ILAC or IEC. As the work of the Metre Convention expands away from physics and engineering, Joint Committees are an excellent way of bringing BIPM together with other bodies which bring specialist expertise —an example being the Joint Committee for Traceability in Laboratory Medicine, established recently with the International Federation of Clinical Chemistry.

3˙4. The introduction of ionizing radiation standards to the work of the BIPM came later even though a local resident, Marie Curie, deposited her first radium standard at BIPM in 1913. As a result of pressure, largely from the USSR delegation to the CGPM, the CIPM took the decision to add ionizing radiation and laboratories in 1964. This was a significant new investment and the CGPM agreed to a doubling of the BIPM budget together with some special "exceptional" contributions to cope with the new work. Since then no budget increases other than allowances for inflation had been made until the General Conference in 2003!

3˙5. At the beginning of the 1960s the BIPM still had only about 25 staff. In the mid-60s, and at the time of the expansion into ionizing radiation, programmes on laser length measurement were also started. These contributed greatly to the redefinition of the metre in 1983, the BIPM acting as the world reference centre for laser comparisons in much the same way as it did for physical artefact-based standards. In the meantime, however, the metre-bar had, however, already been replaced, in 1960, by an interferometric-based definition using optical radiation from a krypton-lamp.

3˙6. The end of BIPM's first 75 years saw a laboratory of about 50 with a clear and unchallenged mission and a steady demand for BIPM's calibration and comparison services.

3˙7. As a result of these expansions, the number of staff rose steadily and reached about 60 in 1985. Within its budget, the BIPM therefore managed to keep technical work going in most of the main areas of interest to the physical and engineering worlds.

3˙8. In 1985, the BIPM absorbed the time section of the Bureau International de l'Heure based at the Paris Observatory and reduced a few other activities so that this could be done within the same global budget.

3˙9. Staff numbers increased to nearly 70 in the mid 1980s but some signs were starting to appear that BIPM could —and should— no longer cover all the main metrological disciplines. Its tasks were also starting to change and its role at that time of a small, research body charged with regular comparisons of high level standards from a relatively stable number of top NMIs as well as a calibration service for Member States of the Metre Convention, was under threat.

3˙10. This period, in retrospect, can be seen as leading in the beginnings of a distinct change in BIPM's work. Areas of activity were removed —either because they became technically redundant or for financial reasons and because they no longer were seen as essential, as the rise in the competence and the number of National Laboratories sometimes made it unnecessary for there to be a unique international resource. Most developing countries also invested in a metrology infrastructure to meet the needs of their industries. This corresponded to a steady increase in the numbers of members of the Metre Convention, to 43, in 1975. However, as nations preferred to invest nationally rather than internationally, the BIPM budget failed to keep pace with demand and work stopped on temperature measurement, basic dimensional measurement, neutrons and some traditional electrical measurements.

3˙11. Apart from its responsibility to maintain the international prototype kilogram —still the only artefact-based unit of the SI— BIPM was therefore no longer the sole repository of an international primary reference standard. However there were, and still are, a number of unique reference facilities at BIPM for secondary standards and quantities of the SI.

3˙12. If it was going to maintain its original mission of a scientifically based organization with the responsibility of co-ordinating world metrology, BIPM recognized that it needed to discharge particular aspects of its treaty obligation in a different way. It also saw the increased value of developing the links needed to establish collaboration at the international and intergovernmental level. It was also faced with the need to provide the Secretariat to ten Consultative Committees of the CIPM as well as an increasing number of working groups. The last ten years have, therefore, seen the start of a significant change in BIPM's way of working. During this period we have also been faced with the need to develop a world metrology infrastructure in new areas of environmental, chemical, medical and food. The shift away from physics and engineering was possible, fortunately, as a result of the changing way in which the SI units can be realized, particularly through the quantum-based standards. Other pressures for co-ordination, rather than independent research, resulted from the increasingly intensive programme of comparisons brought about by the launch of the CIPM's Mutual Recognition Arrangement in 1999s.

3˙13. The most recent consequence of these trends was that the CIPM decided that the Photometry and Radiometry section would close, ending nearly 70 years of such activity at the BIPM. Additional savings would also be made by restricting the work of the laser and length group to a less ambitious programme. This led, in 2006, to the closure of the Length section.

3˙14. A small, 130 year old institution therefore was in the process of "re-inventing" itself so as to take on and develop a changed but nevertheless unique niche role. This was still based on technical capabilities, but was one which had to meet the changing —and expanding— requirements of its Member States in a different way. Much more needed to be done as the benefits of precise, traceable measurement became seen as important in a number of "new" disciplines for metrology. It was, for example, impossible to ignore the real needs which were emerging in medicine and food or to reduce the effort in support of the CIPM-MRA which was already finding application in the reduction of technical barriers to trade.

3˙15. This change of emphasis was endorsed at the 2003 General Conference on Weights and Measures which agreed on a new four-year work programme (2005-9) as well as the first real terms budget increase since the increase agreed in the mid 1960s to finance the expansion into ionizing radiation.

3˙16. Today, the Membership of the Metre Convention stands at 51 Member States.

4. – National structures: The National Metrology Institutes (NMIs)

4˙1. *National Metrology Institutes.* – At the turn of the 19th Century, many of the industrialized countries had set up National Metrology Institutes —generally based on the model established in Germany of the Physikalisch Technische Reichanstalt which had been founded in 1887. The economic benefits of such a national Institute were immediately recognized and, as a result, scientific and industrial organizations in a number of industrialized countries began pressing their Governments to make similar investments. In the UK, the British Association for the Advancement of Science reported that, without a National laboratory to act as a focus for metrology, Great Britain's industrial competitiveness would be weakened. The cause was taken up more widely and the UK set up the National Physical Laboratory in 1900. The US created the National Bureau of Standards in 1901 as a result of similar industrial pressure. A number of other countries already had national laboratories, although they had largely been established for verification and testing of measuring instruments. The major "National Metrology Institutes" (NMIs), however, had a dual role. In general they were the main focus for national research programmes on applied physics and engineering. Their scientific role in the development of the International System of Units, the SI began, however, to challenge and even take over the role of the Universities which were much more concerned with the measurement of the fundamental constants. This was especially true after the development of "quantum physics" in the 1920s and 1930s. Most early NMIs therefore began with two major elements to their mission:

- a requirement to satisfy industrial needs for accurate measurements, through standardizing and verifying instruments; and
- determining the physical constants so as to improve and develop the SI system.

In many ways, their missions now are very similar although the scope and range of measurement responsibilities continues to grow.

4˙2. *NMI science.* – The core role of the NMIs was to realize the definitions of the units and to offer calibration services to industrial and other units. Of course, it was always important to keep one step ahead of industrial need and NMIs were therefore research bodies, often operating at the forefront of science.

5. – Traceability: The birth of accreditation

5˙1. *Traceability.* – Traceability of measurement has been a core concern of the Metre Convention from its inception. Initially a measurement is always made in relation to a more accurate standard reference and these references were themselves calibrated or measured against an even more accurate standard. The chain follows the same pattern until one reached the national standards. The NMIs' job was to make sure that the national standards were accurate enough to meet national needs.

5˙2. *Accreditation bodies.* – Originally NMIs did everything and tested or calibrated all the routine instruments used in industry. Users brought their instruments to NMIs themselves with the result that the calibration load was huge. In the UK, the NPL's own records show that over 563 000 tests and calibrations were done in 1960. This clearly was not efficient, nor did it allow NPL's scientists and engineers to research many of the new techniques that were increasingly relevant to metrology. The solution lay in the creation of national calibration services —networks of competent laboratories which took on the routine work and which released NMI resources. This was the beginning of the accredited laboratory network, which now has a huge role to play in the hierarchy of traceable calibration and quality assurance, and to product quality. In order to enhance efficiency and to innovate, many NMIs turned to automation of measurement so as to relieve the metrologist of much tedious measurement practice. Automation took away the subjectivity of many slow, operator-based, techniques, improved measurement uncertainty and reduced the variation in measurement that came from different operators.

5˙3. *ILAC.* – As the numbers of accredited laboratories increased, largely as a result of new international quality standards, their market for traceable calibrations grew and accreditation is now a major business worldwide. Laboratory accreditation, and the written standards which underpin it, were largely developed in Europe and Australia. Most countries developed a national accreditation service, working to what is now the ISO/IEC 17025 standard, regional groupings emerged, as did the International Laboratory Accreditation Co-operation, ILAC, with some 46 member bodies. The link between accreditation and NMIs is a strong one. Both share a responsibility for accurate, traceable measurements at a national level.

5˙4. *A National Measurement System.* – As NMIs stopped doing all but the highest accuracy measurements and as accredited laboratories, usually in the commercial sector, took on the more routine tasks, the concept of a national hierarchy of traceable measurements became common place, and was frequently called a "National Measurement System". In general the technical capabilities of the intermediate laboratories are assured by their accreditation to ISO/IEC 17025 by a national accreditation body, usually a member of the International Laboratory Accreditation Co-operation (ILAC). At the top of the traceability system, measurements were relatively few in number and had the lowest uncertainty. Progressing down the traceability chain introduced a greater level of uncertainty of measurement and, generally speaking, a larger number of measurements are involved.

5˙5. *Definitions of traceability.* – Traceability itself also needed to be defined. The draft third edition of the *International Vocabulary of Metrology* (VIM) defines traceability as "*The property of a measurement result relating the result to a stated reference through an* unbroken chain of calibrations *or comparisons each contributing to the stated* uncertainty."

The important emphasis is on uncertainty and the need for the continuous, unbroken chain of measurement. Comparisons of standards or references are a common way of demonstrating confidence in the measurement processes and in the reference standards held either in NMIs or in accredited laboratories. The National Accreditation Body usually takes care of these comparisons at working levels, sometimes called interlaboratory comparisons (ILCs) or proficiency testing.

6. – Regional Metrology Organisations

6˙1. *The birth of RMOs.* – The growth of the number of NMIs and the emergence of world economic groupings such as the Asia-Pacific Economic Cooperation and the European Union means that regional metrological groupings have become a useful way of addressing specific regional needs and as a mutual help or support network. The first was probably in Europe, an organization now named "EUROMET" emerging for a lose collaboration of NMIs based on the "Western European Metrology Club". There are now five "Regional Metrology Organizations": APMP (Asian Pacific Metrology Programme with perhaps the largest geographical coverage from India in the west to New Zealand in the east and extending into Asia); COOMET (the Euro-Asian co-operation in Metrology amongst the Central European Countries), EUROMET (the European co-operation in Measurement Standards), SADCMET (the Southern African Development Community Co-operation in Measurement Traceability), and SIM (Sistema Interamericano de Metrologia or Inter-American Metrology System which covers Southern, Central and North America). The RMOs play a vital role in encouraging coherence in their region and between them: without their help the international metrology system would be far more difficult to administer and its outreach to non-members —who may, however, be members of an RMO— would be more difficult.

6˙2. *RMOs today.* – Traditionally, NMIs have served their own national customers. It is only within the last 10 years that Regional Metrology Organizations have started to become more than informal associations of national laboratories and have begun to develop strategies for mutual dependence and resource sharing, driven by concerns about budgets and the high cost of capital facilities. The sharing of resources is still, however, a relatively small proportion of all collaborations between NMIs, most of which are still at the research level. It is, of course, no coincidence that RMOs are based on economic or trading blocs and that these groupings are increasingly concerned with free trade within and between them.

6˙3. *Equivalence of national measurement standards.* – At the NMI level, the framework of the BIPM and the CIPM's Consultative Committees (CCs) took care of the highest level comparisons. However the increased relevance of traceable measurement to trade, and the need for demonstrable equivalence of the national standards held at NMIs, and to which national measurements were traceable, took a major turn in the mid-1990s. This event was stimulated by the need, from the accreditation community as much as from regulators and trade bodies, to know just how well the NMI standards agreed with each other. Unlike much of the work of the Consultative Committees, this needed to involve NMIs of all states of maturity as well as at all levels of accuracy. The task of comparing each and every standard was too great and too complex for the CC network so a novel approach needed to be adopted. In addition, it became increasingly clear that the important concept was one of measurements which were traceable to the SI through the standards maintained at NMIs, rather than to NMIs themselves. Not to develop and work with this concept ran the risk of creating technical barriers to trade (TBTs) if measurements in a certain country were legally required to be traceable to the NMI standards or if measurements made elsewhere were not recognized. The World Trade Organization was turning its attention towards the need for technical measurements to be accepted worldwide and were setting challenging targets for the reduction of TBTs. The metrology community needed to react.

7. – Mutual recognition of NMI standards: the CIPM MRA

7˙1. *The significance of the CIPM MRA.* – The result was the creation, by the CIPM, of a Mutual Recognition Arrangement (MRA) for the recognition and acceptance of NMI calibration and test certificates. The CIPM MRA is one of the key events of the last few years, and one which may be as significant as the Convention du Mètre itself. The CIPM MRA has a direct impact on the reduction of technical business to trade and to the globalization of world business.

7˙2. *Launch.* – The CIPM MRA was launched at a meeting of NMIs from Member States of the Metre Convention held in Paris on 14 October 1999, at which the directors of the NMIs of thirty-eight Member States of the Convention and representatives of two international organizations became the first signatories.

8. – The essential points of the CIPM MRA

8˙1. *Objectives.* – The objectives of the CIPM MRA are:

- to establish the degree of equivalence of national measurement standards maintained by NMIs;
- to provide for the mutual recognition of calibration and measurement certificates issued by NMIs; and
- thereby to provide governments and other parties with a secure technical foundation for wider agreements related to international trade, commerce and regulatory affairs.

8˙2. *Technical competence.* – The procedure through which an NMI —or any other recognized signatory— joins the MRA is based on the need to demonstrate their technical competence, and to convince other signatories of their performance claims. In essence, these performance claims are the uncertainties associated with the routine calibration services which are offered to customers and which are traceable to the SI. Initial claims — called “calibration and measurement capabilities (CMCs)”— are made by the laboratory concerned. They are first reviewed by technical experts from the local Regional Metrology Organization and, subsequently, by other RMOs. The technical evidence for the CMC claims is generally based on the Institute’s performance in a number of comparisons carried out and managed by the relevant CIPM Consultative Committees (CCs) or by the RMO. This apparently complex arrangement is needed because it would be technically, financially, or organizationally impossible for each participant to compare its own SI standards with all others. The CIPM places particular importance on two types of comparisons:

- International comparisons of measurements, known as CIPM key comparisons and organized by the CCs and which generally involve only those laboratories which perform at the highest level. The subject of a key comparison is chosen carefully by the CC to be representative of the ability of the laboratory to make a range of related measurements.

- Key or supplementary international comparisons of measurements usually organized by the RMOs and which include some of the laboratories which took part in the CIPM comparisons as well as other laboratories from the RMO. RMO Key Comparisons are in the same technical area as the CIPM comparison whereas supplementary comparisons are usually carried out to meet a special regional need.

Using this arrangement, we can establish links between all participants and so provide the technical basis for the comparability of the SI standards at each NMI. Reports of all the comparisons are published in the "Key Comparison Data Base" maintained by the BIPM on its web site.

8˙3. *Comparisons and quality systems.* – These comparisons differ from those traditionally carried out by the CCs, which were largely for scientific reasons and which established the dependence of the SI realizations on the effects which contributed to the uncertainty of the realization. In CIPM and RMO key or supplementary comparisons, however, each participant carries out the measurements without knowing the performance of others until the comparison has been completed. They provide, therefore, an independent assessment of performance. The CIPM however took the view that comparisons are made at a specific moment in time and so required participating NMIs to install a quality system which could help demonstrate confidence in the continued competence of participants in between comparisons. All participants have chosen to use the ISO/IEC 17025 standard or ISO Guide 34 for some chemical measurements and have the option of a third party accreditation by an ILAC member or a self-declaration together with appropriate peer reviews.

8˙4. *The NMIs commitment.* – The outcome of this process is that it gives NMIs the confidence to recognize the results of key and supplementary comparisons as stated in the database and therefore to accept the calibration and measurement capabilities and calibration certificates of other participating NMIs.

8˙5. *Associates of the General Conference.* – When drawing up its MRA, the CIPM was acutely aware that its very existence —and the mutual acceptance of test and calibration certificates between its members— might be seen as a technical barrier to trade in itself. The concept of "Associates" of the CGPM was therefore developed. An Associate has, in general, the right to take part in the CIPM MRA but not benefit from the full range of BIPM services and activities which are restricted to Convention Members. The Associate status is increasingly popular with developing countries as it helps them gain recognition world-wide and does not commit them to the additional expense of Convention Membership which may be less appropriate for them at their stage of development.

8˙6. *The Key Comparison Database (KCDB).* – The key comparison database, referred to in the MRA is available on the BIPM web pages (`www.bipm.org`). The content of the database is already evolving rapidly. "Appendix A" lists signatories, and "Appendix B," the set of key comparisons together with the results of from those that have been completed. It will also contain a list of those old comparisons selected by the Consultative Committees that are to be used on a provisional basis. "Appendix C" contains the calibration and measurement capabilities of the NMIs that have already been declared and reviewed within their own Regional Metrology Organization (RMO) as well as those other RMOs that support the MRA.

9. – Take up of the CIPM MRA

9˙1. *Regulators and the CIPM MRA.* – Whilst the KCDB data is, at the moment, largely of interest to metrologists, it is clear that a number of NMIs are keen to see it taken up more widely by regulators and others. This campaign is at an early stage and at the moment, an EU-US trade agreement cites the CIPM MRA as providing an appropriate technical basis for acceptance of measurements and tests and the USA's NIST has opened up a discussion with the Federal Aviation Agency, FAA, and other regulators as to the way in which they can use the KCDB data to help the FAA accept the results of tests and of certificates which have been issued outside the USA.

9˙2. *A market for NMI calibration services.* – There is a range of additional benefits and consequences of the CIPM MRA. Firstly, anyone can use the KCDB to look for themselves at the validated technical capability of any NMI within the MRA. As a result, they can, with full confidence, choose to use *its* calibration services rather than those of their national laboratory *and* have the results of these services accepted worldwide. They can also use the MRA database to search for NMIs that can satisfy their needs if they are not available nationally. This easy access and the widespread and cheap availability of information may well drive a globalization of the calibration service market and will enable users to choose the supplier that best meets their needs. As the CIPM MRA is implemented it will be a real test of market economics.

9˙3. *Consequences for NMI services.* – Secondly, there is the issue of rapid turnarounds. Companies that have to send their standards away for calibration do not have them available for in-house use. This can lead to costly duplication if continuity of an internal service is essential, or to a tendency to increase the calibration interval if calibrations are expensive. NMIs therefore have to concentrate more and more on reducing turnaround times, or providing better customer information through calibration management systems. Some instrument calibrations will always require reasonable periods of time away from the workplace because of the need for stability or because NMIs only can (through their own resource limitations) provide the service at certain times. This market sensitivity is now fast becoming built into service delivery and is, in some cases, more important to a customer than the actual price of a calibration.

10. – Other metrology bodies

10˙1. *The World Metrology System.* – This short review has not permitted me fully to cover the other bodies which contribute to the "World Metrology System." These include the International Organization for Legal Metrology (OIML) which is concerned with metrology which is included in legislation and which has a role in relation to consumer protection. The OIML and the BIPM collaborate closely and share a common concern for international recognition and mutual acceptance of test and measurement results.

10˙2. *Accreditation and ILAC.* – The International Laboratory Accreditation Cooperation, ILAC, coordinates the work of the regional and national bodies which deal with the accreditation of the technical competence of industrial and other laboratories which provide SI traceable calibration services to the day-to-day user. All laboratories accredited to the international standard ISO/IEC 17025 need to demonstrate traceability to the SI units and quantities realized at the relevant NMI.

10˙3. *The importance of MRAs.* – The BIPM, OIML and ILAC have complementary but common purposes. In support of this, they made a common declaration on the importance of Mutual Recognition Arrangements. This declaration, issued in January 2006, can be found on the BIPM web site.

10˙4. *Measurement Accreditation and Standardization.* – No description, however short, of the international metrology structure can ignore the work of the standardization bodies; the International Standardization Organization, ISO, or the International Electrotechnical Commission, IEC. Conformity with these written standards or specifications invariably involves a measurement. There is, therefore, again a close collaboration between the "Metrology, Accreditation and Standardization" bodies in what has come to be called "MAS."

11. – Closing comments

Structures evolve and emerge. The amazing thing about the Metre Convention is that, as a document created in 1875, it is flexible enough to adapt and to serve today's world. The same can be said for NMIs and the increasing number of other laboratories which support traceability and measurement in national and international structures. Collaboration and Coordination are therefore the key to the success of the integrity and coherence of the World Metrology System and to its ability to serve the needs of industry, science and society.

This page intentionally left blank

Measurements and discoveries: The role of instruments

S. Leschiutta

Politecnico di Torino - Torino, Italy
Istituto Nazionale di Ricerca Metrologica - Torino, Italy

1. – Introduction

Students attending this "Enrico Fermi" summer School on Metrology and Fundamental Constants, deserve a dedicated and careful treatment, for a number of reasons:

- The basic notions of Metrology fall in the realm of knowledge provided in the first years of any education system and radicated in a life-long of every-day use, consequently it can appear trivial or obvious to repeat well-known notions.

- Moreover, most of the students attending to this school are not "plain" researchers with a general degree in Physics, Chemistry or Engineering; they are at the moment working in a National Metrological Institution (NMI), and consequently they are proficient or at least conversant in one or more of the "metrologies".

- About 35 "students" gathered here are now working in a NMI, about other 20 in a research institution, devoted to some physical field, about 10 are coming from Universities; more than half of the total population holds a Ph.D. degree or is working toward a Ph.D. degree.

Fundamentals of Metrology consequently will be not repeated here; but, at any rate, a common background is needed and thanks to the Bureau International des Poids et Measures —BIPM— a copy of the last brochure devoted to the SI system, printed 2006-12-16 is provided to all the attendants to this School [1].

The "students" are warmly invited to read that brochure in order to grasp the contents, and to reach easily, if needed, for consultation, any specific point of the text, or tables with definitions or symbols.

The topics covered in the brochure, along with an introduction devoted to Quantities, Units, Dimensions, Legislation, etc., are the SI Units, Decimal multiples and Submultiples of SI Units and writing instructions.There is also an appended list of some Decisions of the General Conference of Weights and Measures (CGPM) or of the International Committee for Weights and Measures (CIPM) that bear directly upon definitions or units, prefixes defined for use as part of the SI, and conventions for the writing of unit symbols and numbers.

A description of the complete machinery of the SI system, will be given in the lessons presented by A. Wallard during this School.

The acronyms quoted in this paragraph are spelled out in the brochure.

While the fifty-odd lessons that will be presented here in Varenna are all aimed to offer a deep consideration of a very specific and in some cases, narrow topic, it seemed worthwhile to provide also a panorama of the multifarious problems linked to modern Metrology, and consequently it is suggest also the reading of a recent paper [2] covering[1]

- The international System of Units,
- Fundamentals Constants and SI Units,
- Metrological Limits,
- Philosophy of Measurement,
- Coherence of the International System of Units,
- The human factors.

To complete the information about the School, as regards the about twenty-five lecturers or tutors, and with some approximations due to a changing or double affiliation, one-fifth was provided by BIPM, another fifth from Universities or Research Laboratories, and the remaining three-fifth from the National Metrological Institutes.

In total 31 Institutions (NMIs, Laboratories, and Universities) participated in this effort, offering Lecturers or sending Students.

As a final remark, the students are urged to take any opportunity offered by this event that was designed, since his inception, as a mean to foster a strong mutual exchange of knowledge. Experience coming of the previous venues of this School tells that in many cases the two weeks of common study resorted in a long-lasting cooperation and friendship.

[1] The permission of J. Valdès and of Elsevier Inc., for the reproduction of that paper is gratefully acknowledged.

It is well known indeed that the Metrology Community is an intertwined one and the occasion of cooperation are in some cases very frequent.

The reader is also referred to a specific lesson given here in Varenna in the year 2000, when a paper on the laws of physics and on the modern forms assumed by the Theory of Measurements, also outside the realm of Metrology was presented. That paper is fitted with an extended bibliography, divided into topics [3].

2. – Summary and aim of this lecture

Looking now at the text of this lesson, one could gain the impression that is just an essay on the History of Science; past events are indeed presented and recollected, but the final aim is to leave the reader with some tenets:

- In any epoch the progress in the knowledge was resting in, or promoted by measurements.
- In any epoch, measuring instruments were the most accurate expression of the current civilisation.
- The existence of a metric system with his set of devices and rules and traditions was always the landmark of a civilisation, as compared to barter or swapping.
- The basic foundations of Physic, such the principle of equivalence are based on measurements.
- The link between Physics and Engineering is offered by the values of the fundamental constants.
- The links between different "chapters" of Physics is offered by the values of the fundamental constants.
- Each society considered a primary duty the selection and training of experts in the measurement art and the reason we are collected in Varenna is also to obey to this duty.

Sir Francis Bacon (1561-1662) was well aware of the importance of the relations between speculation and experiments. He writes in the "Novum Organum," London 1620, few lines that are illuminating about the role of instruments:

"Neither the naked hand nor the understanding left to it can effect much.

It is by instruments and helps that the work is done, which are as much wanted for the understanding as far the hand.

And as the instruments of the hand either give motion or guide it, so the instrument of the mind applies either suggestions to the understanding or cautions."

3. – Some roles of measurements

Being granted that basic information on Metrology is known, or can be traced in the two suggested sources or gathered in conversations, the aims of this report is to present some considerations about the role of instruments and the mutual relations between measurements and discoveries, old and new instruments will be considered and, in particular the role played by vertical and horizontal pendula or pendula-like devices will be considered in fundamental Physics or in the so-called big unifications. An attempt was made to support these points with a reproduction of the original papers.

In Metrology, Physics and Technology, three lines of thought are considered viable as regards the role of Measurements:

- Measurements are not so important in the finding or discovery of a (new) physical law.
- Measurements and instruments are on the contrary essential:
  a) in the "validation" of a new law or of a new instrument;
  b) in selecting between theories that are, apparently, in conflict;
  c) after the big "unifications", when suddenly it is discovered that two separate chapters of Physics, hitherto considered remote, were just two aspects of the same topic.
- Failures in making measurements can lead to the discovery of new effects or phenomena.

Three examples: fall of a body, energy "inside" a moving mass, equivalence principle will now be considered.

3˙1. *Galileo was suddenly loosing any interest in the measurement or in the instruments he was using or improving, as soon the general pattern of the law appeared in his mind.* – An example was the quest of the law giving the time taken by a falling body; Galileo using an inclined plane with a slope of 5% succeeded in slowing down the fall of a sphere rolling down the slope and started to investigate about the duration of the "fall". Many methods were attempted (counting of drops of water, numbering of the beatings between two musical tones([2]), weighting of sand or water, use of a vertical pendulum), until the "law" was guessed. At this moment Galileo was loosing all his interest in the instrument he was designing or improving.

In this case the measurements were aimed to finding the general ratio between fall and time.

([2]) Galileo was an accomplished liutist.

3·2. *What is the form of the "force" in a moving body with mass? Where is the energy of a body at rest stored? What is the difference between the "dead force" and the force "alive", in other terms what is the quantity of interest: kinetic energy, motion quantity, energy, pulse?* – About the "force" or energy inside a moving body, a scientific debate enraged for more than one century and half, between the mathematicians and physicist around the Seventieth Century.

The names are Galileo (1564-1642), Descartes (1596-1650), Newton (1642-1727), Leibnitz (1646-1716), MacLaurin (1698-1746), Euler (1717-1783). The debate was harsh, sometimes rude, also if polite in appearances([3]).

Two series of measurements were performed:

- The collision in a horizontal plane between spheres of different masses and equal diameter.
- The impact of spheres of equal diameter and different masses impinging vertically with different and known velocities on a lead surface. The diameter of the different imprints was measured. (method of Euler).

The solution of this measurement had also the important consequence in helping the Scientist to fix their ideas about potential energy, kinetic energy, motion quantity, concepts hitherto confused.

3·3. *Very early measurements of the equivalence principle.* – In a pendulum, in the falling of masses, is the nature of the mass relevant? (Vertical pendula were used.)

Measurement of Galileo (1602-1638): Blobs of cork or of lead.

Measurement of Newton (1687): Two pendula of equal dimensions and shape (to eliminate the air resistance) and filled with equal masses of gold, silver, sand, glass, barley .... but preserving the position of the barycentre.

Newton, describing carefully the experiment, states also the accuracy of his measurement: 10^{-3}.

4. – Some roles of instruments, the special case of pendula

4·1. *External conditions fostering the birth of a new instrument.* – When a new instrument appears, at least four points are of order:

1) are the external conditions (the paradigms) —that are needed for the development and the growth of the new device, existing?
2) the potential of innovation inside any new instrument;

([3]) Leibnitz was a master in irony: "*Brevis demonstratio erroris memorabilis Cartesii et aliorum*," that can be translated, at the pleasure of the reader, as "short demonstration of a remakable error of Cartesius and others ...", or "short demonstration of an error of the remarkable Cartesius and others ..."

3) versatility of the new device in the hand of scientist;

4) an unexpected need in an external, sometimes remote, field of Science.

Each of these four points would deserve an elaboration, just to provide examples that are well known; for the first point only —the need of an accepted paradigm— few remarks will be presented.

About the paradigms

1. Paradigms are a set of rules that can help the formation and recognition of a community;
2. the concept and the word were introduced around 1960 by Kuhn;
3. the purpose was to make a model and to study what happens during a "scientific revolution" or the "unifications";
4. the concept has roused much controversy, but it is today accepted, at least as a guideline;
5. a circular definition can be found: a paradigm is a set of rules, traditions, suggested readings, scientific meetings that the members of a given scientific community, but only the members of that community, has in common;
6. the adoption of a common paradigm transforms a group of researchers, that for other reasons are different, in a scientific community;
7. various communities, also similar, live separate: who cross the boundaries is considered with suspicion.

The Members of a specific scientific community, for instance the Metrologists:

1. have common features in their education and training to the research,
2. regard themselves as the chief protagonists of their science,
3. are publishing on the same reviews and attending the same scientific meetings.

The importance of the existence of an accepted paradigm is evident when considering the fate of two fundamental instruments of Science, the barometer and the thermometer they

- were born in the same town, Florence;

- appeared in the same period, circa 1650;

- were designed and used by the same group of people; Borelli, Torricelli, Viviani, the last co-operators of Galileo;

– were produced in the same laboratory by the same glass-blower, nicknamed “il gonfia” [4], but

The *barometer* was used correctly and immediately. The paradigm was provided on the spot by Torricelli: the instrument measures the weight of the air, at the bottom of the ocean formed by air. If we take this device in top of a mountain . . . .

The *thermometer*, no paradigm was available and it took some 250 years, to understand the meaning of his readings.

The need of a previously accepted paradigm as a requisite for a smooth acceptance of a new theory was not always tested, but is a very strong requisite.

Along the cases of the two above-mentioned instruments, other vicissitudes can be offered for consideration: those of Ohm, Joule, Wegener, Maxwell (in some Nations), or in my Country, the cases of Marconi, Majorana[5] and Giorgi.

4˙2. *Why also pendula are to be considered.* – One could rightly ask: why in the year 2006 an expert in Metrology has to bother himself with such an archaic and obsolete device?

Pendulum, in his various forms, was and is still fundamental in the conversion between two forms of energy, is not an archaic device, booth for fundamental Metrology[6] and technical applications; in one of his version, torsion pendulum with the shape of a tuning fork made with a piezoelectric material, the daily production runs to the order of 10^5 pieces per day and the accuracy of the worst device is 10^{-5}. In some luxury cars or heavy duty trucks[7], more than twenty torsion pendula, or piezoresonators, are used for a variety of tasks, as clocks, frequency references in display systems and computers, antiskid sensors, rotation sensors, temperature sensors, electrical connections management, navigation systems, etc.

4˙3. *Metrological application of pendula.* – Application of pendula for scientific measurements, can be divided into three periods:

1. vertical pendula, before about 1650,
2. vertical pendula, 1650–1950,
3. horizontal pendula, 1700–now.

For the three periods here mentioned, the highlights only will be pointed out, using tables arranger in chronological order. (Tables I and II.)

[4] The one who blows.

[5] Quirino Majorana (1871-1957), devoted thirty years of his life, in the quest of a possible shielding of gravitational forces.

[6] See, the papers of Quinn, in this volume.

[7] In some trucks, the management of the electric energy performed using clocks and time-ordered attuators, can resort in a saving of more than 25 kg of copper.

Table I. – *Vertical pendula, before about 1650.*

Application	Researchers	Device	Years
medicine	Galileo Mersenne	pulsilogio fever sensor	1601 1636
time interval	Galileo & Viviani	clock with an escapement	1635
fundamental Physics	Galileo	weak principle of equivalence pendula, blobs of cork and lead	1602 1638
geodesy	Riccioli	clock hand regulated to calculate the Earth diameter	1645
speed of sound	Florence Academy	time interval with a pendulum regulated clock	1645

Table II. – *Vertical pendula, after 1650.*

Application	Researchers	Device and topic	Years
length standard	Wren, Hooke, Huygens,	pendulum	1660 1685
light speed	Roemer	pendulum in a clock + telescope	
time interval	Huygens	pendulum in a clock	1685
metric system	Burattini in Latvia	pendulum + binary scaling	1685
fundamental Physics	Newton	weak principle of equivalence, two equal pendula with blobs of sand, barley, glass, gold, silver, etc.	1687
geodesy & geophysics	Richer Kater	pendulum special pendulum for “small g”	1679 c.1920

Table III. – *Four big Unifications and two difficult measurements made using torsion balances as horizontal pendula.*

			Researcher	Year
force	with	mass	Cavendish	1798
force	with	electricity	Coulomb	1785
force	with	magnetic charge	Coulomb	1787
forces	with	magnetism	Gauss	1838
mass	of	the Earth	Eotvos	1888-1908
adjusted	value	for G	Quinn	now

Table IV. – *Some SI units and related physics research.*

Measurement unit	Definition	Age of definition (years)	Realisation/ reproduction	Related physical research	Researcher and organisation
Second (s)	The second is the duration of 9 192 631 770 periods of the radiation . . .	38	Primary frequency standards producing an electromagnetic radiation at the frequency of a transition accuracy 5×10^{-16}	Atomic precision spectroscopy – H maser Nobel Prize 1964, Nobel Prize 1989 Optical Pumping Nobel Prize 1976 Laser cooling of atoms Nobel Prize 1997	Townes, Basov, Prokhorov, Ramsey, Dehemelt, Paul Harvard Univ. Washington Univ., Bonn Chu, Cohen-Tannoudji, Phillips, Stanford, ENS, NIST
Intensity of electrical current (A)	The ampere is that constant current, which if maintained in two straight parallel conductors of infinite length . . . and placed 1 meter apart . . .	60	Volt and Ohm based upon the Josephson and quantum Hall effects-stability better than a few parts in 10^{-7} conventional values for the Josephson constant K_{J} and the von-Klitzing constant R_{K}	BCS theory of superconductivity Nobel Prize 1974 Tunneling phenomena in solids Nobel Prize 1973 Quantised Hall effect Nobel Prize 1985	Bardeen, Cooper, Schrieffer Harvard, Princeton Esaki, Giaever, Josephson Cambridge Univ. Von Klitzing Univ. Wurzburg
Metre (m)	The meter is the length of the path travelled by light in vacuum during a time interval . . .	23	Frequency-stabilized lasers locked to atomic or molecular resonances	Laser Spectroscopy Nobel Prize 1981 Optical frequency comb Nobel Prize 2005	Bloemberg, Schawlow Univ. Toronto, Bell Labs. Hänsch, Max Plank Institute, Hall JILA-Boulder
Mass kilogram (kg)	It is equal to the mass of the international prototype	105			

4·3.1. Early vertical pendula. Vertical pendula were used as a time reference in the astronomical clocks for near three Centuries; in the period 1895–1920, the pendula of Riefler and Shott had a frequency stability approaching 1×10^{-8}/day.

Seasonal variation on the speed of rotation of the Earth was indeed spotted around 1935 by Stoyko, working at the Bureau International de l'Heure at the Paris Observatory using astronomical clocks, all pendula, and measured by prof. Scheibe, using piezoelectric clocks, at PTR in Germany.

4·3.2. Horizontal pendula and torsion balances, 1700–now. Of particular relevance for the foundations of Physics and today Metrology, are the horizontal pendula, in the particular form of torsion balances.

Many of the big "unifications" of science were indeed demonstrated and afterwards measured, using oscillating horizontal pendula or with forces acting on a horizontal plane and balanced by the torsion of a wire. Similar devices are now used for the quest of an adjusted value for the Newtonian Gravitation Constant G.

This subject will be covered by Terry J. Quinn in his lesson on modern methods for measuring G. Table III, spanning the period 1785–now lists some fundamental measurements made using horizontal pendula.

5. – Conclusions

To foster one of the main aims of this report, as stated in the Summary of this paper, namely the perennial strong relations between Metrology and Fundamental Physics, the occurence of some Nobel Prizes related to Metrology is presented in table IV.

For some fundamental quantities, the current definition is sketched, along with the age of the definition and the method used for its reproduction in the laboratory.

Table IV is self-commenting, but one gains, on the one hand, a impression by the number of the Nobel Prizes bestowed directly or indirectly to metrological topics end, on the other, by the amount and quality of research and experimental activities that was and is devoted to the measurement science.

REFERENCES

[1] *Le Système international d'unités – The International System of Units*, 8th edition 2006, Text in French and English— available also in a concise form. Also available in electronic form `www.bipm.org/en/si_brochure/`.

[2] Valdès J., *Features and Future of the International System of Units (SI). Advances in imaging and electron Physics*, Vol **138** (Instituto Nacional de Technologia Industrial and Universidad Nacional de San Martin, Buenos Aires) 2005.

[3] Quinn T. J., Leschiutta S. and Tavella P (Editors), *Recent Advances in Metrology and Fundamental Constants, Proceedings of the International School of Physics "Enrico Fermi", Course CXLVI* (IOS, Amsterdam, SIF, Bologna) 2001.

History of standard definitions: An outline

S. Leschiutta

Politecnico di Torino - Torino, Italy
Istituto Nazionale di Ricerca Metrologica - Torino, Italy

1. – Introduction

Since the most remote antiquity, the turning point between societies based on hunting or berries-picking and the more evolved forms of land-sedentary tribes is considered the end of barter or swapping and the introduction, based on the voluntary or forced acceptance of

- a metric system with its set of devices, scales and rules, modality of operation and traditions;
- an agreed estimate of the energy needed (or the time to be spent) to produce any item;
- a monetary system;
- the availability of some basic, and in some cases very basic mathematical notions.

It would be interesting to dwell on each of these points; a few remarks only are here presented in the next sections, but the interested reader is referred to some books, quoted in what follows and generally available in Universities, Laboratories or easily found in the book-shops. A bibliography, divided into topics relevant to the aims of the present volume, is appended at the end. A special mention is deserved to the BIPM Brochure, updated every few years and available in two languages, French and English. The last version appeared in 2006; to the brochure a short summary is appended.

In the following sections the different kind of the standards will be classified in five categories or families, distinct following some basic criteria such as:

– the energy needed or to be spent to reach a given goal,

– an anthropomorphic approach,

– the use of a material artifacts,

– the respect of a traditional historical definition,

– the adoption of a definition based on any property of Nature or Physics.

The reader is furthermore warned of the fact that some standards, mostly those of length, mass, time and money are usually considered connatural by the people using them; consequently the standards tend to perpetuate themselves and are very difficult to eradicate from the common use, when time is considered ripe for a change.

There are plenty of examples of this behaviour(1); a few of these striking examples will be presented in what follows.

Additional notes of caution must be added.

The measurement standard or units that were or are used are countless; hopefully the libraries are full of conversion tables.

Moreover, dealing with metric systems, the unique part of science with Mathematics entering in strict and direct contact with the commoners, it is mandatory not to remain inside the more scientific or technical aspects, but always to be well aware of the final user of the Measurement Science, the average citizen.

Any metric system must be indeed *universal* in the sense that the universe of its potential users should accept its use.

Other reasons for this enormous number of standards stem from some facts we can recognize in all times and cultures:

– for a given quantity, *e.g.* length of a ribbon, different material standards and units (not only prices) were used in function of the material (wool, linen, silk);

– different measurement technologies were used in function of the material; in some cases, the standard —a wooden rule— was translated along the ribbon, in other instances the shop-keeper had the right to interpose his thumb between two successive positions of the rule, in a third case the wooden rule, without loosing contact with the ribbon laying on a bank, was rotated in vertical direction before laying down again on the fabric;

(1) In the "twenties", the introduction of the Gregorian Calendar in Bulgaria caused severe upheavals.

- common notion that the unit of surface is the square of the unit of length and similarly for the unit of volume, was not common nor practised at all; the unit of surface was the area of a given rectangle and the unit of a volume was the volume of a parallelepiped([2]), with three different edges;
- average users were accustomed to divide by 2, 3, 6; the Pythagorean table was not known and the peasants had to recur a very strange methods to perform the division, frequently with some approximations or using tables prepared on purpose;
- in the trade hundreds of different "standards" were used for some specific items, such smoked fish or fermented cabbage. The reading of past texts of Metrology, penalties included for any no compliance, is sometimes a very lively exercise.

The examples proposed in what follows, are indeed reflecting the Nationality and the culture of the author, but the reader can be assured that very similar phenomena took place in different Countries. The introduction of the Metric System we are using with satisfaction, at least in the Nations that subscribed the Metre Convention signed in 1875, was not eventless in the first part of the XIX century; even if the system SI is nowadays adopted legally, some non-metric units and standards are still resisting in some Nations and fields.

Similarly, being the author familiar with the time and frequency Metrology, the examples are frequently coming from this realm.

2. – Metric systems

A few families of basic criteria, on which construct a metric system can be traced in the history and are listed at the beginning of the previous section; some of these families are still alive or, with the passing of time, a system migrated from one family to another or some families merged.

A couple of examples are of order. The first one considers the permanence, as regards the standard of mass, the kilogram standard, of one basic criterion. The principle, did not change from the Egyptian or Babylonian times until today; the standard to be used is an artifact, a *given unique stone*, to be used and preserved carefully; obviously the "stone" and the technology changed, but, at least for the last five millennia:

- the basic assumption of a given "stone" to be adopted universally as a common reference has held,
- the measurement device and the measurement technique, the scale, using as a reference the local gravity, remained without change.

[2] In Piedmont, the unit of volume for masonry was a wall, 10 ounces —*oncia*— thick (about 45 cm), one foot (*piede liprando*, the piemontese foot, about 52 cm) long, and height one "*trabucco*", 3.083 m high. The three units of length, *oncia, piede, trabucco*, were neither homogeneous nor bearing decimal relations.

The second example, coming from the length metrology, provides an instance of migration from one family to another, namely from a standard based on the Nature, the quarter of Meridian of the Earth, to an artifact, and more recently another move to a physical constant.

2˙1. *Families of standards and units based on the energy needed to reach a given goal.* – Standards belonging to these families were mostly those used in agriculture.

Some examples: the unit for the surface of the fields was the surface that two oxen were able to plough in one day or the surface that a valid peasant was able to mow in a working day, or, again, the area of ground needed in order to fill a "standard" carriage whit hay. These units had different names in various cultures, but are still widely used today in the transactions and are still quoted in the Notary records.

This family with the assumption of a given "energetic" product, presented interesting variations linked to the muscular physical effort (or time) required to obtain the given quantity of final product.

One of these parameters was the different orientations and consequently fertility of soils in the same area. The valleys of North-West of Italy, generally oriented in the West-East direction, offer the example. The northern side of the valley, looking south receives more Sun, while the opposite side, facing north, is less irradiated and consequently less fruitful. To produce the same quantity of hay, the surface on the side looking North, had to be larger than the one on the opposite side and consequently the surface standard to be used had to be larger.

2˙2. *Families of standards based on anthropomorphism.* – Human body, with ratios between its parts, stable enough and the direct availability and the granted portability of these "standards", inch, foot, palm, etc., provided in the past a widespread use of these "human-like natural" standards for length and, in some specific cases, also for time.

As regards length, the system used in the UK, Canada and other parts of the world is still in widespread use; these standards are well known and not requiring here comments or explanations. Also in "metric" countries for some trades, such as plumbing, the "English" system is regularly used.

As regards time, the dimension and shape of the hand, and the length of the human body shadow were used.

In the second instance, the shadow, measured in feet, of a standing man was proportional to the time interval to or from noon. Mnemonic rules (with a regional validity) were linking the shadow length to the hour. In classic Greece, it was customary to fix the hour of a meeting saying: we shall see when the shadow is 6 feet long (early morning or late evening). Human physiology dictates indeed that for each hour during a day, there is a ratio between the length of the gnomon —the stature of the man— and the length of the shadow, measured using as a length standard the length of a foot of the same man.

The constancy of proportion between the dimensions and the shape of the hand was providing another form of time-telling device: the left hand was held open flat with the palm up and a piece of straw with length equal to that of the index was kept vertically at the corner between thumb and palm. The hand, remaining horizontal, was rotated until the shadow of the muscle below the thumb was coincident with the "line of life", easily recognizable on the palm.

At this moment, the point of the palm in which the shadow of the straw tip was cast, was providing the hour([3]).

In booth cases, a horizontal sundial was realized.

2˙3. *Families based on agreed (or imposed) material artifacts.* – Quite often a material standard was imposed as the legal standard; in some cases the use of the device was imposed by a political authority and the standard itself was following the "ruler" in his migrations; this fact offers explanations when we find similar standards in different or remote regions.

As an example, a standard length of about 52 cm was used in the batimetric chart of Greenland and in the region between France and Italy, the Savoy; a possible explanation lies in the fact that at the Viking period strong were the relations, economical and religious, between Greenland and the Viking European areas, and that in the X century the family of Savoy, coming from the modern Schlewig-Hollstein, (North-West of Germany-Denmark) settled down in the Aosta valley, in the part of the Alps, between Italy, Switzerland and France, South of the Geneva lake. Moreover at that time the length standards used in all the Europe were different, in general shorter or far shorter by about half a meter.

Eventually this standard became the basis of the piedmontese length metrology for one millennium([4]), with the name of "piede liprando".

To evaluate the economical relevance of capacity or volume standards, one must remember that the taxes usually were paid in nature, through given amounts of barley, oat, or corn, to be measured using a standard capacity, a bushel, that was always provided by the ruler.

It is evident that the capacity of that bushel and his ownership . . . was a strong and efficient means of political and economical pressure.

The metrological literature of the past centuries is a continuous flurry of complaints about the capacity of the bushel, the missing periodical calibrations of the internal volume, the modalities of filling([5]), the presence, position and form of any hook, how to decide that the device was full, etc.

Another case of use of artificial standards occurs when the experience proves that a definition, a standard, albeit ideal, cannot be used in every-day life for a number of reasons. For these reasons the metre based on a fraction of the Earth's meridian, was

([3]) Dom Pierre de Sainte Marie Magdaleine: *Traitté d'horlogiographie*, Lyon à la Juste Paix. MDCLXXIV.

([4]) The official limit of validity was fixed for the first January 1850.

([5]) It is evident that the modality of filling, the slow pouring, or the spilling from different distances or heights, the presence inside the vessel of a hook intended to clamp the bushel to a scale, have an influence on the final barley quantity inside the vessel.

substituted with a series of man-made mechanical standards, made in copper and later on in an alloy of platinum-iridium. The same process, led from the mass of a cubic decimetre of water at given temperature, to the kilogram standard.

2˙4. *Families based on some traditional historical definition, usually supported by artifacts.* – In some cases an artificial artifact "realized" and made permanent a historical definition; these families are *di per se* not very important, but provided to two important tasks, the dissemination of a unit inside a kingdom and the development of the units of the last family, those based on Nature.

Two examples can be provided, the so-called "pile of Charlemagne" and the "toise".

The pile was a series of tronco-conic vessels, usually in bronze or in silver, the devices being scaled in order to be fitted each one inside a larger one; the devices were providing in France a scale of masses for the dissemination of the units. Similar solutions were adopted in other nations.

The second example, the *toise*, was considered to be the length of the stretched arms([6]), from the tip of one index to that of the other hand, of Charlemagne, founder of the "Holy Roman Empire". This device, better some replicas, were used as a *transfer standard* on the measurements that led to establish the size and shape of the Earth, in campaigns promoted by the Paris Academy of Sciences and performed in the second half of '700, in Finland and in South America. Those measurements were also fundament for the definition of the meter and in the subsequent years to the construction of the Decimal Metric System.

2˙5. *Families based on any property of Nature or Physics.* – The aims of the Metric Decimal System were described many times and its developments are well understood; usually the scientist were admired by its rationality and conceptual elegance, the fundamental reference for length and mass being taken directly from the Nature, the coherence between dimensions and units provided by the decimal approach and the derivation of masses, volumes, surfaces from a unique length standard, without conversions and coefficients.

One of the obstacles to the diffusion of the Metric System in other Countries was its national origin, despite the recognized rationality and elegance.

Countless were the attempts in Europe to provide a French-like approach for the national metric systems. The most common approach was to use numerology to find a hitherto unknown mathematical relation between the previous national standard and the metre or the metric system([7]). In one case, the unit of mass was assumed equal to 500 grams of water, without any mention to the kilogram or to cubic decimetre.

([6]) In latin, "*brachia tensa*", in French, *toise*, in Italian, *tesa.*

([7]) One striking example was provided by the Turin Academy of Sciences that in 1818 stated that the local "piede liprando" was the "length of 1 degree of meridian taken at the latitude of Turin, divided by 60^3" and the relation between the metre and local "piede liprando" was exactly 3^5 to 5^3.

Anyway in the development of the current Metric System a permanent feature can be recognized, *i.e.* the trend to proceed from a number of fundamental and mutually independent units to a reduced number of units directly linked to the fundamental constants, that are properties of the matter.

Examples were the change in the definition of the "candela", previously involving an instrument, in which it was possible to observe the solidification of melted platinum or the changing in role of the metre, now substituted by the value of the fundamental constant c, the velocity of light.

Another radical change will be, as considered in this Varenna School for 2006, the "death" of the last remaining artifact inside the SI, the kilogram, see the contributions by P. Richard and, R. Davis in these proceedings.

3. – General and essential characteristic to be provided by any metric system

To be universally accepted any metric system must offer to the users a number of essential features. The most important ones are here listed, with a few comments; for further news the interested reader is referred to any text of Metrology.

3˙1. *Universality.* – As anticipated the meaning of *universal* is the fact that the system is accepted and consequently used by the largest number of users.

Two remarks are of order:

- in the physics of some centuries ago, the meaning was quite different. In the Galilean and Newtonian Physics till the times of Mach, the adjective was underscoring the validity of something in the whole Universe, known and to be discovered. In this context, the concepts of synchronisation or of a common time scale for events occurring anywhere and anytime, was not controversial and was accepted as a fundamental one.
- In today Metrology, the meaning is far more modest and refers only to the acceptance by many classes of users. This behaviour, however, is sometime difficult to be reached.

An example of such difficulties is provided by the existing divergence as regards the time and frequency standards and definitions, between two classes of very important users in science and technology:

- on the one hand, physicist, radio technologists, some radio astronomers and communications engineers, are interested in the frequency, *i.e.*, on the length of the second and in its stability and accuracy, and not really interested on the epoch (the date) for that second;
- on the other hand, astronomers, surveyors, navigators, space engineers, some other radio astronomers, timekeepers are basically interested in the epoch, the date, the hour; for these categories the time is basically an angle or a difference in longitude.

To accommodate both classes of users with the same unit, the second SI, and with the same time signal, given by unique radio time signals, special arrangements had to be introduced as the adoption of the so-called "leap second"([8]), a second to be periodically removed from the time scales, obtained from stable atomic clocks, in order to reduce the discrepancy between these time scales and the time scales linked to the rotation of the Earth.

The Earth rotation is indeed slowing down for the tidal friction (about one second per year or 3×10^{-8}, as a fractional frequency difference), but the second, as the time interval unit, must remain constant and exact as far is possible (the best atomic clocks have a frequency accuracy of some 10^{-16} and a stability of about 10^{-16}/day).

3˙2. *Uniformity.* – Uniformity means that the scale of that unity proceeds in uniform and regular way, *i.e.* the intervals are all equal, the readings are the result of an integration and the marks are following the order of natural numbers: only if the scale is uniform, an interval can be obtained by subtraction of two readings.

This characteristic is essential in time metrology, geophysics, physics, financial portfolio management, space navigation, etc. As an easy example, consider the age in days on an individual or of any operation on a bank account. The calculation in not difficult, but intricate and prone to be erroneous, with months of 28, 29, 30, 31 days and leap years. The current civil calendar is indeed not a uniform scale.

To cope with these difficulties, astronomers, geophysicists, space navigators, bankers, supermarkets, managers, etc., are using other calendars based on a *decimal count of days*, introduced([9]) in 1624 by an Italian physician, Giulio Cesare Scaligero, and called calendars in "Julian days".

Two of those calendars are in current use:

- Julian days (JD), with origin([10]) at noon of the day 1st January 4713 before Christ,
- Modified Julian day (MJD) with a more convenient origin, moved to a day in the XIX century, to avoid the use of large figures.

To convert one date in JD, to a date in MJD, it is sufficient to subtract the fixed number 2400000.5, the days elapsed between the two conventional origins; following this rule, the beginning of a day is moved to midnight([11]). MJD is currently used in Time Metrology, Geodesy. Geophysics, Space navigation, etc.

([8]) For details see any reference text in time and frequency, as those listed in the bibliography

([9]) Julius Caesar Scaligerus, *De emendatione temporum*, Geneva, 1634.

([10]) The choice of this remote origin was cleverly made by Scaligero, taking into account four different timing systems used in the past, such as the Methon cycle (the periodicity of lunar eclipses), the Olympiads cycle, the "roman indiction", *i.e.* the time intervals after which the tributes were revised, in the Roman Kingdom, Republic and Empire, for a time span of seven centuries, and the date of Easter.

([11]) To the day 11 January 2007 there corresponds the day 54111MJD.

One of the components forming the Julian calendar, the “indizione romana”([12]), a period of 15 years, designed for the real estate taxation, was used for near fifteen centuries and is unique in providing, with some additional information an accurate dating of events in the time span from the III century after Christ to the XVIII century.

3˙3. *Perennity.* – The definition of the unit and consequently of the standard should be perennial; the definition or the device should be stable, well preserved, and carefully used([13]); see in this book the chapters devoted to the mass metrology.

One must be aware that in some cases, such as the mass, the prototype is unique([14]),([15]). An international Committee, the CIPM, (Comité International Poids et Mésures) formed by 18 members, checks once per year how the prototype is conserved at the BIPM, the Bureau International Poids et Mesures. For the organisation of these Bureau and Committee, see the chapters prepared by A. Wallard in this volume.

When perennity is not granted, special cares must be devised to assure the continuity of the dimension (the numerical value) of the standard.

A special case is provided by the construction of the Atomic Time scale TAI—*Temps Atomique Internationale.*

TAI stems by an atomic treatment and dedicated algorithms of the readings of atomic clocks, kept in different laboratories. An atomic clock is a manmade device and can indeed stop or become subject to any calamity. Some news concerning the reliability of atomic clocks used for the formation of the International Time Scale, TAI can be found in the pubblications of the International Telecomunications Union, quoted in the bibliography. By the way, the parameter called Medium Time Between Failures, can reach 25000 hours of continuous service for the atomic clocks used for the formation of TAI and 150000 hours for a piezoelectric clock.

Consequently some redundancy is needed and very accurate means must be available to transfer time between remote laboratories. The reader interested in these topics, is referred to the chapter by F. Arias in this volume.

3˙4. *Accuracy, precision, stability, independence from influence quantities.* – The quoted four characteristics, are just various aspects of the “quality” of a standard.

One initial and important remark: the accuracy of a (written) definition is infinite; one can start to consider the other components —the accuracy of a physical standard

([12]) For the period of one “indizione”, the cadastal taxes were not revised. This computing was characteristic of the Roman Empire, bur survived, at least in the notary activities until the Napoleonic period, and in the Roman church calendar also today.

([13]) The prototype of the kilogram is touched, with a number of precautions, by a human hand, every 40 years, or so.

([14]) In the crime stories or in the fiction novels, a number of entries are dealing with the unique prototype lost, stolen or subject to a ransom. Some unchecked, but circulating stories, regards possible events or adventures of the prototype during the past World War Two.

([15]) This information is based on the present definition of the mass unit; a new definition is underway, as presented elsewhere in this volume.

Table I. – *Trends of the frequency and timing accuracies and their applications. ARGOS: Conic localization System. (From orbiting satellites in polar orbit.) OMEGA: VLF Hyperbolic Navigation system on miriametric waves (VLF- 3–30 KHz). LORAN: LOng RANge Navigator, LF Hyperbolic navigation, in LF band 30–300 kHz. Early LORAN A, accuracy: 10^{-7}–10^{-8}, the most accurate is LORAN-C: 10^{-11}–10^{-12}. GPS: Global Positioning System – GLONASS Global Navigation Satellite System. VLBI: Very Large Base Interferometry, stability 10^{-13}/hour. TRANSIT - TSIKADA: satellite-based hyperbolic navigation Systems, as ARGOS. Disciplined quartz, a piezoelectric frequency standard in which the ageing is removed, using an external reference. For details, see again the pubblications of the International Telecomunications Union, quoted in the bibliography.*

Years	Applications	Accuracy or ageing	Frequencies Standards
1940	Carrier of HF transmitters- radars LORAN A	10^{-5}–10^{-6}	Quartz crystals
1950	Radars – VLBI - DECCA	10^{-6}–10^{-7}	Quartz crystals
1960	LORAN C – radar (interplanetary)	10^{-7}–10^{-8}	Caesium frequency
1970	OMEGA VLBI	10^{-10}–10^{-11}	Maser H for VLBI
1980	Carriers of some TV transmitters early digital communications	10^{-9}	Atomic or disciplined quartz
1990	Conic and Hyperbolic satellite navigation (TRANSIT - TSIKADA – ARGOS)	10^{-9}–10^{-10} 10^{-10}	Disciplined quartz or atomic- rubidium
2000	Circular Satellite navigation (GPS), fundamental metrology, relativity corrections	10^{-12}–10^{-14}	Atomic standards

included— when the definition is transformed in a physical device to be really used in a laboratory, subject to given environmental conditions and to perform a given task. What remains is to express a judgement on the discrepancy between the ideal definition (no error) and the physical realization of that standard.

This discrepancy is conveniently expressed in relative (or fractional) terms; also the sensitivity to any quantity of influence is expressed in relative terms, for instance the frequency sensitivity of a frequency standard to ambient temperature is given as $\Delta f/f/°\mathrm{C}$. Same notation is used for the term stability, that is usually ambiguous, if it is not provided with the statement of the physical quantity concerned.

The statement of the discrepancy between a written definition and its embodiment is a difficult and demanding task, in particular when the aim is to construct a primary standard: there exists no "super-primary" standard to be used as reference and in principle all the realizations of a definition, have the same dignity. This particular case is to be left to the intellectual honesty of the scientist-experimenter.

The needed accuracy or uncertainty[16] to be reached depends on the applications and the status of technology; table I gives an idea of the trend for the frequency and timing accuracies in the period 1940-2000, along some typical applications. In some cases, in lieu of the timing requirements in absolute units, it is given the frequency accuracy or the stability of the clock needed. It is customary to call as ageing, the daily frequency stability versus time, given as $\Delta f/f/d$,

All the figures presented in table I, can be discussed and a justification can be found in the given reference, but the attention of the reader is called to three facts:

- in one metrology, that of frequency and time, in about half a century the requirements as regards frequency accuracy or stability augmented of about one order of magnitude every ten years;
- this trend is observed in frequency standard and in the accompanying instrumentation and comparison and transmission methods;
- any progress in this metrology is immediately giving place to more demanding applications.

Concerning finally the present situation of the atomic frequency standards and their more scientific application, *i.e.* the construction of the international Atomic Time scale TAI, the reader is referred to the chapters prepared respectively by Bauch and Arias and appearing elsewhere in this book.

The last remark to be presented regards "the quality" of the "*written*" definition of a fundamental standard.

This definition should present two characteristics:

- to be well free from ambiguities but as simple as possible,
- to be "wide" enough, without unduly specifications, in order to leave free space to the ingenuity of the researchers that must not be bounded to any existing technology, valid at the moment of writing the definition, but not necessarily permanent in the future.

The present definition of the SI *second*, approaches these two requirements in the sense that it was written about half a century ago, and, as it was seen in the table, the accuracy of the devices embodying the definition, from about 10^{-9} of the Essen caesium beam standard of June 1955, is reaching 10^{-16} in 2005 and was apt to take substantial advantages from scientific discoveries, such the "cold" atoms, hidden in the future at the moment in which the definition was drafted. In other terms, the definition was such to permit, using the same concepts and the same words, an improvement by seven orders of magnitude.

[16] For the meanings of these terms, see the contributions by W. Bich in this volume.

It is somehow ironic that the "fountain" of caesium atoms, now constructing the second SI, was proposed and tested by Zacharias half a century ago: in between four gentlemen were awarded with the Nobel Prize and astonishing improvement in electronics and experimental Physics such as laser cooling and trapping, had to be developed. The reader interested in the history of the progress of the caesium beam standard, can find all the facts on the June 2005 issue of Metrologies, edited by Terry Quinn FRS, the past BIPM Director.

* * *

It is indeed a pleasant duty for the author to recognise the advice and help offered by Dr. Maria Luisa Rastello, the Scientific Secretary for this Varenna School "Recent Advances in Metrology and Fundamental Constants".

Bibliography

General topics: Units, quantities, coherence inside the Systems, nomenclature, fundamental Constants, the Science of Measurements.

- *Le Système international d'unités – The International System of Units SI*, BIPM, Pavillon de Breteuil, F-92312 Sèvres, Cedex, France.

- Quinn T. J., Leschiutta S. and Tavella P., *Recent advances in Metrology and Fundamental Constants, Proceedings of the International School of Physics "Enrico Fermi", 2000*, Course CXLVI (SIF, Bologna, IOS Press, Amsterdam) 2001.

- ISO – International Standards Organization, *Quantities and Units* in *The ISO Standards Handbook*, 3rd edition, 1993.

- *Quantities, Units and Symbols in Physical Chemistry*—the IUPAC Green Book (Blackwell Science) 2nd edition, 1993.

- *The International Vocabulary of Metrology* (usually called "the VIM", from the French name *Vocabulaire Internationale de Metrologie*) (International Standard Organization) 2nd edition, 1993.

- Denis-Papin M., *Metrologie Génèrale; grandeurs, unites et symbols* (Dunod Editeur, Paris) 1960.

- Danlous-Dumesnil N., *Etude critique du sistème métrique* (GauthierVillars, Paris) 1962.

- Sena L. A., *Units of Physical Quantities and Their Dimensions* (MIR Publishers, Moscow) 1972. This small book, was written in Russian, and translated into English, it is remarkable for his clarity.

- Kula W., *Measurements and Humanity.* This book, written in Polish, is available also in French, English and in Italian (*Le misure e gli uomini dalla antichità ad oggi* (Rome) 1987), it is really impressive for the wealth of information, mostly regarding France and Poland, in particular for the use of bushels in the countryside and related controversies. The book is somehow undermined by the author's approach to give a sociological interpretation to any fact.

- Petley B. W., *Fundamental Physical Constants and the Frontiers of Knowledge* (Adam Hilger Ltd, Bristol) 1985.

- Wise Norton M., *The values of Precision* (Princeton University Press, Princeton) 1995.

- Klein H. A., *The Science of Measurement* (Dover Publications, New York) 1988.

- Sydenham P. H., *Measuring Instruments: Tools of Knowledge and Control* (Sydenham, Stevenhage) 1979.

- Testut C., *Mémento du pesage* (Hermann & C. Edit., Paris) 1946.

Note: The specific theme of the Coherence inside the Units is well covered by Mills I. M. : "Physical Quantities and Units" in *Recent advances in Metrology and Fundamental Constants* (Course CXLVI).

Note: Pertinent observations about the measurement Theory can be found in a number of papers by L. Finkelstein, that were instrumental for a general awareness about the measurement theory, previously confined to some restricted circles. A list of these papers can be found in a rewiev by Leschiutta S. : "Metrology and the Logic of Physics" in *Recent advances in Metrology and Fundamental Constants* (Course CXLVI).

Additional Reference to Table I

- International Telecommunications Union: *Handbook: Selection and use of Precise Frequency and Time Systems* (ITU-Radiocommunication Bureau, Geneva) 1997.

History of Metrology

- Jedrzejewski F., *Histoire universelle de la mesure* (Editions Ellipses, Paris) 2002. Very well documented and of agreeable reading.

- Berriman G., *Historical Metrology* (Dent, London) 1936. This booklet is out of print, but, if found in a library, is a real mine of facts and data.

- Hocquet J. C., *La Métrologie Historique* (Presses Universitaires de France, Paris) 1995.

- Moreau H., *Le Système Mètrique* (Chiron, Paris) 1975.

– Witthoft H. *et al.*, *Die Historische Metrologie in dem Wissenschaften* (Scripta Mercaturae Verlag, St. Katharinen) 1986.

– Machabey A., *Histoire des Poids et Mésures dépuis le* 13° *Siècle* (Imprimerie de Troyes, Troyes) 1962.

Physical quantities

T. Quinn

Bureau International des Poids et Mesures, Pavillon de Breteuil, F-92312 Sèvres Cedex, France

1. – Introduction

The International System of Units (SI) is the universal language of science. It allows measurement results and predictions of theories to be compared worldwide and over different epochs. It also allows international trade to be carried out with goods and services being measured on a common basis. The SI was formally established only in 1960 but it was the culmination of nearly one hundred years of discussion on how best to build a system of units suitable for all areas of science and everyday life. Underlying all the discussions on units there was at the same time a parallel discussion on the meaning of and the best way to treat the concept of physical quantity. While this is the main topic of these lectures, it is impossible to treat physical quantities without at least mentioning the units in which they are measured, but no systematic account is given here of the SI.

In one sense, the details of the meaning and the correct use of quantities and units in science are comparable to the details of the grammar and usage in a language. For those whose mother tongue it is, correct grammatical use is absorbed from parents, friends and relations while very young and it is tempting to say that this is sufficient. Indeed, in some countries the teaching of grammar in schools has almost disappeared following the doctrine that the language is what people speak and it should not be constrained. While this is not the occasion to go into this, the result is a significant loss in the precision of meaning and in clarity of expression in human discourse. In science, lack of precision in meaning and in clarity of expression is not acceptable. As is the case for a language, however, many of the basic concepts of quantities and units and how they are used come very early in life and become part of the body of scientific knowledge that we all absorb. Nevertheless, great efforts are now made to draw up clear rules for the definitions and use of quantities, units and their symbols to promote precision and clarity in scientific discourse.

In these lectures entitled “Physical quantities”, I shall begin with a brief outline of the history of ideas concerning physical quantities and then present the current understanding and usage and indicate the essential sources of up-to-date information. Although physical quantities and units are often written and talked about and handbooks and websites produced, there are relatively few, what I would call, modern authoritative sources in the English language where the subject is treated in depth. Among these I would include *Units and Dimensions* by E. A. Guggenheim, 1942 [1]; *On the History of Quantity Calculus and the International System* by J. de Boer, 1995 [2] and *Physical Quantities and Units* by I. Mills, 2000 [3]. The last of these is closely modelled on, but bringing up to date, the article by Guggenheim and was the basis of the lectures given by I. Mills at the previous Varenna Metrology Summer School. The present text draws heavily on the articles by J. de Boer and I. Mills both of whom presented very clear expositions of the subject. Pertinent texts also are the IUPAC Green Book *Quantities Units and Symbols in Physical Chemistry*, 2nd edition, 1993 [4], the ISO Standards Handbook *Quantities and Units*, 1993 [5], and the *International Vocabulary of General and Basic Terms in Metrology*, 3rd edition, 2006 [6] (in press at the time of writing these lectures). In the particular context of the Varenna Summer School, I refer also to the two lectures on recent developments in the SI, the first given by me at the 2000 School and the second given here by Andrew Wallard.

As regards formal texts published by official bodies, the *SI Brochure* [7] published by the BIPM is the official SI handbook and gives a complete description of the International System of Units and is now in its 8th, 2006, edition.

In these lectures, the names of a number of international organizations and their specialized committees are mentioned, principal among these are the following:

The International Bureau of Weights and Measures, BIPM. The BIPM was created by the Metre Convention, in 1875. It operates under the supervision of the International Committee for Weights and Measures CIPM, which itself comes under the authority of the General Conference on Weights and Measures, CGPM, consisting of delegations from the Governments of the Member states of the Convention. The CIPM has created a number, at present ten, of Consultative Committees to advise it on matters related to metrology. The Consultative Committee for Units, CCU, is responsible for drawing up the text of the SI Brochure. The ensemble of the these bodies make up the Intergovernmental Organization of the Metre Convention.

The International Organization for Standardization ISO and the International Electrotechnical Commission IEC. These are the two international (non-governmental) organizations responsible for drawing up international standards (specifications, *norms* in French). Of particular interest in the present context are the ISO and IEC Technical Committees ISO/TC12 and IEC/TC 25 which are responsible for ISO/IEC Standard 31 on quantities and units. This standard is shortly to be superseded by a new joint standard ISO/IEC 80000, *Quantities and Units.* Both ISO and IEC are much involved with terminology and also with standards and guides on the expression of uncertainties and with statistics (see the JCGM below).

The International Organization for Legal metrology OIML is the other intergovernmental body whose principal task is related to metrology. It was created in 1955 and is concerned with international specifications concerning measuring instruments used in those areas of metrology subject to legal requirements.

The Joint Committee for Guides on Metrology JCGM is an *ad hoc* group of organizations that have agreed to cooperate in drawing up international documents related to metrology. It is made up of the three organizations mentioned above plus the International Laboratory Accreditation cooperation, ILAC; the International Federation for Clinical Chemistry and Laboratory Medicine, IFCC; the International Union of Pure and Applied Physics, IUPAP and the International Union for Pure and Applied Chemistry, IUPAC. The two publications of the JCGM at present are the International Vocabulary of Basic and General Terms in Metrology, the VIM and the Guide to the Expression of Uncertainty in Measurement, the GUM.

2. – History

Mathematics and the manipulation of numbers goes back to the time of the Babylonians but it was not until the third century AD that the Greek mathematician Diophantus of Alexandria first made use of written symbols to represent unknown numbers. In the 9th century, the systematic use of symbols, as opposed to the numbers, is due to the Arab mathematician Al-Khwarizimi who also demonstrated the solutions for a number of different types of quadratic equation. As is well known, the word algebra came from the title of his famous book. Much later, Francois Viète in the 16th century and René Descartes in the 17th century made important advances in the use of symbols in algebra. None if this, however, was related to the symbolic use of what we would call today physical quantities, indeed the term concept of physical quantity itself did not appear until late in the 19th century at the hand of James Clerk Maxwell.

Notwithstanding Newton's magnum opus *Philosophiae Naturalis Principia Mathematica (Mathematical Principles of Natural Philosophy)*, the mathematization of the physical sciences really got underway at the beginning of the 19th century. Progressively during this century a number of important advances in mathematics were made with respect to the application of algebraic operations to objects other than pure numbers. The interpretation of the operations of addition and multiplication was generalized and applied to entirely different objects. De Moivre and Euler and later Gauss in 1831 and the Irish mathematician Hamilton in 1833 applied addition and multiplication to complex numbers. Then followed their application by Cayley and Sylvester to matrices and then to linear substitutions and permutations by Abel, Galois and Cauchy in the 1840s. This led to the calculus of abstract groups where multiplication and division were applied to elements of a group thus giving these operations a new interpretation. In 1850, Boole extended the use of elementary mathematics to the theory of logic by applying the addition and multiplication operations to classes or sets and to logical propositions. This was the start of mathematical logic and later to the theory of sets by Cantor and Schroeder at the end of the century.

The modern term “quantity calculus” refers to the application of mathematical operations to symbols representing physical quantities. As has been remarked by Mills, however, the term “algebra of quantities” or “quantity algebra” would in fact better describe what is actually done, but it is too late to change the term now. The starting point for any discussion on the meaning of quantity calculus is the meaning of the term physical quantity itself. In the past much has been written about this and it goes to the very heart of the epistemology, *i.e.*, our basic theory of knowledge, of physics. Maxwell introduced the concept of physical quantity in his 1873 *Treatise on Electricity and Magnetism* [8] in the following terms:

“Every expression of a quantity consists of two factors or components. One of these is the name of a certain known quantity of the same kind as the quantity to be expressed, which is taken as a standard of reference. The other component is the number of times the standard is to be taken in order to make up the required quantity.

The standard quantity is technically called the Unit and the number is the Numerical Value of the quantity.”

Thus

$$\text{physical quantity} = \text{numerical value } x \text{ unit},$$

which is expressed symbolically as

$$Q = \{Q\} \cdot [Q]. \tag{1}$$

Note that in writing symbols for physical quantities italic characters are used.

Maxwell clearly states that the unit is also a quantity and he uses the descriptive term “a quantity of the same kind as the quantity to be expressed”. It is this that allows the comparison in a quantitative way of other quantities of the same kind and makes the measurement of physical quantities possible. We come back later to the meaning of the term quantity of the same kind. The unit is seen as a reference or a standard and can be interpreted as a material object using the word standard, in the French sense *etalon.* Maxwell does not attempt to define his concept physical quantity other than by giving a description of how it is used. This is consistent with the modern view that the concept of a physical quantity is a basic concept, sometimes termed a “primitive”, which cannot be defined in terms of something else but the use of which is explained by the rules of quantity calculus.

The great advantage of expressing the results of physics in terms of physical quantities is that it gives a representation that does not depend upon the choice of units. If a particular length L is expressed in two different units $[L]'$ and $[L]''$ by the two expressions $L = \{L\}' \cdot [L]'$ and $L = \{L\}'' \cdot [L]''$, the numerical values satisfy the “inverse proportionality rule”

$$\{L\}'/\{L\}'' = [L]'/[L]' \,, \tag{2}$$

but the physical quantity L is itself invariant. This is the essential reason why the use of physical quantities and not numerical values is to be preferred in theoretical description of physical phenomena. Using physical quantities gives a representation which is invariant with respect to the choice of units.

In 1887, Helmholtz [9] published the results of an investigation into the meaning of the elementary operations of the mathematical formalism of physics. He had been greatly inspired by the earlier mathematical investigations of H. Grassmann (1861) on mathematical operations with various objects in geometry. Helmholtz identified Maxwell's concept "numerical value of a physical quantity" as "a concrete number" in German "*benannte Zahl*". He wrote: *Objects or attributes of objects, which —compared with others of the same kind— permit the distinctions "larger", "equal" or "smaller" are called quantities. If these can be expressed by a concrete number then we call this the "value of the quantity".* Multiplication of quantities with numbers and addition of quantities of the same kind were related to and largely explained on the basis of physical combination or concatenation of physical objects corresponding to quantities of the same kind. Thus for Helmholtz the possibility of direct empirical measurement was considered to be an essential property of a physical property. Helmholtz and the mathematician Hölder [10] are usually seen as initiators of the axiomatic treatment of what is often called the theory of measurement. In the development of quantity calculus, German physicists and mathematicians have played an important part and I mention in this respect particularly two books, the first by J. Wollot, *Grössengleichungen, Einheiten und Dimensionen* (1957) [11] and U. Stille, *Messen und Rechnen in de Physik* (1955) [12]. Unfortunately neither have ever been translated into English.

Maxwell's generalization of the application of the usual mathematical operations to such objects as physical quantities was new and was by no means accepted by all scientists at the time —or even later. Many took the view that the only part of the value of a physical quantity that it made any sense to operate on as a mathematical quantity was the numerical value part. One of the strong proponents of Maxwell's ideas, however, was the English mathematician Alfred Lodge who wrote in 1888 [13] "*The equations in mechanics and physics express relations between quantities and are independent of the mode of measurement of such quantities, much the same as one can say that two lengths are equal without enquiring whether these are going to be measured in feet or metres.*" He was very active in promoting the use of physical quantities as opposed to numerical values and, in engineering, quantity calculus rapidly became used and known under the name of the Stroud system. The advance of quantity calculus in physics did not, however, advance rapidly and during the first quarter of the twentieth century some important physicists and philosophers were strongly against it. Notable among these were N. Campbell in his book *Physics: The Elements* (1920) [14] and Bridgman in his book *Dimensional Analysis* 1922 [15]. They both expressed opposition to the interpretation of the symbols of mathematical physics other than just as numbers or numerical values. Bridgman in considering unit conversion admits, however, to the tacit requirement of the validity of the inverse proportionality rule for concrete units. He considered the common notation

$$88\,\mathrm{ft/s} = 88 \times (1/5280)\,\mathrm{mile}/(1/3600)\,\mathrm{h} = 60\,\mathrm{mile/hour} \tag{3}$$

as a kind of shorthand and wrote "*In treating the dimensional formula in this way we have endowed it with a certain substantiality, substituting for the dimensional symbol of the fundamental unit the name of the concrete unit employed and then replacing this concrete unit by another to which it is physically equivalent* (*e.g.*, 1 ft = (1/5280) mile). *That is we have treated the dimensional formula as if it expressed operations actually performed on physical entities, as if we took a certain number of feet and divided them by a certain numbers of seconds. Of course we actually do nothing of the sort. It is meaningless to talk of dividing a length by a time: what we actually do is to operate with numbers which are the measures of these quantities.*" He then continues "*We may however use the shorthand method of statement if we like with great advantage in treating problems of this sort but we must not think we are actually operating with physical things in other than a symbolic way*" but this of course opened the way to the development of the calculus with these symbolic quantities. The crucial step was to accept this "shorthand" notation, regardless of one's philosophical hesitations as to its meaning, as a valid symbolic notation for expressing quantities in terms of units in the physical sciences and to develop rules for the algebraic use in the formalism of theoretical physics. The subject was then further advanced and formalized by Guggenheim [1] and Wallot [11].

From the end of the nineteenth century right up until the adoption of the SI by the 10th CGPM in 1960, the question of the definition of electrical quantities and how best to measure them from both the theoretical and practical points of view were much debated. Already in 1863 Maxwell said "the phenomena by which electricity is known to us are mechanical in origin and therefore they must be measured by mechanical units or standards." The problem was to find appropriate electrical quantities and relate them to mechanical quantities in a convenient way. I present a brief outline of the development of ideas on electrical measurements in these lectures without attempting to give an exhaustive account since this would take up too much time.

While it was formally stated by Maxwell that a unit is no more than a particular example of a quantity of the same kind as the one to be expressed, the development of ideas and the practice related to units followed a path rather separate from that of concepts related to physical quantities. This was because the establishment of appropriate units for human activities long predated any ideas on the meaning of the concepts of physical quantity and was in fact for most of human history a mundane practical matter. Indeed, even today, very little of the activities of the national metrology institutes are related to the formal algebraic structure of quantity calculus, instead, it concerns the practical establishment of units and their dissemination whether they be related to mass standards or the quantum-Hall effect.

The most recent formal discussion of quantity calculus that I am aware of is in the article already cited by J. de Boer in 1995. In this he lays out a sketch of the formal algebraic structure of quantity calculus in terms of group and set theory.

One final point in the history of the subject concerns nomenclature and vocabulary. Since 1984 there has existed an International Vocabulary of Basic and General Terms in Metrology, universally known as the VIM [3]. This was originally drawn up jointly by the four international organizations having a particular interest in metrology, namely the

BIPM, the OIML, the ISO and the IEC that created the JCGM (see above). The third edition is now in press and it differs significantly from the first two in that the formalism of international standards on systems of nomenclature and vocabulary are much more evident. I shall say a few words about the new edition of the VIM and draw to your attention the main points related to the modern ideas concerning the vocabulary to be used in discussing quantities and units.

3. – Quantities and quantity equations

3˙1. *Definitions of quantities.* – The mathematical description of the natural world is in terms of equations between symbols representing physical quantities. Although much debated in the past, this understanding of the equations of physics is now accepted by the great majority of scientists. The term "physical quantity" is intended to include quantities in all areas of science. The present definition of physical quantity from the VIM is the following: *a property of a phenomenon, body or substance, to which a number can be assigned with respect to a reference.* This is perfectly consistent, of course, with the relation $Q = \{Q\} \cdot [Q]$.

A distinction exists between what are called quantities in a general sense, such as mass, length and time, and quantities in a particular sense such as the mass of this piece of brass, the length of this table or the time taken to fly from Paris to Turin. While in a formal sense this is an important distinction, and is strongly reflected in the new edition of the VIM, in practice it is so obvious that it is not necessary to be explicit. For example, we can write the relation between force, F, mass, m, and acceleration, a, as

$$F = m \cdot a \tag{4}$$

without necessarily having to state explicitly that the force F is the force that is applied to this particular mass m and that the acceleration a is the acceleration of this same mass. It would be a perverse interpretation of the equation to understand it to mean that the acceleration a was not that of the mass m that was subject to the force F! In order to make the meaning explicit, however, we would have to write

$$F_j = m_j \cdot a_j\,, \tag{5}$$

where F_j is the force applied to a mass m_j leading to an acceleration a_j of that same mass. In general, therefore, the equations of physics are equations between particular quantities although this has rarely to be stated explicitly. Particular quantities are also referred to as instances of quantities in the general sense.

By convention, a set of quantities has been chosen that are known as base quantities, these are shown below in table I.

Derived quantities are all other quantities that are expressed as products of powers of the base quantities. The set of derived quantities can be extended without limit. Base quantities are conventionally considered to be mutually independent, since one base

Table I. – *Base quantities, their dimensions and units in the International System.*

Base quantity	Symbol	Dimension	Base unit	Symbol
length	l, x	L	metre	m
mass	m	M	kilogram	kg
time	t	T	second	s
electric current	I	I	ampere	A
thermodynamic temperature	T	Θ	kelvin	K
amount of substance	n	N	mole	mol
luminous intensity	I_v	J	candela	cd

quantity cannot be expressed by a product of the powers of the other base quantities. Each base quantity also has, by convention, its own independent dimension, shown also in table I. The dimension of a derived quantity, $\dim Q$ is given in terms of the dimensions of the base quantities by the product

$$\dim Q = \mathsf{L}^{\alpha}\mathsf{M}^{\beta}\mathsf{T}^{\gamma}\mathsf{I}^{\delta}\mathsf{O}^{\varepsilon}\mathsf{N}^{\xi}\mathsf{J}^{\eta}\,, \tag{6}$$

where the exponents, indicated by Greek letters, are generally small integers which may be positive, negative or zero and are known as the dimensional exponents. Table II gives some examples.

One can check the dimensional consistency of a quantity equation by verifying that both sides have the same dimension.

There are some derived quantities for which the defining equation is such that all the dimensional exponents in the expression for the dimension are zero. This is true in particular for any quantity which is defined as the ratio of two quantities of the same kind. Such quantities are described as being dimensionless or of dimension one.

Table II. – *Some examples of derived quantities and their dimensions.*

Quantity	Dimension
velocity	LT^{-1}
angular velocity	T^{-1}
force	LMT^{-2}
energy	$\mathsf{L}^2\mathsf{MT}^{-2}$
entropy	$\mathsf{L}^2\mathsf{MT}^{-2}\Theta^{-1}$
electrical potential	$\mathsf{L}^2\mathsf{MT}^{-3}\mathsf{I}^{-1}$
permittivity	$\mathsf{L}^{-3}\mathsf{M}^{-1}\mathsf{T}^4\mathsf{I}^{-2}$
magnetic flux	$\mathsf{L}^2\mathsf{MT}^{-2}\mathsf{I}^{-1}$
illuminance	$\mathsf{L}^{-2}\mathsf{J}$
molar entropy	$\mathsf{L}^2\mathsf{MT}^{-2}\Theta^{-1}\mathsf{N}^{-1}$
relative density	1

3˙2. *Quantities of the same kind.* – Quantities in a general sense are also known as "kinds of quantity". Thus quantities of the same kind, which are the only ones that can be added and subtracted, are instances of the same "quantities in a general sense" or instances of the same "kinds of quantity". Quantities of the same kind have the same dimension but quantities having the same dimension are not necessarily of the same kind. Quantities of the same kind have the same units but the converse does not necessarily apply. For example, moment of force and energy have the same dimension, namely $\mathsf{L}^2\mathsf{M}\mathsf{T}^{-2}$ and the same SI unit but are not regarded as quantities of the same kind, *i.e.*, it makes no physical sense to add a moment of force to an energy. All energies, however, such as heat, kinetic energy and potential energy, are considered quantities of the same kind and have the same dimension. It is worth pointing out here that because different quantities can have the same unit, even quantities not of the same kind, one should avoid trying to identify a quantity by its unit. One might ask: how does one decide whether or not two quantities are quantities of the same kind? There is no recipe for this, since it is a matter of understanding the physical significance of the quantities in question.

Physical quantities of all kinds can, on the other hand, be multiplied and divided according to the equations of physics. Despite whatever philosophical hesitations that may still exist, we all take it as understood that it makes perfect sense to divide a length by a time to give a new quantity known as a velocity.

One can define new quantities suitable for the application in hand. For example, one can define the distance between Varenna and Turin via Milan as $D(TM)$ and use as the unit the kilometre. One can also define a new quantity $D(TR)$ as the distance from Varenna to Turin via Rome and also use the same unit since $D(TM)$ and $D(TR)$ are quantities of the same kind. Another example is to be found in the case of temperature. The fundamental physical quantity is the thermodynamic temperature symbol T whose unit is the kelvin symbol K. We have also defined another quantity known as International Practical Temperature symbol T_{90}. This is defined by the International Temperature Scale of 1990, ITS-90. Since T and T_{90} are quantities of the same kind, we are entitled to specify that the unit of T_{90} is also the kelvin. It follows that, because these two temperatures are quantities of the same kind, they can be added and subtracted and thus it makes sense to write, for example, that at about 20°C, $T - T_{90} = 5\,\text{mK}$. If the quantities T and T_{90} were not of the same kind, we would not be able to write down an expression giving the difference between a thermodynamic temperature and an International Practical Temperature.

There exists a distinction between those quantities whose magnitudes are additive for subsystems, which are known as extensive quantities and those that are not and which are known as intensive quantities. Examples of extensive quantities are length, mass and electric current and examples of intensive quantities are pressure, temperature and chemical potential (partial molar Gibbs energy).

All of this is part of the formal grammar of quantity calculus. Much of it is so obvious that it can seem trivial when stated explicitly but it underlies all of our mathematical representation of the natural world.

3˙3. *Quantity equations and numerical value equations.* – There exists an important distinction between quantity equations and numerical value equations due to the fact that the first is independent of the units use and the second is not. For example, the quantity equation for the velocity, v, of an object in uniform motion is

$$v = l/t\,, \tag{7}$$

where l is the distance travelled in a time t. This equation remains, of course, valid for any system of units. If, however, we wish to have the velocity expressed in kilometres per hour when we express the distance travelled in metres and the time in seconds we can construct a numerical value equation that will do this for us. We write

$$\{v\}_{\mathrm{km/h}} = 3.6\{l\}_{\mathrm{m}} \cdot \{t\}_{\mathrm{s}}\,, \tag{8}$$

where the subscripts represent the units of the particular numerical value in the curly brackets. If we remove the subscripts we have

$$\{v\} = 3.6\{l\} \cdot \{t\}\,, \tag{9}$$

which is a perfectly correct numerical value equation but valid only for the particular units kilometres per hour, metres and seconds for the quantities velocity, distance and time, respectively. There used to exist many physics texts where the equations were all numerical-value equations and this has rendered them much more difficult to use when, for example, we wish to convert the equations which may originally have been drawn up in imperial units or c.g.s. units, into SI.

4. – The equations of physics, definitional constants

Equations between quantities that represent the physical world often contain numerical factors, these are known as definitional constants. For example, the area A is related to the length l and height h by the equation

$$A = hl. \tag{10}$$

The area of a triangle of base length l and height h is

$$A = 1/2\,hl. \tag{11}$$

And the area of an ellipse whose major axis has length l and minor axis h is

$$A = (\pi/4)hl. \tag{12}$$

The numerical factors are all known as definitional constants because they define the quantity being represented by the equation. The quantity is area and it must be the same quantity in all three equations, *i.e.*, the three equations must be consistent. However, we have the choice to define these quantities in different ways, for example we could decide that the definitional constant in the last of these equations should be 1 instead of $(\pi/4)$, but by doing so we would be required to include a factor $(\pi/4)$ in the first two. If we did not do this, any calculation in which we try and find the relative masses of rectangular, triangular and elliptic sheets of steel would not work out. In other words, the definitional constants are necessary in order to make the equations for the areas of different shapes physically self-consistent.

A more instructive example is the numerical factor of 1/2 in the equation

$$T = (1/2)\, mv^2 \,, \tag{13}$$

relating kinetic energy T to mass m and velocity v. Remember here, the previous discussion about quantities in the general sense and quantities in the particular sense. Strictly, we should write this equation as $T_j = 1/2\, m_j {v_j}^2$ and state that we are referring to a particular body or particle designated by the subscript j. But in the ordinary course of physical discussion this is not necessary and could even obscure the essential point being made. This equation is related to that expressing the change in potential energy U as the product of a force F acting over a distance x:

$$U = -\int F\, \mathrm{d}x. \tag{14}$$

It would in principle be possible to re-define energy so that the numerical factor in eq. (13) became 1, but then the factor 2 would appear in (14), so that the two formulae would take the form $T' = mv^2$ and $U' = -\int 2\, F\, \mathrm{d}x$. A prime is added to the symbols for kinetic and potential energy here in order to emphasize that they are not the same quantities that appear on the left of (13) and (14), because we have defined energy in a different way, so that $T' = 2T$ and $U' = 2U$. Changing the value of a definitional constant in any equation corresponds to changing the definition of one of the quantities in the equation, and to avoid confusion we should always then adopt a new symbol for the re-defined quantity.

As a third example of a definitional constant we may consider the commutator of the operators for coordinate q and momentum p in quantum mechanics, given by the relation

$$qp - pq = ih/2\pi, \tag{15}$$

where h is the Planck constant. The factor $1/2\pi$ in this equation may be regarded as a definitional constant. Dirac chose to re-define the constant h in this equation so that the definitional constant became 1, by writing (15) in the form

$$qp - pq = i\hbar, \tag{16}$$

where it is clear that $\hbar = h/2\pi$. In this case both of the expressions (15) and (16) are in common use today, but there is no confusion because the different symbols h and $\hbar$ are used for the two forms of the Planck constant. The change in the definitional constant between (15) and (16) is sometimes called rationalization. A further example of rationalization occurs in the equations of electromagnetic theory, as we shall discuss below in sect. **5**.

The equations of physics play a key role in defining the quantities that we use, and hence in defining units. These equations are for the most part well established today, so that definitional constants such as the factor 1/2 in (13) are taken for granted, and we never even think of defining things another way.

5. – The problem of electrical quantities and units

Electrical and magnetic phenomena have long been known. There is article in the *Philosophical Transactions of the Royal Society* in 1776, on the electrical phenomena associated with electric eels. The first serious attempt to quantify magnetic and electrical effects was, however, that of Gauss in 1832. He devised an absolute system of units based on the millimeter, milligram and second for the measurement of magnetic phenomena. He was followed by Weber who extended Gauss' system to other electrical measurements. It was found that magnetic and electrical phenomena could be described by two mutually incompatible systems of quantities and units depending upon whether one began with the inverse square law of force between two magnetic poles or from that between two electrical charges. For each of these two systems of equations Weber in 1851 defined a coherent absolute system unit system based on the centimetre, gram and second as fundamental mechanical units. These became known as the c.g.s. electromagnetic and c.g.s. electrostatic systems. Weber carried out experiments to realize the electromagnetic unit of electric resistance which, in terms of the fundamental mechanical units was found to have to be equivalent to one centimetre per second. In the 1850s William Thompson, later Lord Kelvin, used a similar system but based upon the then British units. At about this time the British Association for the Advancement of Science (BAAS) began to be interested in electrical units. A Standards Committee was created and in 1863 adopted the first among what came to be an important series of decisions concerning electrical quantities and units. It concerned the method devised by William Thompson for realizing the absolute unit of electrical resistance. The BAAS went on to recommend practical units that were more appropriate for electrical engineering; it had been found that the c.g.s. electromagnetic units for voltage and resistance were many orders of magnitude smaller than what in practice was required. The BAAS practical units for resistance and voltage were made to be 10^9 and 10^8 times larger than the corresponding electromagnetic units. The importance of reaching world-wide agreement on units for electromagnetic measurements led to the first International Congress of Electricians which was held in Paris in 1881. This was the forerunner of a series of such Congresses that resulted in the creation, in St. Louis, USA, in 1904, of the International Electrotechnical Commission.

The BAAS practical units were approved by the 1881 Congress and later at the Congress held in Chicago in 1893, were formally adopted. Although these units were in principle related to the c.g.s. absolute units by exact factors of ten, as time went on the material realizations of these units, namely the column of mercury for the ohm and the Westin Cell for the volt became the *de facto* world standards for electrical units, they became known as the International Units. This was not considered a satisfactory situation. At the beginning of the 20th century a fundamental change in thinking was proposed by the Italian scientist and engineer, Giovanni Giorgi. He realized that many of the difficulties of the c.g.s. system when applied to electrical measurements would be resolved if a fourth fundamental unit of an electrical nature were to be added. The electrical quantities would then no longer be defined solely in terms of mechanical units as Maxwell had said. Instead, and this is the crucial innovation, in the force law of magnetic interaction there would be a constant of proportionality known as the vacuum permeability. With Giorgi's choice of mechanical units being the kilogram, metre and second (the MKS system) with the fourth unit the ampere, the constant of proportionality would be 10^{-7} MKS units of force per ampere. This was not all, the addition of the fourth electrical unit did away with another of the consequences of the c.g.s. electrical units, namely fractional exponents in expressions for certain quantities. For example, electric current was $\mathrm{cm}^{1/2}\,\mathrm{g}^{1/2}\,\mathrm{s}^{-1}$ and the unit for electrical resistance was the cm/s. All of this seen with hindsight is an indication that three fundamental units were not sufficient and that a fourth was needed.

Another change in electrical quantities and units that took place during the first half of the twentieth century was the so-called rationalization. It had been remarked by Oliver Heaviside in the 1880s that electrical science was full of factors of 4π that seemed to appear at random in equations. He proposed to begin by including a factor of $1/4\pi$ in the basic expression of Coulomb's law. The problem arises of course when moving between systems with spherical symmetry and planar symmetry, both of which are needed at various times. Giorgi also took this up and by the time the future SI system had taken shape in the 1930s there was general agreement that Giorgi's proposal for what became known as rationalization of electrical quantities should take place. The SI is a system of units based on rationalized system of quantities and quantity equations. The comparison between the units of the four-dimensional rationalized SI and those of the three-dimensional non-rationalized c.g.s. is not simple. The basis of any comparison must be the comparison of the quantities in the two systems because these are different. In principle the comparison can be made by comparing the corresponding quantity equations in the two systems. For example, the basic equations for the force between two parallel straight conductors in the two systems are

$$F/L = \mu_0\, I I'/2\pi r \tag{17}$$

and

$$F/L = 2\, I_m {I'}_m/r \tag{18}$$

for the four-dimensional and three-dimensional systems, respectively. From this we see that

$$I_m = \sqrt{(\mu_0/4\pi)} \cdot I. \tag{19}$$

From this, together with the relation that $\varepsilon_0 \cdot \mu_0 \cdot c^2 = 1$, the relations between other corresponding quantities can be derived. Because in practice this is not a straightforward operation, I refer, for a more detailed explanation, to the Varenna lectures of I. Mills in 2000 [3].

6. – The International System of Units, the SI, a coherent system of units

Although these lectures are mainly concerned with physical quantities, the intimate link that exists between quantities and units calls for a short digression on the SI. The formal description of the SI is to be found in the SI Brochure A shortened pocket version is also now widely available so I do not need to elaborate on the details of the structure of the SI as an outline has been given by A. Wallard in one of the preceding lectures at this School. There are, however, a number of points that it is useful to highlight in the context of our discussion on physical quantities.

In the same way that derived quantities are defined as the products of the powers of the base quantities, derived units of the SI are defined as products of the powers of the base units of the SI. When the product of the powers includes no numerical factors other than the number one, the derived units are called coherent units. The base and derived units of the SI form a coherent set known as the coherent SI units. The word coherent is used in the following sense: when coherent units are used, equations between numerical values of quantities take exactly the same form as the equations between the quantities themselves. Thus, if only coherent units are used, conversion factors between units are never required. The expression for the coherent unit of a derived quantity may be obtained from the dimensional product of that quantity by replacing the symbol for each dimension by the symbol of the corresponding unit. For example, the defining equation for a force is $F = m\,a$, whose dimension is MLT^{-2} so that its SI unit is $\mathrm{m\,kg\,s^{-2}}$ see the table III below where the formal correspondence between dimensional and unit symbols is the following:

$$\begin{array}{lll} \mathsf{L} \rightarrow \mathrm{m}, & \mathsf{I} \rightarrow \mathrm{A} & \mathsf{J} \rightarrow \mathrm{cd} \\ \mathsf{M} \rightarrow \mathrm{kg}, & \Theta \rightarrow \mathrm{K} & \\ \mathsf{T} \rightarrow \mathrm{s}, & \mathsf{N} \rightarrow \mathrm{mol} & \end{array}$$

Each physical quantity has only one SI coherent unit. This may be written in different forms, however, if it includes derived SI units having special names and symbols. For example, the SI unit of energy is written in terms of the base units as $\mathrm{m^2\,kg\,s^{-2}}$, but this has a special name, the Joule symbol J where by definition $\mathrm{J} = \mathrm{m^2\,kg\,s^{-2}}$.

It is often an advantage to use special names and symbols for expressing compound units, for example molar entropy can be expressed more simply than in table III by using the Joule:

$$\text{molar entropy} = \mathrm{J\,K^{-1}\,mol^{-1}} \text{ instead of } \mathrm{kg\,m^2\,s^{-2}\,K^{-1}\,mol^{-1}}.$$

Table III. – *Examples of derived units with their dimension and with their expression in terms of base units.*

Quantity	Dimension	SI unit in terms of SI base units
velocity	LT^{-1}	$\mathrm{m\,s^{-1}}$
angular velocity	T^{-1}	$\mathrm{s^{-1}}$
force	LMT^{-2}	$\mathrm{kg\,m\,s^{-2}}$
energy	$\mathsf{L}^2\mathsf{MT}^{-2}$	$\mathrm{kg\,m^2\,s^{-2}}$
entropy	$\mathsf{L}^2\mathsf{MT}^{-2}\Theta^{-1}$	$\mathrm{kg\,m^2\,s^{-2}\,K^{-1}}$
electrical potential	$\mathsf{L}^2\mathsf{MT}^{-3}\mathsf{I}^{-1}$	$\mathrm{kg\,m^2\,s^{-3}\,A^{-1}}$
permittivity	$\mathsf{L}^{-3}\mathsf{M}^{-1}\mathsf{T}^4\mathsf{I}^{-2}$	$\mathrm{A^2\,s^4\,kg^{-1}\,m^{-3}}$
magnetic flux	$\mathsf{L}^2\mathsf{MT}^{-2}\mathsf{I}^{-1}$	$\mathrm{kg\,m^2\,s^{-2}\,A^{-1}}$
illuminance	$\mathsf{L}^{-2}\mathsf{J}$	$\mathrm{cd\,m^{-2}}$
molar entropy	$\mathsf{L}^2\mathsf{MT}^{-2}\Theta^{-1}\mathsf{N}^{-1}$	$\mathrm{kg\,m^2\,s^{-2}\,K^{-1}\,mol^{-1}}$
relative density	1	1

The SI is based upon seven base quantities each with its corresponding base unit. The question might be asked as to how the choice of seven base quantities and units was made. Of course, since our system of weights and measures has its origins in the distant past, it is not realistic to pretend that it was drawn up on wholly logical grounds following a set of carefully though out principles. We know very well that mass, length and time were the fundamental or basic (as opposed to the base) quantities and the units of mass length and time were those that had been used from time immemorial for human commerce. It became clear that at least one additional unit would be needed towards the end of the 19th century with the rise of electrical engineering. Chemists in the middle of the 20th century pleaded for a unit to be used with chemical reactions in which it is not the mass of the reacting substances that matters but more something related to their valency and hence the mole was introduced. The important position of light and the lighting industry led to the call for a special quantity and unit to express the visible effects of light.

The SI was originally adopted and our ideas of quantities and units developed before relativistic effects were known. In the most recent editions of the SI Brochure it is made clear that the quantities and units referred to are proper quantities and proper units. That it to say, they are quantities and units that are defined in the local small spatial domain, so that effects of general relativity are negligible and the only relativistic effects that need to be taken account of are those of special relativity such as the second-order Doppler effect in frequency standards. It must be remembered, however, that if experiments are being carried out or measurements are made from a distance where either the velocity or the gravitational potential is significantly different, then both quantities and units will be subject to the effects of general and special relativity. Today this is most evident in observations using satellite clocks, such as those in the GPS satellites.

7. – Biological quantities and units

The SI brochure now also makes special reference to what are often known as biological quantities and units. Biological quantities are those defined in terms of the response of the human body or biological tissues to outside physical factors or to therapeutic substances. The quantitative measurement of such quantities is often very difficult because the mechanism of the specific biological effect is not sufficiently well understood for it to be fully describable in terms of physico-chemical parameters. In the cases of the response of the human body to outside factors, the biological effect involves weighting factors that may not be well known or defined and are often both energy and frequency dependent. Optical radiation, depending upon its frequency, causes chemical changes in living and non-living tissues. This property is called actinism and radiation capable of causing such changes is known as actinic radiation. In some cases the results of measurements of photochemical and photobiological quantities of this kind can be expressed in SI units. Photometric quantities can be expressed in SI units (see below) but other actinic effects such as the reddening of the skin as a result of UV solar radiation, sunburn, are quantified in terms of action spectra defined by the International Commission for Illumination, CIE. Sound pressure waves, if of sufficiently high frequency and intensity can also cause physical changes in human tissues as well as to the human auditory function. The effects on hearing as well as the effects of high intensity ultrasonic radiation are important in diagnosis and therapy. It is well known that ionizing radiation deposits energy in irradiated living and non-living biological tissues. High doses of ionizing radiation kill cells and this is used in radiotherapy. Appropriate weighting factors have been devised in order to compare therapeutic effects with absorbed dose as a function of type of radiation. Very low doses of ionizing radiation are known to cause damage to the DNA of living cells. Different weighting factors are used to estimate the effects of low doses for radiation protection purposes.

7˙1. *Photobiological quantities.* – For all of these different situations, special weighting factors have been established, perhaps the most well known of which is that used for quantifying the response of the human eye to light. These so-called photometric quantities must take into account both the purely physical characteristics of the radiant power stimulating the visual system and the spectral responsivity of the latter, see ref. [16]. The subjective nature of the spectral responsivity of the eye sets photometric quantities apart from purely physical ones. The most important photometric quantities are luminous flux, luminous intensity, luminance and illuminance. These are all based on the fundamental definition of luminous flux φ_v which is derived from radiant flux φ_c by evaluating the radiation according to its action upon what is known as the CIE standard observer. For convenience, however, the base quantity is luminous intensity and its SI unit is the candela. For photopic vision luminous flux is defined by the relation

$$\varphi_v = K_m \int_A \varphi_{c,\lambda} \, V(\lambda)\mathrm{d}\lambda, \tag{20}$$

where φ_v is the luminous flux in lumens, $\varphi_{c,\lambda} = \mathrm{d}\varphi_c/\mathrm{d}\lambda$ the spectral concentration of radiant flux (radiant power) in watts per metre, $V(\lambda)$ is the spectral luminous efficiency function for photopic vision and $K_m = 683\,\mathrm{lm\,W^{-1}}[V(\lambda_m)/V(555.016\,\mathrm{nm})] \approx 683\,\mathrm{lm\,W^{-1}}$.

A similar but slightly different relation is given for scotopic (night time) vision.

7˙2. *Quantities for ionizing radiation.* – For ionizing radiation, the quantities needed to describe in a quantitative way the interaction with the human body are all related to amount of energy deposited by the radiation in tissue. The effect on the tissue depends on the type and energy of the radiation. The basic quantity is the absorbed dose symbol D. For any ionizing radiation this is the mean energy imparted to an element of irradiated tissue divided by the mass of the element. The SI unit is the grey which is a special name for the Joule per kilogram. Another quantity is the dose equivalent H. This is the product of D and a factor Q at the point of interest in the tissue where Q is known as the quality factor for that radiation. The unit for dose equivalent is the sievert, also equal to Joule per kilogram. For uncharged particles, *i.e.*, indirectly ionizing radiation, the relevant quantity is known as kerma, symbol K (a name derived from Kinetic Energy Released in Matter). Kerma is defined as the sum of the initial kinetic energies of all charged particles liberated in an element of matter divided by the mass of that element. The unit of kerma is the grey. The establishment of appropriate quality factors for different types and energies of radiation is the key to proper control both of ionizing radiation therapy for cancer and for health protection in the case of very low doses.

7˙3. *Biological quantities for medicine.* – For a very wide range of therapeutic and other biological products where the biological effect is not sufficiently well understood for the effects to be quantifiable in physico-chemical terms, the World Health Organization, WHO, establishes the so-called WHO International Units. These result from studies of the therapeutic effect of particular samples of the substance in question. The therapeutic effects are quantified by bioassays and subsequently maintained and distributed through stocks of original substances. These original stocks are held by a laboratory designated by the WHO for this purpose. For example, one of the largest is the National Institute for Biological Standards and Control (NIBSC) in London. The obvious problem with such a method is that when the original stock of substance runs out, the International Standard no longer exists. This problem is dealt with (but it is also much discussed) by carrying out studies to compare samples from the original stock before it has run out with a new stock made to be as close as possible to the original. But as bioassays are not precise quantitative operations there is no doubt that successive versions of an International Standard do not precisely repeat the original.

Major areas of activity in this field are now for example DNA vaccines and gene therapy where the quantities to be defined and units of measurement are highly complex and for the moment beyond the range of methods that link them to the SI. Experience has shown, however, that as the science of molecular biology advances, it becomes possible progressively to define the specific therapeutic or biological quantity such that measurements in terms of SI units, namely the mole or the kilogram, become possible. When first discovered, for example, insulin and penicillin had to be quantified by biological assays but their action is now sufficiently well understood for them to be prepared as pure chemical substances and their therapeutic effects directly related to the quantity present measured in moles or grams. This is not the case for such things as prions, whose biological effects cannot be related yet to any measurable physico-chemical property of the protein in question. Molecular biology is a very active area of science and the metrology of highly complex biological systems is becoming increasingly important. There are still many hundreds of WHO International Standards.

8. – The use of the fundamental constants as reference quantities for the definition of units

It is generally understood that in setting up a system of quantities and units, one begins by choosing a set of base units and then for each of these one chooses a convenient example of each as the unit. However, we have already seen that in reality our present system of quantities and units did not evolve quite like this. The basic units for mass length and time evolved in a pragmatic way over the centuries and millennia long before precise ideas concerning physical quantities crystallized.

Today, we have the opportunity of looking anew at our definitions of the base units of our system of measurement. This has arisen because we are now at the point where it becomes possible to use the fundamental constants as references for the base quantities rather than specifying a particular example of each quantity as the unit. In the case of the unit of length, the metre, we have already done this in the definition of the metre of 1983:

The metre is the length of the path travelled by light in vacuum during a time interval of 1/299 793 458 *of a second.*

This definition is followed in the latest edition of the SI Brochure by the statement "It follows that the speed of light in vacuum is 299 793 458 metres per second exactly".

In a recent paper in *Metrologia* some of us [17] have suggested that the definitions of the base units can all be formulated in terms of a set of fundamental constants as references. These are shown here in table IV.

In our proposal we then go further and suggest that the International System of Units, the SI, is the system of units scaled so that the

1) the ground-state hyperfine-splitting transition frequency of the caesium 133 atom $\Delta\nu(^{133}\mathrm{Cs})_{\mathrm{hfs}}$ is 9 192 631 770 hertz;

2) the speed of light in vacuum c_0 is 299 792 458 metres per second;

3) the Planck constant h is $6.626\ 069\ 3 \times 10^{-34}$ joule second;

4) the elementary charge e is $1.602\ 176\ 53 \times 10^{-19}$ coulomb;

5) the Boltzmann constant k is $1.380\ 650\ 5 \times 10^{-23}$ joules per kelvin;

6) the Avogadro constant N_A is 6.022 141 5 × 10^{23} per mole;

7) the spectral luminous efficacy of monochromatic radiation of frequency 540 × 10^{12} hertz $K(\lambda_{540})$ is 683 lumens per watt.

Accompanying this definition of the SI would be a list of representative units, together with a representative list of the quantities whose values could be expressed in those units. This list would include, of course, the metre for length, the kilogram for mass, the second

Table IV. – *The definitions of the kilogram, ampere, kelvin, and mole that link these units to exact values of the Planck constant h, elementary charge e, Boltzmann constant k and Avogadro constant N_A, respectively.*

kilogram	ampere	kelvin	mole
(kg-1a) The kilogram is the mass of a body whose equivalent energy is equal to that of a number of photons whose frequencies sum to exactly $[(299\ 792\ 458)^2/662\ 606\ 93]$ $\times 10^{41}$ hertz. (kg-1b) The kilogram is the mass of a body whose de Broglie-Compton frequency is equal to exactly $[(299\ 792\ 458)^2/(6.626\ 069\ 3\times 10^{-34})]$ hertz.	(A-1) The ampere is the electric current in the direction of the flow of exactly $1/(1.602\ 156\ 53\times 10^{-19})$ elementary charges per second.	(K-1) The kelvin is the change of thermodynamic temperature that results in a change of thermal energy kT by exactly $1.380\ 650\ 5\times 10^{-23}$ joule.	(mol-1) The mole is the amount of substance of a system that contains exactly $6.022\ 141\ 5\times 10^{23}$ specified elementary entities, which may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles.
(kg-2) The kilogram, unit of mass, is such that the Planck constant is exactly $6.626\ 069\ 3\times 10^{-34}$ joule second.	(A-2) The ampere, unit of electric current, is such that the elementary charge is exactly $1.602\ 176\ 53\times 10^{-19}$ coulomb.	(K-2) The kelvin, unit of thermodynamic temperature, is such that the Boltzmann constant is exactly $1.380\ 650\ 5\times 10^{-23}$ joule per kelvin.	(mol-2) The mole, unit of amount of substance of a specified elementary entity, which may be an atom, molecule, ion, electron, any other particle, or a specified group of such particles, is such that the Avogadro constant is exactly $6.022\ 141\ 5\times 10^{23}$ per mole.

for time, the ampere for electric current, the kelvin for thermodynamic temperature, the mole for amount of substance and the candela for luminous intensity, as well as the current 22 SI-derived units with special names and symbols such as the radian, newton, volt, lumen and katal and some of their corresponding quantities. Such a list could in fact be taken from, for example, tables I-IV in the BIPM SI Brochure. This single definition and list, together with the same system of quantities and laws of physics upon which the present SI rests, establishes the entire system without the introduction of base units and derived units —all units are on an equal footing. Further, there is no need to be concerned about whether or not adopting exact values for these seven constants fully specifies the SI, for we know that these constants define the seven SI base units and that the SI as presently constructed is fully specified by those units([1]). This version of the SI is only a mild departure from the guiding assumption discussed in the first paragraph of sect. 1.2 of the Mills *et al.* paper [17], inasmuch as the quantities and units on which it is based are the same as the current SI; the only difference is that the categorization of units as "base" or "derived" is no longer applicable and this we see as a logical extension of current thinking.

The practical realization of any unit of this new version of the SI, whether it is one of the present base or derived units or not, would be by employing a method (a primary method) defined by an appropriate equation of physics linking the unit in question to one or more of the fixed constants. For example, the volt and ohm would be realized through the equations of the Josephson and quantum-Hall effects using the exact values of h and e; the kelvin through a primary thermometer using the exact values of k or R, and so on. The user would be at liberty to use whichever equation of physics and method is considered most appropriate. The CIPM could decide, however, to formalize certain of these methods as a *mise en pratique*.

Looking further to the future, it is of interest to speculate about eventually replacing the definition of the second based on $\Delta\nu(^{133}\mathrm{Cs})_{\mathrm{hfs}}$ with a definition that links the second to an exact value of the familiar and highly important Rydberg constant R_∞. In this case, entry 1) in the numbered list above, including the three words that precede it, would read "such that the 1) Rydberg constant R_∞ is 10 973 731.568 525 inverse metres". At present, the theory and experimental determination of hydrogen and deuterium transition frequencies are not sufficiently accurate to do this, but they could be in the future. In the formulation of the SI considered here, such a replacement could simply be made with no other change. The fact that the Rydberg constant has the unit of inverse metre and would replace a constant that has the unit of inverse second would not matter; the product $c_0 R_\infty$ would be an exactly known frequency that could be related by theory to an accurately measurable transition frequency in hydrogen.

([1]) However, the choice of constants used to define the SI is not unique. For example, statements 3) and 4), which fix the values of h and e, could be replaced by statements that fix the values of the Josephson and von Klitzing constants $K_{\mathrm{J}} = 2e/h$ and $R_{\mathrm{K}} = h/e^2$. If this were to be done, statements 5) and 6) could then be replaced with statements that fix the Stefan-Boltzmann constant $\sigma = (2/15)\pi^5 k^4/(h^3 c^2)$ and the Faraday constant $F = N_{\mathrm{A}} e$.

A major advantage of the proposed new approach is that it does away entirely with the need to specify base units and derived units, and hence the confusion that this requirement has long been recognized to engender, not the least of which is the arbitrariness of the distinction between base units and derived units. This need is eliminated by no longer having a unique, one-to-one correspondence between a particular unit and a particular fundamental constant. It thus does away with a situation such as exists with the explicit-constant definition for the ampere, (A-2) in section 2.3 of the Mills *et al.* paper, in which the unit of current is defined in terms of a constant, the elementary charge, the unit of which is not the ampere but the ampere second, or coulomb. Such cross-referencing between units in definitions can be avoided by not linking particular constants to particular units.

We emphasize that no matter which direction the CIPM chooses to take —the explicit-unit approach, the explicit-constant approach, or this last approach that defines the entire SI without linking a particular unit to the exact value of a particular constant— the same measurement system will result. In practice, if not formally, the base units and derived units will be indistinguishable and the seven constants listed above, that is, $\Delta\nu(^{133}\mathrm{Cs})_{\mathrm{hfs}}$, c_0, h, e, k, N_{A} and $K(\lambda_{540})$, will form the basis of the system.

9. – Final remarks

At the beginning I said that the SI is the international language of science. In these lectures I have attempted to give an outline of the underlying structure of the grammar of this language, namely quantity calculus, and to present some very recent proposals for basing the SI on the fundamental constants of physics. These proposals are the subject of much discussion and will certainly evolve over the next few years.

REFERENCES

[1] Guggenheim E. A., *Philos. Mag.*, **33** (1942) 479.

[2] de Boer J., *Metrologia*, **31** (1995) 405 (this article contains an extensive bibliography on quantity calculus and its history).

[3] Mills I. M., *Physical Quantities and Units* in *Recent Advances in Metrology and Fundamental Constants*, *Proceedings of the International School of Physics "E. Fermi"*, Course CXLVI, edited by Quinn T. J. and Leschiutta S. (IOS, Amsterdam, Italian Physical Society, Bologna) 2001.

[4] *Quantities, Units and Symbols in Physical Chemistry*, IUPAC, 2nd edition (Blackwell Publishing) 1993.

[5] *Quantities and Units*, ISO Standards Handbook, ISO/IEC 31 (1993). Due to be replaced by a new edition as ISO/IEC 80000 in 2007.

[6] *International Vocabulary of General and Basic terms in Metrology, VIM*, 3rd edition, to be published in 2007.

[7] *SI Brochure*, 8th edition (BIPM) 2006.

[8] Maxwell J. C., *Treatise on Electricity and Magnetism* (Oxford University Press, Oxford) 1873.

[9] HELMHOLTZ H. V., *Zahlen und Messen erkenntniss-theoretis-betrachtet*, in *Leipzig, Wiss. Abh.* III (1887), pp. 356-391.

[10] HÖLDER O., "Die Axiome de Quantitat und Lehre vom Mass", *Ber. Verhy. Sachs. Ges. Wiss. Leipzig. Math Phys. K1*, **53** (1901) 1-64.

[11] WALLOT J., *Grössengleichungen, Einheiten und Dimensionen* (1957).

[12] STILLE U., *Messen und Rechnen in der Physik* (Vieweg, Braunschweig) 1955 and 1961.

[13] LODGE A., *Nature*, **30** (1888) 283.

[14] CAMPBELL N., *Physics, the Elements* (Cambridge University Press) 1920.

[15] BRIDGMAN P. W., *Dimensional Analysis* (Yale University Press) 1922 and 1931.

[16] *Principles Governing Photometry*, BIPM, Monograph (1983).

[17] MILLS I. M. *et al.*, *Metrologia*, **42** (2006) 227.

Recent developments in uncertainty evaluation

W. Bich

Istituto Nazionale di Ricerca Metrologica, INRIM - Torino, Italy

The publication in 1993/1995 of the Guide to the Expression of Uncertainty in Measurement, GUM, stimulated a great development of research concerning measurand estimation and uncertainty evaluation. The application of the GUM method has brought into light its merits and limits. This paper reviews them and discusses the recent evolution of concepts and methods in the specific field of uncertainty evaluation.

1. – Introduction

The publication of the Guide to the Expression of Uncertainty in Measurement in 1993 [1], with its reprint in 1995, marked a significant progress in the field of uncertainty evalution. For the first time, a method was available, capable to treat random and systematic contributions to uncertainty in a unified, plausible and consistent way. The method proposed in the GUM has deep, sound foundations and far-reaching consequences. During the last decade, both aspects were extensively explored. On the one hand, the probabilistic foundations of the GUM have been better understood, on the other, its framework has been successfully applied to a wide range of experimental situations (for a non-exhaustive list, see [2]; see also [3-9]). These investigations contributed to bring into light some internal inconsistences and to better specify limits to the applicability of the GUM method. This paper reviews merits and limits of the method proposed in the GUM, and discusses the recent evolution of concepts and methods in the specific field of uncertainty evaluation.

2. – The present GUM

2˙1. *Fundamentals.* – The GUM framework is well known. The value y of the quantity Y intended to be measured, the *measurand* [10], is not obtained directly, but from the values $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$ of other *input quantities* $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)^{\mathrm{T}}$ to which the measurand is related by a functional relationship or *model* $Y = f(\boldsymbol{X})$. Most measurements can be described by this indirect scheme. The notation uses capital symbols for quantities and lower-case symbols for quantity values. Although it is accepted that the the experimental outcome concerning a quantity, be it a single value or a series of indications, is uncertain, "it is assumed that the physical quantity itself can be characterized by an essentially unique value" (GUM, 4.1.1, Note 1). Therefore, in the GUM framework the input quantities all have fixed values to which a unique value corresponds for the measurand. The GUM gives a measure for the uncertainty of the input quantity values and a recipe to propagate these input uncertainties through the model in order to obtain the corresponding uncertainty associated with the measurand value. This framework can be viewed under different angles. In its simplest interpretation, it is the old law of propagation of errors of Gauss, modernly written in terms of variances and covariances, which are better measures than errors. As such, it is based on a first-order approximation, and therefore holds for reasonably linear models, being exact in the linear case. However, variances and covariances are properties of random variables. These, in the GUM framework, are associated to the quantities and represent "the possible outcome of an observation of that quantity" (GUM, 4.1.1). To further confuse the picture, the same capital symbol is adopted for both the quantity and the associated random variable. In modern terms, the propagation is based on the approximation

$$f(\boldsymbol{X}) \approx f\left[\mathrm{E}(\boldsymbol{X})\right] + \sum_{i=1}^{N} \left.\frac{\partial f}{\partial X_i}\right|_{X_i=\mathrm{E}(X_i)} \left[X_i - \mathrm{E}(X_i)\right], \tag{1}$$

from which, by squaring and taking the expectations of both members, the well-known law of propagation of uncertainties follows:

$$\mathrm{V}(Y) \approx \sum_{i,j=1}^{N} \left.\frac{\partial f}{\partial X_i}\frac{\partial f}{\partial X_j}\right|_{X_i,X_j=\mathrm{E}(X_i,X_j)} \mathrm{Cov}\left(X_i, X_j\right), \tag{2}$$

or, in matrix terms,

$$\mathrm{V}(Y) \approx \boldsymbol{J}_{\boldsymbol{X}} \mathbf{V}_{\boldsymbol{X}} \boldsymbol{J}_{\boldsymbol{X}}^{\mathrm{T}}, \tag{3}$$

where

$$\boldsymbol{J}_{\boldsymbol{X}} = \left(\partial f/\partial X_1, \ldots, \partial f/\partial X_N\right) \tag{4}$$

is the $(1 \times N)$ Jacobian matrix and $\mathbf{V}_{\boldsymbol{X}}$ is the variance-covariance matrix of the input random vector $\boldsymbol{X}$ with element

$$\{\mathbf{V}_{\boldsymbol{X}}\}_{i,j} = \mathrm{Cov}\left(X_i, X_j\right). \tag{5}$$

Equation (2), or its matrix equivalent (3), is valid to the first order for random variables. It is distribution-free, that is, it holds for any probability density functions (PDFs) of the variables, however it needs some adaptation to be useful in the real world.

First, expectations are not known and only estimates for them are available. Second, also variances and covariances are unknown. In the following section, I will review how the present GUM deals with these issues.

2˙2. *Merits and limits.* – As concerns the first problem, that is, the fact that expectations $\mathrm{E}(X_i)$ are unknown and only estimates x_i are available for them, a further approximation is accepted, under the assumption that the estimates are good, *i.e.*, not too far away from the expectations. The derivatives are thus calculated in the estimates. This approximation is no longer distribution-free, as it holds for reasonably symmetric PDFs. For a non-symmetric PDF the estimate can be considerably far from the expectation and its value should be adequately shifted. Extensive guidance is given on this crucial point [GUM, F.2.4.4].

2˙2.1. The GUM is Bayesian. The second issue, that is, the fact that also the variances and covariances or, equivalently, that the covariance matrix $\mathbf{V}_{\boldsymbol{X}}$ is unknovn, has two aspects. Usually, some of the input estimates to a model are not obtained from series of repeated indications, but are unique values obtained by non-statistical ways. This is, for example, the case of a physical constant, or a material property, or the value of a reference standard. In this case, the notion of random variable, conventionally used to describe the behaviour of a population, seems not applicable, since no population can be imagined for these quantities. However, the mathematical properties of random variables are general and independent of the intepretation one gives to probability. It is thus sufficient to adopt a broader view of probability in which the term "random" does not apply strictly to populations, but in a broader sense to whatever is not perfectly known, that is to say, is uncertain [11]. As a matter of fact, in the GUM a distinction is made between PDFs "derived from an observed frequency distribution" and those "based on the degree of belief that an event will occur [often called subjective probability]". In any case, "both approaches employ recognized interpretations of probability." [GUM, 3.3.5]. In conclusion, as concerns the PDFs based on the the degree of belief, the GUM is Bayesian (or subjective).

2˙2.2. The GUM is frequentist. Whatever the origin of the PDF, guidance is given in the GUM on how to obtain adequate measures of uncertainty from the available knowledge, the only trace of the different views on probability being the well-known distinction between type-A and -B evaluations. These measures are the standard deviations (more exactly, the variances) of the PDFs. In type-A evaluations, for an input quantity X for which a sample of values $(x_1, \ldots, x_n)$ is available, one takes as an estimate for X the average $\bar{x} = 1/n \sum x_j$ [GUM, 4.2.1]. Accordingly, one usually takes the experimental standard deviation $s_{\overline{x}}$

$$s_{\bar{x}} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \frac{(x_j - \bar{x})^2}{n-1}} \tag{6}$$

as an estimate of the standard deviation σ [GUM, 4.2.3].

It is worth noting that the standard deviation given by eq. (6) is viewed as an estimate, with $\nu = n-1$ degrees of freedom [GUM, 4.2.6], of the "true" parameter $\sigma/\sqrt{n}$. Therefore it is considered as a realization of a random variable.

As a consequence of this approach, $s_{\bar{x}}$ has itself a standard deviation [GUM, E.4]. This conveys the idea that, for type-A evaluations, the standard uncertainty $u(x_i)$ associated with an estimate x_i of the corresponding input quantity X_i is only an estimate of its "true" uncertainty and therefore has itself an uncertainty.

In this respect, the GUM adopts a frequentist attitude.

2˙3. *Confidence interval, interval of confidence and expanded uncertainty.* – As far as only the standard uncertainty $u(y)$ associated with the measurand value y is involved, the two views, Bayesian (or subjective) and frequentist, coexist in the GUM in reasonable harmony. Unfortunately, although the standard uncertainty $u(y)$ is a well-established scale measure, in many (probably most) instances it is unfit to the purpose. In the GUM it is recognized that "in some commercial, industrial, and regulatory applications, and when health and safety are concerned, it is often necessary to give a measure of uncertainty that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand." [GUM, 6.1.2]. Many standards consider this specific measure of uncertainty. As an example, a regulation widely adopted in legal metrology requires that "the expanded uncertainty, U, for $k = 2, \ldots,$ shall be less than or equal to one-third of the maximum permissible error" [12].

The expanded uncertainty U mentioned in the example "is obtained by multiplying the combined standard uncertainty $u_c(y)$ by a *coverage factor k*" [GUM, 6.2.1]. However, "It should be recognized that multiplying $u_c(y)$ by a constant provides no new information but presents the previously available information in a different form". Based on these remarks, it can be concluded that U is not only of little usefulness, but even misleading, therefore its use should be avoided, especially in standards and regulations.

In the GUM a further uncertainty measure is introduced, namely, U_p, the expanded uncertainty U at a stipulated coverage probability p. This measure responds to the needs outlined above and, being more useful, is more demanding in terms of the effort needed to evaluate it. Historically, recognising the difficulties involved in this uncertainty measure, the Comité International des Poids et Mesures, CIPM, in its Recommendation 1 (CI-1986), requested that "in giving the results of all international comparisons or other work done under the auspices of CIPM and the Comités Consultatifs. . . the combined uncertainty of type A and type B uncertainties in terms of *one standard deviation* should be given" [GUM, A.3]. However, this pre-GUM recommendation was superseded by the Mutual Recognition Arrangement, MRA [13], according to which "The degree of equivalence of each national measurement standard is expressed quantitatively by two terms: its deviation from the key comparison reference value and the uncertainty of this deviation (at a 95% level of confidence)." [MRA, T.2]. Also the Calibration and Measurement Capabilities (CMCs) of the National Metrology Institutes are declared at a 95% level of confidence [MRA, T.7].

The problem of evaluating U_p is treated in Annex G [GUM, G.4.2], and is the following.

If all the input estimates and their associated uncertainties were obtained by type-A methods, standard statistical tools could be adopted [14,15]. Detailed information would in any case be needed on the PDF for the output estimate y. The adopted *escamotage* assumes that, if the input estimates are independent and no non-Gaussian component dominates, then, by virtue of the Central-Limit Theorem, the output PDF is reasonably Gaussian and, by multiplying $u(y)$ by a factor $k = 2$, an approximated 95% confidence interval can be obtained. In order to have a more accurate confidence interval, it should be considered that $u(y)$ is only an estimate of the "true" uncertainty, its accuracy being defined by its *effective degrees of freedom* ν_{eff}. The effective degrees of freedom ν_{eff} of $u(y)$ depend on the different degrees of freedom ν_i of the input uncertainties $u(x_i)$, and on the magnitudes of the input uncertainties themselves, according to the Welch-Satterthwaite formula [16-18]. Therefore, under the conditions of applicability of the Central-Limit Theorem, the quantity $(y - Y)/u(y)$ is a random variable with a standard Student's t-distribution (GUM formulation), or, (preferred wording), Y is a random variable following a scaled-and-shifted t-distribution, scaled by $u(y)$ and shifted by y. The relevant k factor can thus be applied to obtain the required *confidence interval.*

Unfortunately, for most practical models at least some of the input quantities are not evaluated from a sample, and only one value is available. The uncertainty of this value is evaluated by a type-B method, so that the uncertainty $u(y)$ associated with the measurand value is a combination of type-A and -B components. In this situation, a choice has to be made between the frequentist and the Bayesian attitude. The choice made in the GUM is frequentist, in that it is chosen to keep the effective degrees of freedom of $u(y)$, to be calculated with the Welch-Satterthwaite formula.

"The question arises as to the degrees of freedom to assign to a standard uncertainty obtained from a type B evaluation" [GUM, G.4.2]. In the GUM, by considering that the degrees of freedom can be viewed as a measure of the uncertainty of the uncertainty, a "subjective" degrees of freedom is associated to a type-B component as a measure of its reliability. However, degrees of freedom, a natural measure for type-A evaluations, presents a number of difficulties when applied in a subjective context. The main objection is that assigning a subjective degrees of freedom is by far more an ambitious and problematic operation than assigning a subjective uncertainty. As a matter of fact, very little or no guidance is given in the GUM concerning this problem. In addition, since type-B evaluations tend to be safe, in practical applications type-B components are often assigned very large degrees of freedom, which dominate the effective degrees of freedom of $u(y)$.

At a deeper scrutiny, the very concept of reliability of a type-B evaluation looks hard to understand. In a frequentist framework, the existence of a parent population with a distribution whose parameters are unknown and estimated by the corresponding (random) sample estimators gives a solid motivation to degrees of freedom as a measure of reliability of the estimates. In a subjective context, the idea of reliability of a "guess" looks artificial. Rather than spending efforts to evaluate the reliability of his own guess, one should be able to formulate a guess whose reliability is not an issue. As mentioned above, this tends to be the experimenters' attitude in most practical situations.

Since the standard uncertainty $u(y)$ has an effective degrees of freedom, the idea is conveyed that a true uncertainty exists for a given measurand estimate, and that the value $u(y)$ obtained is only one of the possible values for the uncertainty of that estimate. In this view, the outcome of a measurement is twofold: a measurand estimate and an uncertainty estimate. Although this viewpoint is respectable and can probably be motivated, many people (including the author) think that it is not useful in measurement science.

As a curiosity, I will mention that, since the standard uncertainty $u(y)$ contains type-B components, the interval obtained by multiplying the standard uncertainty by the coverage factor k is not called in the GUM "confidence interval", a specific term in statistics [15] but "interval of confidence". This diplomatic distinction is impossible to render in many languages. It can be viewed as the symbol of the difficulties in which the GUM occurs as a consequence of the adoption of a frequentist view in the construction of an interval of confidence.

Beside some concerns on the Welch-Satterthwaite formula [19, 20], a further major limitation in the GUM approach to U_p is its lack of generality. In many measurements the conditions of applicability of the Central-Limit Theorem are not fulfilled and the method cannot be applied. For example, the input estimates may be correlated, or a non-Gaussian input component (say, a uniform arising form a type-B evaluation) may be dominant. In this case the output distribution is trapezoidal-like, and therefore, for $p = 0.95$ in general $k \leq 2$. Last but not least, very little guidance is provided in the GUM on how to obtain an expanded uncertainty with asymmetric distributions [GUM, G.5.3].

2˙4. *Arbitrary number of measurands.* – In the GUM, there is no explicit treatment of the case in which two or more measurands are determined from a common experimental setup, *e.g.*, using the same instruments and/or procedures and/or standards. However, this case is very frequent in the practice of calibration laboratories. As instances, sets of weights, standard gauge blocks and standard resistors and capacitors, and in general all artefact standards may be calibrated with respect to a common (set of) reference standard(s) by using the same (set of) comparator(s). This *multivariate* case requires a generalisation of the GUM treatment. The latter can be viewed as a special case in which the output vector has dimension one, *i.e.*, is a scalar. It happens that special cases may hide some features which only appear when dealing with a general treatment, and uncertainty makes no exception, the most important missing feature being here the correlation between the measurand estimates. This is important when using linear combinations of the estimates (typically sums). For example, when two mass standards, which have been calibrated with respect to a single reference standard, are used together as reference for a further calibration, the uncertainty of the sum of their estimates is affected by their covariance, the latter depending upon the relative magnitude of the systematic effects in the uncertainty of the first calibration [21, 22].

2˙5. *Comments.* – In conclusion to this section, the GUM estabishes a framework which has solid foundations. Some issues remain outstanding, namely:

1) the construction of an interval of confidence (GUM wording) or (preferred wording) *coverage interval* at a stipulated *coverage probability*, that is, of an interval containing the value of a quantity with a stated probability (for a discussion on the term "coverage interval", see [23]);

2) the evaluation of standard uncertainties and covariances for an arbitrary number of measurands;

3) the construction of a coverage interval at a stipulated coverage probability for an arbitrary number of measurands.

3. – Supplements to the GUM

In 1997 the same seven organizations which had participated in the preparation of the GUM established the Joint Committee for Guides in Metrology (JCGM). The following year, the International Laboratory Accreditation Cooperation, ILAC, joined them. The JCGM has two working groups [24]. WG1 "Expression of uncertainty in measurement", has the task "to promote the use of the GUM and to prepare Supplements for its broad application". WG2 "On International vocabulary of basic and general terms in metrology", has the task "to revise and promote the use of the VIM".

To fulfill its task, WG1 is preparing a number of documents under the common banner "Evaluation of measurement data". For a review of the motivation and structure of each individual document, see [25].

Among the documents under preparation by the JCGM-WG1, the most advanced are two Supplements to the GUM, intended to overcome the difficulties outlined in the previous section. The first, "Supplement 1 to the 'Guide to the expression of uncertainty in measurement' — Propagation of distributions using a Monte Carlo method", deals, primarily but not uniquely, with the construction of a coverage interval for the measurand in a general case, that is, it aims at fulfilling the task of item 1 in subsect. **2**˙5. The second, "Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Models with any number of output quantities", (provisional title) will address the same issue for the multivariate case, that is, items 2 and 3 in the mentioned subsection. In the following of this section, I will describe the main features of Supplement 1, the one at the most advanced stage.

3˙1. *Propagation of distributions.* – A coverage interval at a prescribed coverage probability is a known fraction of a PDF, or of the corresponding Cumulative Distribution Function (DF), its integral. This suggests that, were the PDF $g_Y(\eta)$ for the measurand Y available, any desired coverage interval could be obtained by simply selecting from $g_Y(\eta)$ the appropriate fraction(s). From this more general viewpoint, the GUM law of propagation of uncertainties can be interpreted as a shortcut in which only first (expectations/estimates) and second (variances/squared uncertainties) moments are propagated.

The PDF $g_Y(\eta)$ for the measurand can be calculated in principle from the model $Y = f(\boldsymbol{X})$ and the joint PDF $g_{\boldsymbol{X}}(\boldsymbol{\xi})$ for the input variables. A formal definition [26] for the PDF for Y is

$$g_Y(\eta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_{\boldsymbol{X}}(\boldsymbol{\xi})\delta(\eta - f(\boldsymbol{\xi}))\,\mathrm{d}\xi_N \cdots \mathrm{d}\xi_1, \tag{7}$$

where $\delta(\cdot)$ denotes the Dirac delta function.

Equation (7) is obtained from the method of Jacobians (see, *e.g.*, ([14], 1.2), or ([27], Chapt. 1)), a useful tool in mathematical statistics. However, it solves only formally the problem of the PDF $g_Y(\eta)$ for Y, since the multiple integrals in it can hardly be calculated analytically except in the simplest cases, and numerical integration techniques must be in general adopted. In addition, the joint PDF $g_{\boldsymbol{X}}(\boldsymbol{\xi})$ for the input variables has to be specified.

3˙1.1. PDFs for input quantities. Let us first consider the second issue. In a frequentist framework, PDFs characterize existing populations. Therefore, the task is to guess from a data sample the appropriate PDF and estimate its parameters. Parameter estimates are uncertain by definition. In a Bayesian context, PDFs can be viewed as a formal tool to represent one's knowledge about a particular quantity. Therefore, as far as one has a sensible way to associate PDFs and relevant parameters to a given knowledge, there is no uncertainty on parameters.

It has been seen in subsection **2˙3** that to obtain, from the standard uncertainty $u(y)$, U_p, the expanded uncertainty U for the measurand estimate y at a stipulated coverage probability p, a choice has to be made between the frequentist and Bayesian approach adopted for type-A and -B evaluations, respectively, of the input components $u(y_i)$. It has been seen also that the GUM choice is frequentist, and the difficulties inherent in this choice have been outlined.

In Supplement 1 the alternative choice has been made, on the grounds that it is easier and more intuitive to encompass data arising from populations in a Bayesian scheme than *viceversa.* To further simplify, only the case of independent input quantities has been considered, so that the joint PDF $g_{\boldsymbol{X}}(\boldsymbol{\xi})$ for the input vector $\boldsymbol{X}$ factorizes in the product of N-independent PDFs for the input variables X_i

$$g_{\boldsymbol{X}}(\boldsymbol{\xi}) = \prod_{i=1}^{N} g_{X_i}(\xi_i). \tag{8}$$

The only joint PDF considered in Supplement 1 is the multivariate Gaussian

$$g_{\boldsymbol{X}}(\boldsymbol{\xi}) = \frac{1}{\left\{(2\pi)^N \det \mathbf{V}_{\boldsymbol{X}}\right\}^{1/2}} \times \exp\left[-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{V}_{\boldsymbol{X}}^{-1}(\boldsymbol{\xi}-\boldsymbol{\mu})\right], \tag{9}$$

in which $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ and

$$\mathbf{V}_{\boldsymbol{X}} = \begin{bmatrix} \mathrm{V}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_N) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{V}(X_2) & \cdots & \mathrm{Cov}(X_2,X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_N,X_1) & \mathrm{Cov}(X_N,X_2) & \cdots & \mathrm{V}(X_N) \end{bmatrix} \tag{10}$$

is the *covariance matrix* of X. A covariance matrix is by construction a *positive definite* matrix [28].

Therefore, our problem reduces to the assignment of N appropriate scalar PDFs to input estimates and to obtaining from these the PDF for the output quantity.

The way to assing PDFs within type-B evaluations is described to a considerable extent in the GUM. For example, when an estimate x_i is available from an handbook with, say, maximum and minimum possible values x_i+a and x_i-a, the GUM prescription is to adopt a uniform distribution with mean x_i and standard deviation $s = a/\sqrt{3}$. This is just an example of assignment based on the available information of this kind. In general, a suitable criterion in this field comes from application of the Principle of Maximum Entropy, a variational principle well-known in information theory, suitably adapted to measurement theory [29].

As concerns type-A evaluations, that is, those for which a sample of n data is available, Bayes' Theorem helps in establishing the appropriate PDF [11]. Let us assume that the unknown input quantity X (we drop here the subscript for brevity) has unknown value μ. While taking a series of indications x_j $(j = 1, \ldots, n)$, random effects cause a scattering which is conventionally described by a Gaussian distribution with expectation μ (the "true value" of X) and standard deviation σ (which has nothing to do with the quantity, being caused by superposed random effects). It is believed that, under suitable conditions, scatter experienced by the sample mean $\bar{x} = 1/n \sum x_j$ is narrower by a factor $\sqrt{n}$. Therefore, in both frequentist and Bayesian approaches one would normally take the sample average and its standard deviation as given by eq. (6). However, in a frequentist scheme these two quantities are (random) statistics estimating the corresponding population parameters μ and σ/n, as already outlined in subsect. **2**·2.2. In a Bayesian approach, on the contrary, these two quantities are used to construct a suitably scaled-and-shifted t-distribution with $\nu = n-1$ degrees of freedom. This distribution or, equivalently, the corresponding PDF, describe the knowledge available on X.

The PDF of a standard t-distribution is

$$g(t,\nu) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\Gamma(\frac{\nu}{2})} \left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \tag{11}$$

where

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t, \qquad z > 0,$$

is the gamma function.

The t-distribution has expectation $\mathrm{E}(t) = 0$ and variance $\mathrm{V}(T) = \nu/(\nu - 2)$ $(\nu > 2)$. For $\nu \to \infty$ it degenerates in the standard Normal $\mathrm{N}(0, 1)$ with expectation $\mu = 0$ and variance $\sigma^2 = 1$.

The appropriate PDF for the example given above is the t-counterpart to a Gaussian $\mathrm{N}(\bar{x}, s^2/n)$, namely,

$$g_X(\xi) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)\sqrt{(n-1)\pi}} \times \frac{1}{s/\sqrt{n}} \times \left(1 + \frac{1}{n-1}\left(\frac{\xi - \bar{x}}{s/\sqrt{n}}\right)^2\right)^{-n/2}, \tag{12}$$

which is obtained from PDF (11) by using the shifting-and-scaling variable transform

$$\xi = \bar{x} + \frac{s}{\sqrt{n}} t. \tag{13}$$

PDF (12) has expectation and variance

$$\mathrm{E}(X) = \bar{x}, \qquad \mathrm{V}(X) = \frac{n-1}{n-3}\frac{s^2}{n}, \tag{14}$$

which tend to μ and s^2/n, respectively, for $\nu \to \infty$. The expectation is only defined for $n > 2$ and the variance for $n > 3$. From eq. (14), the best estimate x for the quantity X and its standard uncertainty $u(x)$ are taken as

$$x = \bar{x}, \qquad u(x) = \sqrt{\frac{n-1}{n-3}\frac{s^2}{n}}, \tag{15}$$

respectively. The expression for the standard uncertainty is the Bayesian counterpart to the sample standard deviation, given by eq. (6), which is recommended in the frequentist context of the present GUM. The former is larger than the latter, the difference decreasing for increasing n. However, as already mentioned, the latter is a statistic with $\nu = n - 1$ degrees of freedom, the former is the scale parameter of a subjective PDF. In a sense, the information concerning reliability has been transferred to the PDF itself.

For $n \leq 3$, the standard deviation diverges and cannot be taken as the standard uncertainty for x. *Ad-hoc* solutions have been proposed [30,31]. However, this feature of the standard deviation of a t-distribution does not affect the propagation of a PDF such as that described by eq. (12), the only requirement on it being that $\nu > 0$. Therefore, with as little as two data (the minimum meaningful value to form an average and a standard deviation) the recommended PDF can be assigned and propagated.

3˙1.2. Counterpart to degrees of freedom for type-B evaluations. In subsect. **2**˙3 the GUM procedure of assigning degrees of freedom to type-B evaluations of uncertainty as a measure of reliability has been criticised, on the grounds that type-B evaluations look intrinsically reliable. At a closer look, however, a specific concern about reliability arises. This relates to the following: when taking from a certificate or a handbook a value x for a quantity X for which a tolerance is given in terms of an interval $X = x \pm \Delta$, the GUM recommendation is to assign a uniform PDF [GUM, 4.3.7]

$$g_X(\xi) = \begin{cases} 1/2\Delta, & x - \Delta \leq \xi \leq x + \Delta, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

X has expectation and variance

$$\mathrm{E}(X) = x, \qquad \mathrm{V}(X) = \frac{(2\Delta)^2}{12}\,. \tag{17}$$

The point is here that Δ is never known exactly, due to the finite number of decimal digits with which it is specified, so that it is affected by an uncertainty (which can be viewed as a genuine uncertainty of a tolerance) depending upon the number of decimal digits with which the tolerance is given. According to the Principle of Maximum Entropy, the appropriate PDF for this case turns out to be a *curvilinear trapezoid.* This PDF is trapezoidal-like, but has flanks that are not straight lines. What matters here is that it has the same expectation as the uniform PDF (16) whereas its variance is

$$\mathrm{V}(X) = \frac{(2\Delta)^2}{12} + \frac{d^2}{9}\,, \tag{18}$$

where d is half the least-significant digit of Δ. Variance (18) is larger than that given by eq. (17) and reduces to it for $d = 0$.

This way of treating the available information by selecting the suitable PDF is the Bayesian counterpart to the assignment of degrees of freedom to a uniform PDF required by the frequentist scheme of the GUM.

3˙1.3. Multivariate example. To give a last, multivariate example of the way in which available information is embodied in a PDF, let us come back to the case in which an estimate $\boldsymbol{x}$ is available for the vector quantity $\boldsymbol{X}$, with its standard uncertainties and covariances, collected in the variance-covariance matrix as in eq. (10). This matrix is preferably called in this context the *uncertainty matrix* $\boldsymbol{U_x}$ and written as

$$\boldsymbol{U_x} = \begin{bmatrix} u^2(x_1) & u(x_1, x_2) & \cdots & u(x_1, x_N) \\ u(x_2, x_1) & u^2(x_2) & \cdots & u(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ u(x_N, x_1) & u(x_N, x_2) & \cdots & u^2(x_N) \end{bmatrix}. \tag{19}$$

The recommended PDF would be in this case the multivariate Gaussian, here written as

$$g_{\boldsymbol{X}}(\boldsymbol{\xi}) = \frac{1}{\left\{(2\pi)^N \det \boldsymbol{U}_{\boldsymbol{x}}\right\}^{1/2}} \times \exp\left[-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{x})^{\mathrm{T}}\boldsymbol{U}_{\boldsymbol{x}}{}^{-1}(\boldsymbol{\xi}-\boldsymbol{x})\right]. \tag{20}$$

Again, the estimate $\boldsymbol{x}$ and its uncertainty matrix $\boldsymbol{U}_{\boldsymbol{x}}$, viewed as statistics in the frequentist approach, are here the expectation and the covariance matrix, respectively, of a subjective PDF.

3·1.4. Propagation. Considering that each of the input PDFs can be viewed as a filter through which one can access the (virtual) population of values concerning the relevant quantity X_i, the propagation of the N input PDFs through the model is obtained by use of Monte Carlo simulation [32-34], implemented in the following way:

- Draw simultaneously from each population a quantity value x_i. These values are taken at random, according to the density assigned to their population. The outcome of this drawing is a vector $\boldsymbol{x}_r$ $(r = 1, \dots, M)$.
- Propagate vector $\boldsymbol{x}_r$ through the model to obtain a possible value y_r for the measurand Y.
- Iterate M times the steps above to obtain a numerical simulation of the population of possible values for the measurand Y.

Once the population has been obtained, an estimate and its standard uncertainty can be calculated by taking the population average $\widetilde{y}$

$$\widetilde{y} = \frac{1}{M}\sum_{r=1}^{M} y_r \tag{21}$$

and standard deviation $u(\widetilde{y})$

$$u^2(\widetilde{y}) = \frac{1}{M-1}\sum_{r=1}^{M}(y_r - \widetilde{y})^2. \tag{22}$$

A coverage interval at the prescribed coverage probability p (the main motivation for all the effort described above) can be easily constructed by removing the appropriate fractions from the tails. The coverage interval is given in terms of its endpoints. Since in general the PDF for Y is not symmetric, the interval for a stated probability is not unique. Therefore a choice has to be made, the two most popular alternatives being the *probabilistically symmetric* and the *shortest* coverage intervals. The endpoints of the former are the $(1-p)/2$- and $(1+p)/2$-quantiles, providing a $100p\%$ coverage interval. The latter has the property that, for a unimodal, or single-peaked distribution, it contains the mode, the most probable value.

4. – Conclusions

Supplement 1 to the GUM, based on the same foundations on which the GUM was built, presents a consistent method to construct a coverage interval for a measurand, thus obviating a major difficulty of the GUM itself. Point estimates are still possible, although the results would in most cases differ from the corresponding GUM result, due to the lack of generality of the GUM approach. This circumstance will lead to a revision of the GUM, in order to resolve its internal inconsistencies and re-align it with more recent documents, while preserving its standing as the leading document in the field of uncertainty evaluation.

REFERENCES

[1] BIPM, ISO, IEC, IFCC, IUPAC, IUPAP, OIML, *Guide to the Expression of Uncertainty in Measurement* (International Organization for Standardization, Geneva) 1993-95.

[2] Joint Committe for Guides in Metrology - Working Group 1, `http://www.bipm.org/en/committees/jc/jcgm/wg1_bibliography.html`.

[3] Bich W., Cox M. G. and Harris P. M., *Metrologia*, **30** (1993/94) 495.

[4] Stuart P. R., *Metrologia*, **30** (1994) 727.

[5] Decker J. E. and Pekelsky J. R., *Metrologia*, **34** (1997) 479.

[6] Tsai B. K. and Carol Johnson B., *Metrologia*, **35** (1998) 587.

[7] Hamilton C. A. and Tang Y. H., *Metrologia*, **36** (1999) 53.

[8] Lira I. and Kyriazis G., *Metrologia*, **36** (1999) 163.

[9] Irikura K. K., Johnson R. D. III and Kacker R. N., *Metrologia*, **41** (2004) 369.

[10] BIPM, ISO, IEC, IFCC, IUPAC, IUPAP, OIML, *International vocabulary of basic and general terms in metrology (VIM) - Third edition* Final Draft, 2006.

[11] Press S. J., *Bayesian Statistics: Principles, Models, and Applications* (John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore) 1989.

[12] Organisation Internationale de Métrologie Légale, *International Recommendation R 111*, OIML, Paris 2004, `http://www.oiml.org/publications/R/R111-1-e04.pdf`.

[13] Comité International des Poids et Mesures, *Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes* BIPM, Paris 1999, `http://www.bipm.org/utils/en/pdf/mra_2003.pdf`.

[14] Bickel P. J. and Docksum K. A., *Mathematical statistics: Basic Ideas and Selected Topics* (Prentice Hall, Englewood Cliffs, New Jersey) 1977.

[15] ISO, *ISO 3534-1. Statistics Vocabulary and symbols Part 1: Probability and general statistical terms* (International Organization for Standardization, Geneva) 2006.

[16] Satterthwaite F. E., *Psychometrika*, **6** (1941) 309.

[17] Welch B. L., *Biometrika*, **34** (1947) 28.

[18] Welch B. L., *Biometrika*, **29** (1938) 350.

[19] Ballico M., *Metrologia*, **37** (2000) 61.

[20] Hall B. D. and Willink R., *Metrologia*, **38** (2001) 9.

[21] Bich W., *Metrologia*, **27** (1990) 111.

[22] EA, *Doc. 4/02, Expression of the Uncertainy of Measurement in Calibration*, European co-operation for Accreditation, 1999, `http://www.european-accreditation.org/n1/doc/EA-4-02.pdf`.

[23] Willink R., *Metrologia*, **41** (2004) L5.

[24] BIPM, `http://www.bipm.org/en/committees/jc/jcgm/`.

[25] Bich W., Cox M. G. and Harris P. M., *Metrologia*, **43** (2006) S161.

[26] Cox M. G. and Siebert B. R. L., *Metrologia*, **43** (2006) S178.

[27] Gibbons J. D. and Chakraborti S., *Nonparametric statistical inference*, 3d ed. (Marcel Dekker, New York, Basel, Hong Kong) 1992.

[28] Golub G. H. and Van Loan C. F., *Matrix Computations* (North Oxford Academic, Oxford) 1983.

[29] Weise K. and Woeger W., *Meas. Sci. Technol.*, **4** (1993) 1.

[30] Kacker R. and Jones A., *Metrologia*, **40** (2003) 235.

[31] Kacker R., *Metrologia*, **43** (2003) 1.

[32] Fishman G. S., *Monte Carlo* (Springer, New York) 1996.

[33] Manly B., *Randomization and Monte Carlo methods in Biology* (Chapman and Hall, London) 1991.

[34] Noreen E., *Computer intensive methods for testing hypotheses* (John Wiley & Sons, New York) 1989.

Electrical metrology

H. Bachmair

Physikalisch-Technische Bundesanstalt - Braunschweig, Germany

1. – Introduction

Electrical metrology covers a broad field which describes a wide arc from the realisation of the electrical units to their use in many different fields of application. It is the intent of this paper to provide an overview of the general status and capability of electrical metrology. It also intends to illustrate the scope of electrical subject areas, the complexity of individual activities and the relationships with the rest of metrology. And it outlines the main measurement philosophies and techniques and typical equipment for high-accuracy measurements. It will start with the base units of electrodynamics and the relationship between mechanical and electrical units and continue with the definition, realisation, reproduction, maintenance and dissemination of a unit shown by the example of the units of current, voltage and resistance. Impedance measurements, bridge and ratio techniques, AC/DC transfer and power and energy measurements confirm the broad range of application of electrical metrology and at the same time provide the basis for AC measurements of any kind. International comparisons guarantee the consistency of measurements world-wide and are the prerequisite for a mutual recognition of national standards.

2. – SI and electrical units

The *General Conference on Weights and Measures* (CGPM) adopted the *International System of Units* (SI) [1] in 1960. It consists of 7 base units, 4 of them forming the base units of electrodynamics (fig. 1), deduced from the *Giorgi* system and further developed to the *MKSA System* (m*eter*, k*ilogram*, s*econd*, a*mpere*) approved by the *In-*

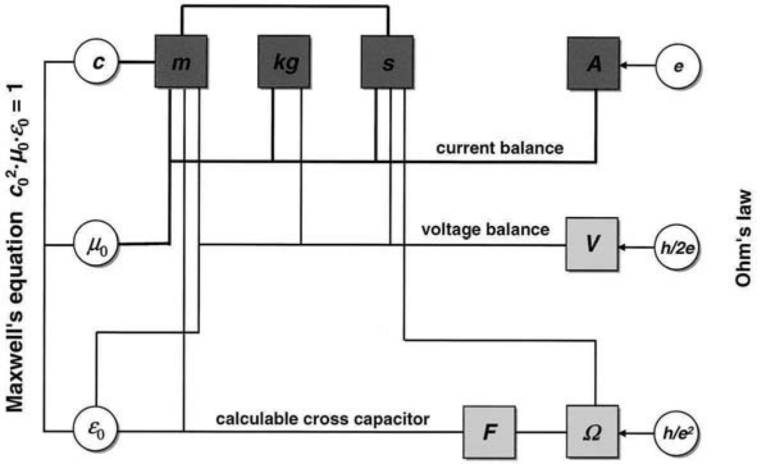

m
kg
s
A
c
e
current balance
voltage balance
V
h/2e
μ0
ε0
calculable cross capacitor
F
Ω
h/e²
Maxwell's equation $c_0^2 \cdot \mu_0 \cdot \varepsilon_0 = 1$
Ohm's law

Fig. 1. – The units of electrodynamics.

ternational Committee for Weights and Measures (CIPM) in 1946. All electrical and magnetic units can be expressed by means of these four base units of electrodynamics.

The *meter* is defined as the length of a path travelled by light in vacuum during a time interval of 1/299 792 458 of a second. This definition has the effect of defining the speed of light in vacuum as a fixed quantity. The *kilogram* is the unit of mass and is equal to the mass of the *International Prototype* of the kilogram. It is the only unit of the SI which is still defined by means of a material measure. The *second* is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the caesium-133 atom. The ampere is the only electrical base unit of the SI; its definition assumes equality of mechanical and electrical energy. The ampere definition has the effect of defining the permeability of the free space μ_0. It is realised by means of a so-called current balance. A current balance is an electrodynamic current-to-force transducer which compares an electrodynamically generated force with the force of a mass due to gravity.

The unit of voltage can be derived from the mechanical units meter, kilogram and second in connection with the permittivity of the free space ε_0. It is realised by means of a so-called voltage balance. This is an electrostatic voltage-to-force transducer which compares an electrostatically generated force with the force of a mass due to gravity. ε_0 can be derived from the velocity of light in vacuum c and the permeability of the free space μ_0 using *Maxwell's* equation which combines these three quantities. c, μ_0 and ε_0 are fundamental constants with fixed values without any measuring uncertainty.

The ohm can be realised by means of a calculable cross capacitor. For such a capacitor, the capacitance can be calculated from a single length and the permittivity ε_0 of the free space. The impedance of this capacitor $1/\omega C$ can be compared with a resistance R, the uncertainty being a few parts in 10^8. *Ohm's* law links the three quantities current,

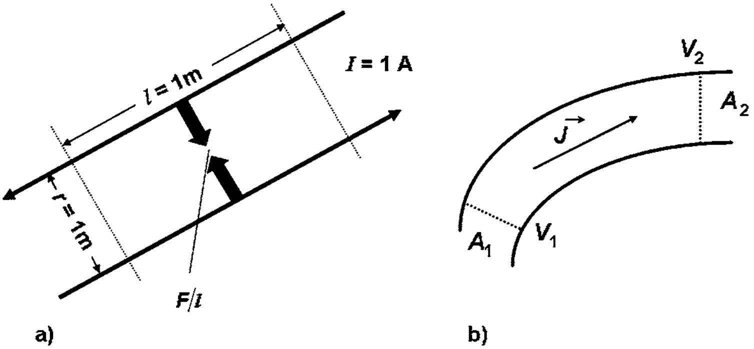


Fig. 2. – a) Definition of the ampere. b) Definition of the volt and the ohm.

voltage and resistance and therewith allows one of these units to be determined when the two others are known. It can also serve to control the experiments for the determination of the three units.

The units of voltage, resistance and current are reproduced, using macroscopic quantum effects. The equation of the *Josephson* effect links a voltage with the quotient $h/2e$ (e elementary charge, h *Planck's* constant) and a frequency. For the reproduction of the unit of resistance, the quantum *Hall* effect is used. It allows a resistance to be derived from the quotient h/e^2. The ampere can be reproduced by means of a controlled flow of single charge quanta. Counting single electrons one by one is still in an experimental stage, because the desired uncertainty cannot yet be achieved.

3. – Definition of a unit

The general definition of a unit is a physical experiment or artefact, based on well-established principles (*e.g.*, *Newton's* or *Maxwell's* laws). This means that the definition of a unit is its exact establishment with zero uncertainty. It always starts from ideal assumptions. The SI is a coherent system. In this context, the term "coherent" has a specific meaning: any unit should be related to the base units by only a multiplicative or divisive combination, *i.e.* the derived units are products of the base units with integer exponents and with a numerical prefactor of unity. For example, $1\,\mathrm{V} = 1\,\mathrm{m}^2\cdot\mathrm{kg}\cdot\mathrm{s}^{-3}\cdot\mathrm{A}^{-1}$, and $1\,\Omega = 1\,\mathrm{V}/1\,\mathrm{A} = 1\,\mathrm{m}^2\cdot\mathrm{kg}\cdot\mathrm{s}^{-3}\cdot\mathrm{A}^{-2}$.

At the 9th CGPM in 1948, a couple of definitions for electrical units were ratified, among them the definitions for the ampere, the volt, the ohm and the watt. The *ampere* is that constant current which, if maintained in two straight parallel conductors of infinite length of negligible circular cross-section, and placed 1 m apart in vacuum, would produce between these conductors a force equal to $2 \cdot 10^{-7}$ newton per metre of length (fig. 2a):

$$(1)\qquad \frac{F}{l} = \mu_0 \cdot \frac{I^2}{2\pi r} = 2 \cdot 10^{-7}\,\mathrm{N/m}, \quad \mu_0 = 2\pi \cdot \frac{F}{I^2} \cdot \frac{r}{l} = 4\pi \cdot 10^{-7}\,\mathrm{N/A}^2.$$

Looking at the definition of the ampere, terms like *straight parallel*, *infinite*, *negligible*, and *vacuum* show that the experiment described is defined under ideal physical conditions. The ampere definition has the effect of assigning an exact value to the magnetic field constant: $\mu_0 = 4{\cdot}\pi{\cdot}10^{-7}\,\mathrm{H/m}$. As the speed of light is determined by the meter definition, this also establishes ε_0 as a defined quantity. The underlying physical principle is the force between two current-carrying conductors, the direction depending on the direction of the two currents. The force is repulsive if the two currents flow in the same direction, and attractive if they flow in opposite directions as shown in fig. 2a.

The unit of potential difference and of electromotive force —the *volt*— is the potential difference between two points of a conducting wire carrying a constant current of 1 ampere when the power dissipated between these points is equal to 1 watt (fig. 2b). The power dissipated in a conductor between two cross-sections A_1 and A_2 can be calculated as the integral of the vector product of current density J and electric field strength E across the volume G of the conductor

$$P_J = \oiiint_G \vec{J} \cdot \vec{E} \, \mathrm{d}v = (V_1 - V_2) \cdot I. \tag{2}$$

After some modifications, the dissipated power is given by the potential difference $(V_1 - V_2)$ multiplied by the current I flowing through the conductor. This is exactly the definition of the volt.

The unit of electric resistance —the *ohm*— is the electric resistance between two points of a conductor when a constant potential difference of 1 volt, applied to these points, produces in the conductor a current of 1 ampere, the conductor not being the seat of any electromotive force (fig. 2b). The potential difference is the line integral of the vector of the electric field strength E between the two cross-sections A_1 and A_2:

$$V = V_1 - V_2 = \int_1^2 \vec{E} \, \mathrm{d}s = \int_1^2 \frac{\vec{J}}{\sigma} \, \mathrm{d}s = I \int_1^2 \frac{1}{\sigma(s_\mathrm{J}) \cdot A(s_\mathrm{J})} \, \mathrm{d}s_\mathrm{J}. \tag{3}$$

The general expression for the resistance is the integral of the inverse product of conductivity $\sigma(s_\mathrm{J})$ and cross-section $A(s_\mathrm{J})$ along the path of the wire. For a linear wire with constant cross-section and homogeneous conductivity, one gets the well-known expression $R = \ell/\sigma \cdot A$, with ℓ being the length of the wire, A the cross-section of the wire and σ its conductivity. In contrast to what one might expect from the ampere definition, the most commonly used units are the volt and the ohm, because it is much easier to maintain and compare voltages and resistances than currents.

4. – Realisation of a unit

In contrast to the definition, the realisation of a unit is always bound to an experiment with an uncertainty larger than zero, because experiments are never ideal. The difference between a theoretically designed experiment (definition) and its practical realisation is

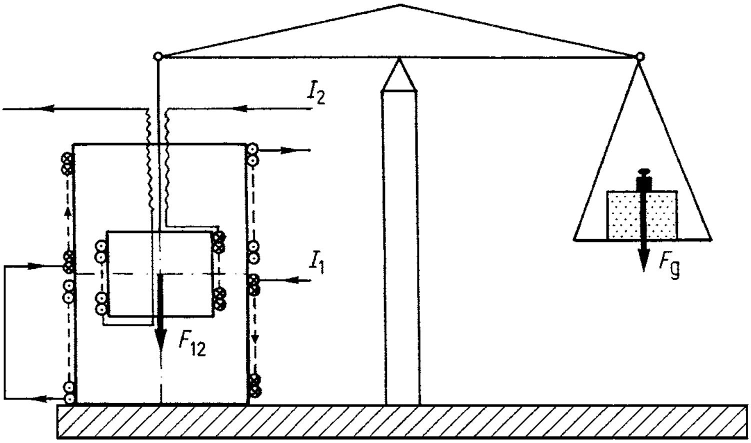


Fig. 3. – Principle of a current balance.

the reason why a distinction is made between the definition and realisation of a unit. Experiments for realising a unit are usually difficult, time-consuming and slow (sometimes decades long) to produce results. Therefore, they are normally only established and maintained at the National Metrology Institutes (NMI's).

A current balance [2] compares an electrodynamically generated force F_{12} with the force of a mass due to gravity F_g (fig. 3). F_{12} is generated by means of a coil system consisting of fixed and movable coils carrying currents I_1 and I_2. F_{12} is given as the product of the two currents multiplied by the partial derivative of the mutual inductance L_{12} in the direction of the force. At equilibrium,

$$I_1 \cdot I_2 \cdot \partial L_{12}/\partial z = m \cdot g \tag{4}$$

is obtained. If the same current flows through both coils, the current is given by the square root of the mass m multiplied by the local acceleration due to gravity g divided by $\partial L_{12}/\partial z$. As these quantities can be determined in SI units, the ampere can also be determined in SI units. $\partial L_{12}/\partial z$ depends in a complicated way on the dimensions of the coils and their position to each other. Therefore, the uncertainty of the classical current balances was limited to a few parts in 10^6.

In 1975, Kibble [3] proposed a new type of current balance, which overcomes the problems with the coil dimensions. The experiment is performed in two parts: the weighing part (static mode, fig. 4a) and the moving part (dynamic mode, fig. 4b). In the static mode, the force F of a magnetic field with flux density B on a coil carrying a current I is compared with the force of a mass due to gravity.

$$F = I \cdot G(B, l) = mg. \tag{5}$$

The current is measured as the voltage drop across a resistor using the *Josephson* and

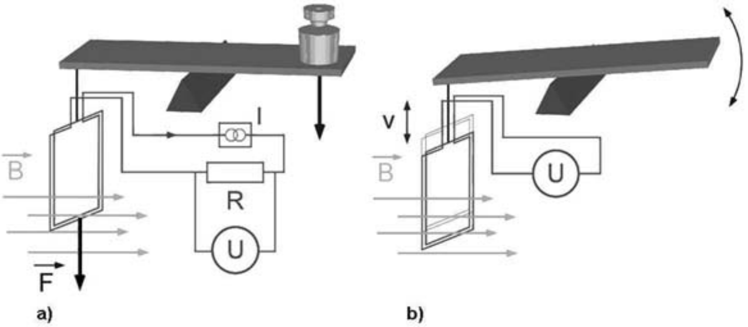


Fig. 4. – Principle of a watt balance: a) weighing part, b) moving part.

quantum *Hall* effects. The function $G(B, \ell)$, ℓ being the width of the coil exposed to the magnetic field, is determined in the dynamic part of the experiment. In this part of the experiment, the coil moves slowly with a velocity v through the magnetic field B around the centre position of the coil during the weighing part of the experiment. The voltage induced in the coil is proportional to the velocity v and the function $G(B, \ell)$ which can thereby be determined in terms of the induced voltage and the velocity,

$$U = v \cdot G(B, l). \tag{6}$$

The velocity can be measured with high accuracy by means of a laser interferometer, and the induced voltage is measured against a *Josephson* voltage standard. Combining the two parts of the experiment, one gets two expressions for the function $G(B, \ell)$, one from the weighing part (static mode, mg/I) and one from the moving part (dynamic mode, U/v). By remodelling of the equation, one obtains an expression for the electrical power $U \cdot I$ on the left-hand side of the equation and for the mechanical power $m \cdot g \cdot v$ on the right-hand side of the equation. Therefore, this balance is called a watt balance. If voltage U and current I are measured by means of quantum standards, m becomes proportional to the *Planck* constant h.

World-wide, there are several experiments in operation or in a status of being established. The first result was obtained with the NPL watt balance which furnished an uncertainty of $2 \cdot 10^{-7}$. The uncertainty of the NIST watt balance is now approaching $2 \cdot 10^{-8}$ and therewith is at present the most accurate experiment [4]. The watt balance of METAS is in operation and about to furnish first results. LNE is assembling its watt balance, and a test set-up is presently established at BIPM so that in the near future 5 independent experiments will be available. Due to the very small uncertainty, watt balances are top candidates for a possible redefinition of the unit of mass, the kilogram, by means of a fixed value for the *Planck* constant h.

Besides the current balances, voltage balances have been established. They are electrostatic voltage-to-force converters and allow a voltage to be determined directly in SI units [5]. A change in electrical energy is compared with the corresponding change in

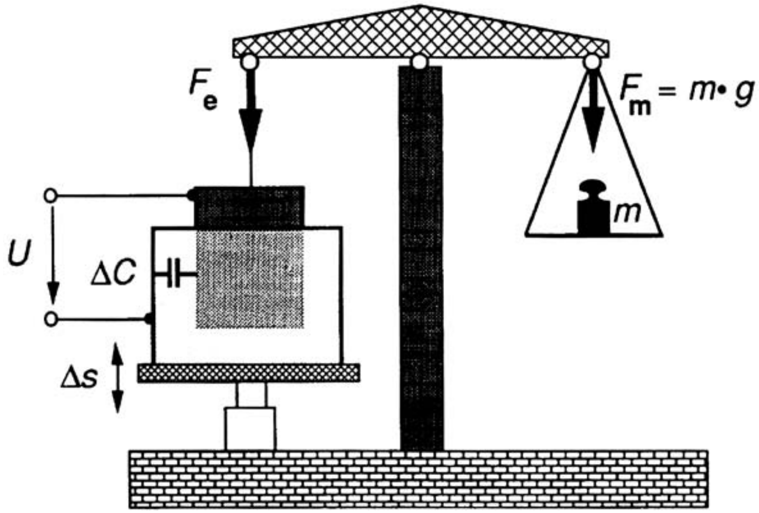


Fig. 5. – Principle of a voltage balance.

mechanical energy (fig. 5). The energy stored in a capacitor is changed by moving one of the electrodes by a distance Δs in the direction of the force F_{e}, $\Delta W_{\mathrm{e}} = -1/2 \cdot U^2 \cdot \Delta C$. The corresponding change in mechanical energy amounts to $\Delta W_{\mathrm{m}} = m \cdot g \cdot \Delta s$. In equilibrium, ΔW_{e} is equal and opposite to ΔW_{m}, and the voltage can be calculated directly in SI units. The various types of voltage balances differ from their voltage-to-force transducers: a parallel-plate capacitor where the force is proportional to the distance s of the plates and an immersion capacitor as in fig. 5 where the generated force is nearly constant, independent of the position of the moving electrode. In the eighties of the last century, four voltage balances had been in operation at LNE (the former LCIE), NMIA (former CSIRO NML), PTB and the University of Zagreb (Croatia). At that time, the uncertainty was comparable to the uncertainty obtained with watt balances. Meanwhile, voltage balances are no longer in operation, because the uncertainties achieved with watt balances are by one order of magnitude smaller that those of the voltage balances.

A realisation of the ohm as given by its definition is not possible, because a resistance cannot be calculated from its dimensions and the conductivity of the material with the required low uncertainty. However, the Australian scientists *Thompson* and *Lampard* discovered a new theorem of electrostatics [6] which allows the capacitance of a capacitor to be calculated independent of its cross section only from the length of the electrode system. The theorem in its general form allows the capacitance per unit length of an indefinite long cylinder of arbitrary cross section to be calculated which is divided into four parts by means of infinitesimally small gaps (fig. 6a). The cross capacitances C_1' and C_2' per unit length of opposite parts of the cylinder are given by

$$e^{-\pi C_1'/\varepsilon_0} + e^{-\pi C_2'/\varepsilon_0} = 1. \tag{7}$$

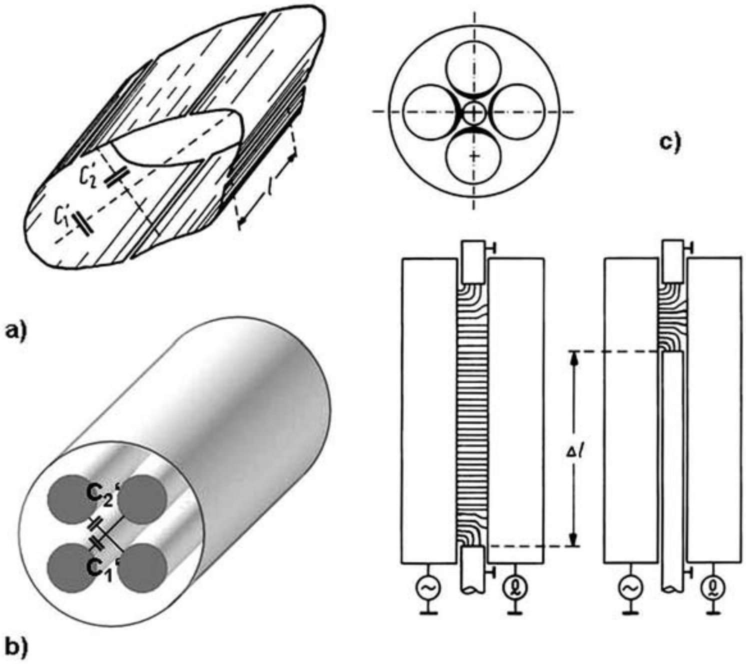


Fig. 6. – Calculable cross capacitor: a) theorem, b) realisation, c) electrode arrangement.

In practice, a symmetrical system in form of 4 cylinders arranged at the corners of a square is used with equal cross capacitances (fig. 6b). This simplifies the relation for the cross capacitance:

$$C_0' = \frac{\varepsilon_0}{\pi} \cdot \ln 2 \approx 2\, pF/m, \quad \text{with } \mathrm{C}_1' \approx \mathrm{C}_2' \approx \mathrm{C}_0'. \tag{8}$$

As the electrostatic field does not only expand in the inner of the electrode system but also outside, it must be shielded by means of an electrostatic screen which surrounds the electrode system.

In practice, two problems must be solved: the infinitesimally small gaps between the electrodes and the limitation of the length of the electrode system. With a geometry of the capacitor as shown in fig. 6b, the feed-through of the electrostatic field becomes negligibly small even with gap widths in the order of a millimetre. This is true for both, the gaps between adjacent electrodes and between the electrodes and the screen. The so-called guard electrodes can limit the length of the capacitor (fig. 6c). In the vicinity of the guard electrodes, the electric field is distorted (a three-dimensional field instead of a two-dimensional one) and does no longer fulfil the theorem. The whole length of the capacitor can be divided into three parts, two inhomogeneous parts around the guard electrodes and a homogeneous part in the middle of the capacitor. By moving one of the

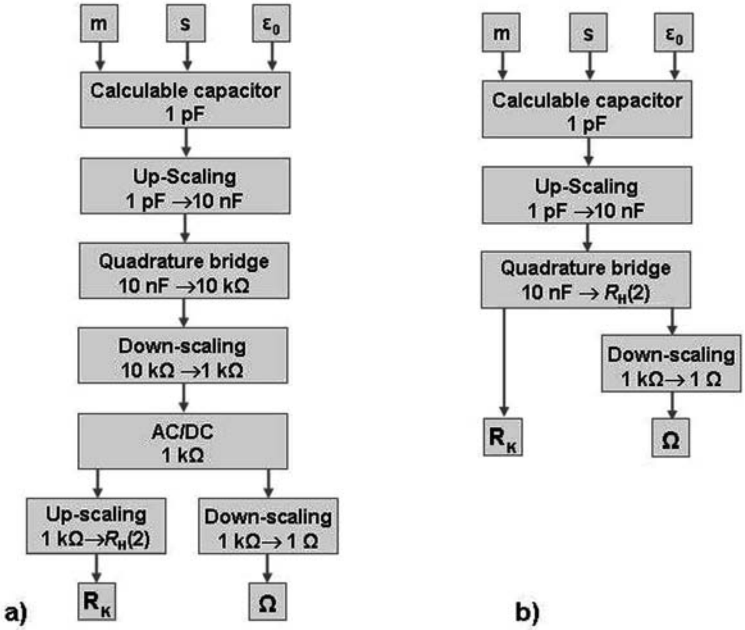


Fig. 7. – Determination of the unit of resistance from the unit of capacitance: a) conventional scheme, b) scheme using the quantised *Hall* resistance.

guard electrodes on a straight path without changing the geometry between the guard electrodes and the main electrodes, one cuts a length $\Delta\ell$ out of the homogenous part of the field which corresponds to a change in capacitance of

$$\Delta C_0 = \frac{\varepsilon_0}{\pi} \cdot \ln 2 \cdot \Delta l. \tag{9}$$

This method has another advantage, because a shift in length can be measured by means of a laser interferometer with a much larger accuracy than an absolute length. Calculable cross capacitors are operated at several NMI's and allow the unit of capacitance to be determined with an uncertainty of about $2\cdot 10^{-8}$. New improved capacitors are presently built at NMIA, Australia; NRC, Canada and at the BIPM.

By means of a couple of AC bridges and DC comparators, the as-maintained unit of resistance can be traced back to the unit of capacitance (fig. 7a), and at the same time, the value of the *von Klitzing* constant can be determined in SI units [7]. With cross capacitors, capacitances in the order of 0.2 to 1 pF can be determined. This value must be scaled-up to 1 nF or 10 nF before a comparison with a 10 kΩ or 100 kΩ AC resistance can be made in a special bridge (quadrature bridge). The AC resistance must be scaled-down to 1 kΩ, a level where the most accurate AC/DC resistors with calculable AC/DC characteristic are available. The last step in the whole chain is to scale-down the DC resistance to the as-maintained unit of resistance or to scale it up to determine a quantum

Hall resistance at the second plateau (12.9 kΩ) in SI units. The AC measurements are performed at an angular frequency of 10^4 Hz which corresponds to a frequency of 1.6 kHz. For some time, a second path has been opened (fig. 7b) using a quantum *Hall* resistance as AC/DC transfer resistance. This shortens the chain and allows a direct link to the *von Klitzing* constant. Presently, the quantum *Hall* resistance agrees at AC and DC within about 2 parts in 10^8.

5. – Reproduction of a unit

Since about 1970, quantum-physical experiments have been available for the reproduction of the electrical units, especially those which involve macroscopic quantum effects like the *Josephson* and quantum *Hall* effects and single electron tunnelling. These effects allow a reproduction of the electrical units based on fundamental constants which are believed to be constant in time and space, and, therewith, the macroscopic quantum phenomena allow these units to be reproduced at any place at any time. There are two macroscopic quantum effects, the *Josephson* effect and the quantum *Hall* effect which allow the units of voltage and resistance to be reproduced with an uncertainty much smaller than the uncertainty with which these units can be determined in SI units. The ampere can be reproduced by means of a transfer of single charges with a defined frequency f. The realisation of this effect is still in an experimental stage. While the *Josephson* and quantum *Hall* effects allow voltages of the order of 1 V to 10 V and resistances of about 10 kΩ to be reproduced, the currents generated by the transfer of single electrons are of the order of pA and cannot be reproduced with the desired uncertainty.

In 1962, *Josephson* predicted a quantum-mechanical effect which occurs on two weakly coupled superconductors separated by a thin normally conducting or insulating barrier, called a *Josephson* junction [8]. If such a *Josephson* junction is biased at a non-zero DC voltage, the supercurrent across the junction oscillates with a frequency proportional to that voltage. By phase locking the *Josephson* oscillator to an external frequency-stabilised microwave source, constant voltage steps are generated in the DC characteristic of the *Josephson* junction, given by

$$V_{\mathrm{J}}(n) = \frac{n \cdot f}{K_{\mathrm{J}}} \equiv n \cdot \frac{h}{2e} \cdot f \qquad \text{with} \qquad K_{\mathrm{J}} \equiv \frac{2e}{h}, \tag{10}$$

with $h/2e$ being the inverse of the *Josephson* constant K_{J}, f the frequency of the irradiated microwave and n the step number of the plateau in the *Josephson* characteristic. At a frequency of about 70 GHz, the voltage at the first step is about 145 μV, adjustable to the same resolution as the frequency. The *Josephson* effect is a very universal effect [9]; its voltage is independent of the material of the superconductors, the type of the *Josephson* junction and its geometry, the temperature, the frequency of the microwave and the irradiated microwave power. As a voltage can be reproduced by means of the *Josephson* effect with an uncertainty of at least two orders of magnitude smaller than the uncertainty with which the unit of voltage is known in SI units, the *Consultative Committee for Electricity and Magnetism* (CCEM) recommended, and the CIPM decided, to fix a

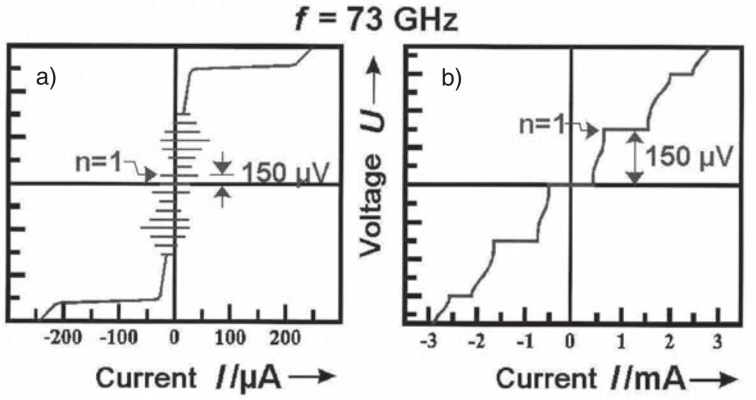


Fig. 8. – Characteristic of a *Josephson* junction: a) underdamped junction, b) highly damped junction.

numerical value for the *Josephson* constant $K_{\text{J-90}} = 483\,597.9\,\text{GHz/V}$ to make use of the small uncertainty for comparison and calibration purposes [10]. The value was chosen to be in as close an agreement as possible with the SI value and was put into force on 1 January 1990.

Today, large arrays of multilayer *Josephson* junctions are manufactured with up to 250000 junctions connected in series for applications as DC or programmable voltage standards. For completely underdamped SIS tunnel junctions, constant voltage steps are generated in the sub-gap part of the quasiparticle branch of the DC characteristic and, therefore, cross the voltage axis (fig. 8a). This allows the standard to be operated at zero bias current. The main advantage of the SIS arrays is the high output voltage which allows a voltage of 10 V to be generated with less than 20000 junctions at a microwave frequency of 70 GHz. A disadvantage is that the output voltage is multi-valued. Together with the relatively small current width of the step voltage this lead to stability problems as regards the adjusted DC voltage. Arrays with highly damped *Josephson* junctions can be rapidly switched from one voltage step to another and are therefore best suited for programmable *Josephson* voltage standards (fig. 8b). As they are only switched between zero and the first step, a larger number of junctions is required to obtain a specific output voltage, *i.e.* about 80 000 junctions for 10 V at a microwave frequency of 70 GHz.

In 1980, *von Klitzing* discovered a quantum-mechanical effect, which occurs when a two-dimensional electron gas (2DEG) in a semiconducting device is exposed to a strong traverse magnetic field at very low temperatures [11]. Depending on the gate voltage (Si-MOSFET) or the flux density (heterostructures), the *Hall* resistance shows plateaux at integer fractions of the quotient h/e^2,

$$R_{\text{H}}(i) = \frac{1}{i} \cdot R_{\text{K}} \equiv \frac{1}{i} \cdot \frac{h}{e^2} \qquad \text{with} \qquad R_{\text{K}} \equiv \frac{h}{e^2}\,, \tag{11}$$

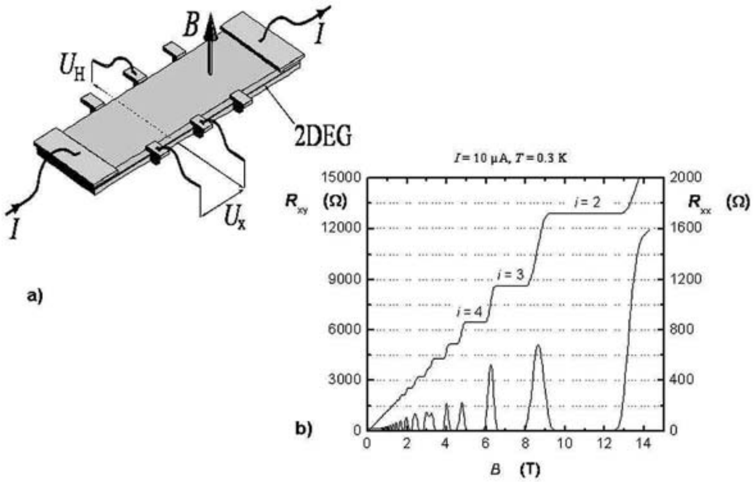


Fig. 9. – a) Quantum *Hall* probe (heterostructure). b) *Hall* resistance and longitudinal resistance as a function of flux density.

with h/e^2 being the *von Klitzing* constant R_K and i an integer which denotes the plateau number. The quantum *Hall* resistance is to a large extent independent of the device, the semiconducting material and the width of the sample, the plateau number and the direction of the magnetic field [12]. On the first plateau, the *Hall* resistance amounts to 25.8 kΩ. As a quantised *Hall* resistance can be reproduced by means of the quantum *Hall* effect with an uncertainty of about one order of magnitude smaller than the uncertainty with which the unit of resistance is known in SI units, the CCEM recommended, and the CIPM decided, to fix a numerical value for the *von Klitzing* constant $R_{\text{K-90}} =$ 25 812.807 Ω to make use of the small uncertainty for comparison and calibration purposes. The value was chosen to be in as close an agreement as possible with the SI value.

The 2DEG of a semiconducting device is located in the inversion layer of the semiconductor device, formed at the interface between a semiconductor and an insulator or two semiconductors, one of them playing the role of the insulator. Figure 9a shows a quantum Hall probe as is used today by many NMI's. The light grey area forms a GaAs heterostructure in which the 2 DEG is embedded at the interface of the GaAs and the AlGaAs layers. It is supplied with a specific number of ohmic contacts, two for the injection of the current and six in total for monitoring of the *Hall* voltage U_H and the longitudinal voltage U_X. Heterostructures show clearly distinct plateaux as can be seen from the characteristic in fig. 9b. It shows the quantised *Hall* resistance R_{xy} and the longitudinal resistance R_{xx} as a function of the flux density. At the plateaux, the longitudinal resistance vanishes which is a necessary condition for the transversal resistance which is exactly the quantised *Hall* resistance.

In ultra-small metallic tunnel contact circuits, tunnelling of single electrons can be observed as a macroscopically observable effect if their *Coulomb* interaction dominates

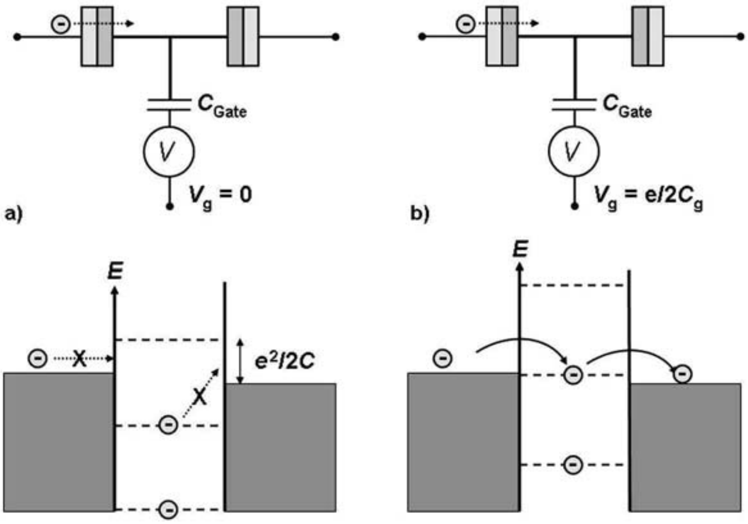


Fig. 10. – SET transistor: a) electron tunnelling blocked, b) electron tunnelling allowed.

on the energy scale [13]. This means that the effect can only be observed at structures with nm-dimensions and at very low temperatures in the mK-range. The resulting single charge effect allows a controlled manipulation of single charge quanta and therewith opens up unique applications in electric metrology. A single electron circuit consists of an ultra-small conducting island which is connected to two electron reservoirs by means of two tunnel barriers. The tunnelling of one electron onto the island requires an energy of

$$\Delta E = E_\mathrm{C} = \frac{e^2}{2C} \gg k_\mathrm{B} \cdot T \tag{12}$$

which is equal to the increase in the *Coulomb* energy of the island, as long as the temperature is small enough. Note that small dimensions mean a small capacitance C and therewith a high *Coulomb* energy E_C. Typical values are 100 nm for the dimensions of the tunnel barrier and 1 fF for the effective capacitance. The resulting *Coulomb* energy corresponds to a thermal energy at a temperature of 1 K so that the experiment must be performed at temperatures below 100 mK, preferably 10 mK.

As long as the *Coulomb* energy is much larger than the thermal energy, tunnelling of an electron from the reservoir onto the island is suppressed for energetic reasons as long as the gate potential equals zero (fig. 10a). At the same time, an electron cannot leave the island due to the potential barrier between the island and the reservoir. By means of a capacitively coupled gate electrode, the electrostatic potential of the island can be continuously varied so that for certain values of the gate voltage an electron can be transferred to the island or leave the island (fig. 10b). Such a SET transistor works similar to a field effect transistor (FET), whereby the conductance is modulated by the gate charge. This allows a SET transistor to be operated as an extremely sensitive electrometer with a resolution of a fraction of an electron.

Just as well as *Ohm's* law is valid in the macroscopic world, it can be used to check the three quantum effects in the microscopic world. These combined experiments are called "closing the metrological triangle". If the triangle closes within the combined uncertainty of the three quantum standards, there is good confidence in the three experiments. It must be tested whether the voltage drop across a quantised *Hall* resistance of a current generated by means of the SET effect is equal to a voltage generated by means of the *Josephson* effect,

$$n \cdot \frac{h}{2e} \cdot f_U = e \cdot f_I \cdot \frac{1}{i} \cdot \frac{h}{e^2} \qquad \text{but} \qquad U_{\mathrm{J}} = R_{\mathrm{K}} \cdot I_{\mathrm{SET}} \approx 40\,\mathrm{nV}\,! \tag{13}$$

The disadvantage of this combined experiment is the low signal level which can hardly be measured with an uncertainty in the order of $1 \cdot 10^{-8}$. To solve this problem, the current can be scaled up by means of a cryogenic current comparator. But even with a current ratio of 10 000 : 1, the voltage becomes only $400\,\mu\mathrm{V}$ and is still too small to be measured against a *Josephson* voltage with the desired uncertainty.

Another and presently more promising solution is to extend the triangle to charge a capacitor using the relation $Q = C \cdot U$. The charge can be determined by counting electrons one by one, so that $Q = n \cdot e$. As has already been shown, a quantised *Hall* resistance can be compared with a capacitance with very small uncertainty so that C is given by $C = i \cdot e^2/\omega \cdot h$. Therewith the equation reads:

$$N \cdot e = \frac{i \cdot e^2}{\omega \cdot h} \cdot n \frac{h}{2e} f_{\mathrm{J}} \qquad U_{\mathrm{J}} = (1\,\mathrm{pA} \cdot 1\,\mathrm{s})/1\,\mathrm{pF} = 1\,\mathrm{V}. \tag{14}$$

Assuming a current of 1 pA which charges a capacitor of 1 pF over a time period of 1 s, a voltage of 1 V is generated which can easily be compared with a *Josephson* voltage with the desired uncertainty. These experiments have also their bottlenecks: The capacitor must be operated at very low temperatures in a cryostat and is charged with a few million electrons, none of which must be lost! During the charging process, the voltage across the capacitor increases linearly, while a sinusoidal voltage is used to compare it with R_{K}. Therefore, the frequency behaviour of the capacitance must be exactly known.

6. – Maintenance and dissemination of a unit

Standard cells, *Zener* references and standard resistors are widely used to maintain and disseminate the units of voltage and resistance. As there is no current standard available for maintenance of the ampere, this unit is maintained and disseminated by a combination of voltage and resistance standards. For most exact applications, the secondary standards are directly compared with the corresponding quantum standards.

Until 1972, the members of the *Meter Convention* regularly compared their as-maintained units of voltage with the unit kept at the BIPM to get an idea of the differences between their as maintained units (fig. 11). The national units were realised by a bank of standard cells and corrected from time to time when the comparisons showed

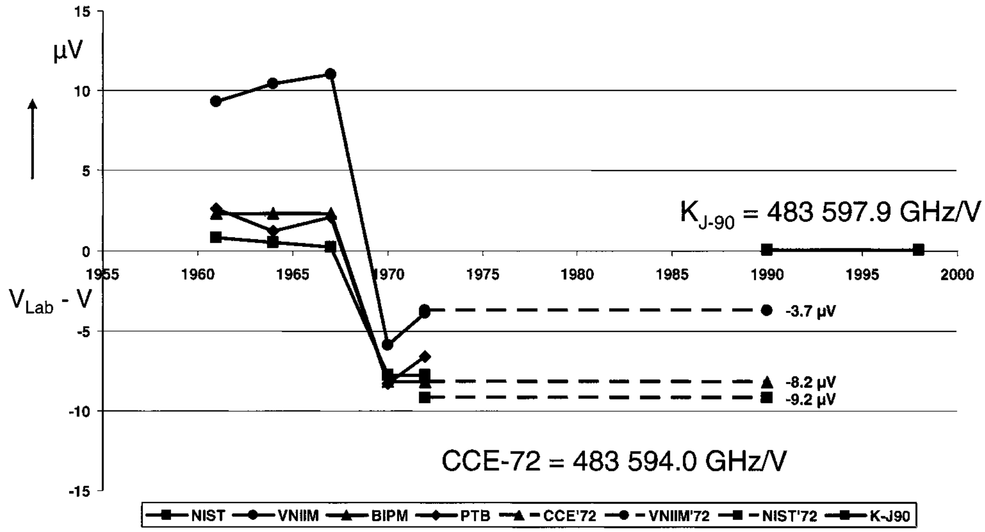


Fig. 11. – As-maintained unit of voltage before and after fixing a value for the *Josephson* constant.

that the drift of the as-maintained units was too large. In 1970, most of the NMIs, BIPM included, corrected their as-maintained units by about $10 \cdot 10^{-6}$ following the latest results obtained from the least squares adjustment of CODATA which itself was based on the results of current balance experiments performed at different laboratories. In 1972, the CCE (now CCEM) recommended to use a value for $2e/h = 483\ 594.0\,\text{GHz/V}$ for comparison purposes, because at that time several NMIs already run a *Josephson* voltage standard. Unfortunately, not only one value was fixed but three, because NIST and VNIIM choose there own values corresponding to their as-maintained units. In 1990, $K_{\text{J-90}}$ was introduced which forced the NMIs to change their as-maintained units again by about 8 ppm in the opposite direction, so that they came close to the value they already had before 1970. All members of the *Meter Convention* accepted $K_{\text{J-90}}$. Use of the same conventional value for the *Josephson* constant considerably improved the maintenance of the national units.

Standard cells and *Zener* references are the most commonly used voltage standards. Standard cells have an output voltage of 1.018 V with an internal resistance which ranges from $200\,\Omega$ to $1200\,\Omega$. The temperature coefficient is approximately $40 \cdot 10^{-6}$/K and is formed by the difference of the temperature coefficients of the positive and negative poles which are about ten times as large. Standards cells are characterised by a very low noise ($4\,\text{nV}/\sqrt{\text{Hz}}$) and a good long-term stability ($10^{-7}/\text{a} \ldots 10^{-6}/\text{a}$). They must, however, be handled with extreme care, because they are sensitive to loading as well as to shock and vibration. *Zener* references have output voltages in a range from 1 V to 10 V with a very small internal resistance of a few mΩ at 10 V and 1 kΩ at the 1 V level. The temperature coefficient ranges from $5 \cdot 10^{-7}$ to $5 \cdot 10^{-8}$ at 10 V, and $1 \cdot 10^{-6}$ at the 1 V level. The

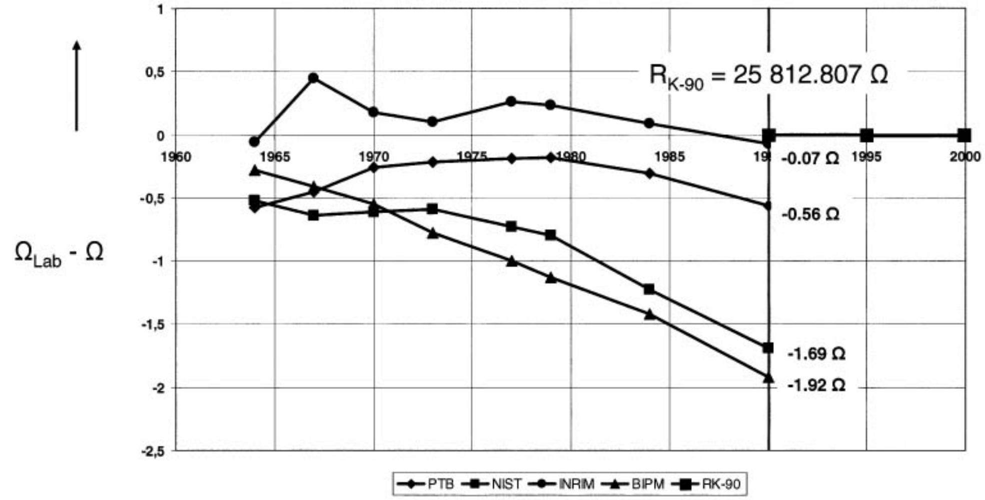


Fig. 12. – As-maintained unit of resistance before and after fixing a value for the *von Klitzing* constant.

output noise of *Zener* references is about ten times higher than that of standards cells ($30\ldots 50\,\text{nV}/\sqrt{\text{Hz}}$). The long-term stability is comparable with standard cells (10^{-6}/a). Due to the low internal resistance, *Zener* references are insensitive to loading. Due to the sensitivity of standard cells and the robustness of *Zener* standards, nowadays mostly *Zener* references are used. The most stable output is the 10 V output which is directly derived from the output amplifier of the *Zener* diode. The 1 V and 1.018 V outputs are not as stable as the 10 V output, because both are scaled down from the 10 V level by means of resistive dividers which are characterised by an additional drift. It is, therefore, not unusual when different output voltages show different drift characteristics.

Besides the as-maintained units of voltage, the members of the *Meter Convention* regularly compared their as-maintained units of resistance (fig. 12) with the unit kept at the BIPM to get an idea of the differences between the different as-maintained units. They were realised by a bank of standard resistors and corrected from time to time when the comparisons showed too large a drift of the as-maintained units. In 1990, $R_{\text{K-90}}$ was introduced which forced the NMIs to change their as-maintained units by different amounts so that they came closer to the value of the SI ohm. All the members of the *Meter Convention* accepted $R_{\text{K-90}}$ and thus considerably improved the maintenance of the national units.

Secondary resistance standards are mostly single element resistors of different kind which depend on their value and range of application. A group of Thomas type-1 Ω resistors is used by many laboratories to maintain the unit of resistance. A standard resistor is expected to have an excellent long-term stability, small temperature, pressure and humidity coefficients and a small thermal emf to copper. Widely used resistance alloys are

Zeranin, *Manganin*, *Evanohm* and *Isaohm*. Their temperature coefficient is small and can be positive or negative, depending on the heat treatment during artificial ageing and the characteristic of the resistivity as a function of temperature. To insulate the resistive element against its housing, insulators like polyethylene, polystyrene, *Teflon*, sapphire or quartz are used. Note that the ratio of resistivity of insulators to conductors is 10^{24}!

Why are resistance measurements so important for metrology? Besides voltage, resistance is one of the few electrical quantities which can easily be realised and measured. The calibration range extends over 21 decades from $10\,\mu\Omega$ to $10\,\mathrm{P}\Omega$. It is this enormous dynamic resistance range, available even in common objects, which makes resistance so useful as an indicator of other physical parameters, *i.e.* temperature, force and pressure. Resistors are most widely used in sensor applications, *i.e.* strain gauges and thermometers. Resistance is the electrical quantity most widely used by other disciplines.

Generally, a resistance is defined by the voltage drop V caused by a current I flowing through the resistor, similar to the definition of the ohm. The ratio of an unknown resistor R_{X} and a standard resistor R_{S} is therefore determined by the voltage ratio of the voltages across the resistors and the inverse current ratio of the currents flowing through the resistors. Resistors are rarely measured by accurately measuring the voltage and current. Usually, we use a potentiometer, a bridge or a current comparator to scale from a known resistance to an unknown resistance. Potentiometers, bridges or current comparators convert the accurate measurement of the voltage and the current into the ratio of two voltages, resistances or currents. If the same current flows through both resistances, we are talking about a potentiometric method. It is especially suited for medium and high-ohm resistances. The resistance ratio is proportional to the voltage ratio. If the voltage drop across the two resistors is the same, we are talking about a comparator method. It is mostly used for medium and low-ohm resistances. The resistance ratio is proportional to the inverse current ratio or proportional to the ratio of the number of turns of the two windings of the current comparator.

7. – Impedance measurements

There is a lot of confusion and intimidation about impedance measurements, the different types of connectors and the measurement configurations. Before an impedance of any kind —resistance, capacitance, inductance— can be measured properly, it must be clearly defined by fixing the boundary conditions, *i.e.* the potentials and currents at its ports [14]. The starting point always is a 2-terminal impedance, as resistors, capacitors and inductors may internally be considered as 2-terminal devices. A 2-terminal impedance is defined by the voltage across and the current through those 2 terminals (fig. 13a). Influence parameters affecting the internal impedance, such as temperature, pressure, humidity and others, are generally related to the choice of materials and the mechanical design and can usually be improved only by improving the environmental conditions, especially their stability with time. Electrical parameters like dissipated power (self-heating!), voltage, frequency, and external electric or magnetic fields may also have a significant effect on the accuracy and repeatability of an impedance measurement.

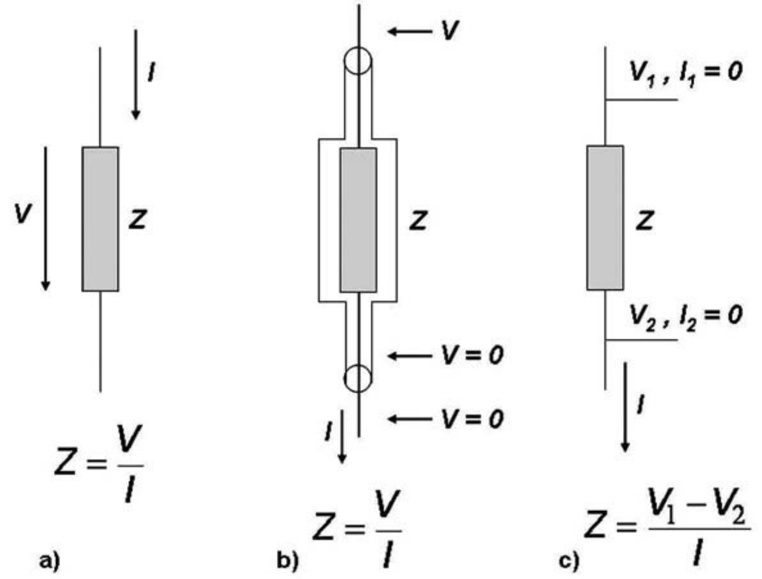


Fig. 13. – Definition of a 2-terminal (a)), 3-terminal (b)) and 4-terminal (c)) impedance.

As high-value resistors, capacitors and inductors are very susceptible to external currents or electromagnetic fields, they will be normally screened to protect them against these environmental influences. Conductive shields are used to protect high-value resistors and capacitances against external currents and electric fields. Magnetic shields are used for the protection against magnetic fields; sometimes they are used in combination with electric shields. The shields are normally connected through a low impedance to a fixed potential, preferably zero potential and will therewith eliminate any external currents and stabilise the internal leakage currents of an impedance. A 3-terminal impedance is defined as the voltage across the impedance divided by the current out of the lower potential terminal (fig. 13b).

The 4-terminal configuration is used for most accurate impedance measurements. Low-value impedances suffer from poorly defined potentials, especially if current is flowing along the leads that are measuring the potential. As a solution, the potential is measured with leads that carry no current. The potential junctions must be defined with low impedance invariant to the current flowing through the impedance. A 4-terminal impedance is defined as the potential drop between the high and low potential terminals divided by the current out of the low current terminal (fig. 13c). 4-terminal impedances are commonly used for resistance measurements of less than 100 kΩ.

2-terminal pair impedances are most appropriate for reactance standards. They are defined as the voltage at the high terminal divided by the current out of the low terminal when the low potential is zero (fig. 14). As the current of the inner conductor is forced to flow back on the outer conductor, the impedance of a 2-terminal pair is the

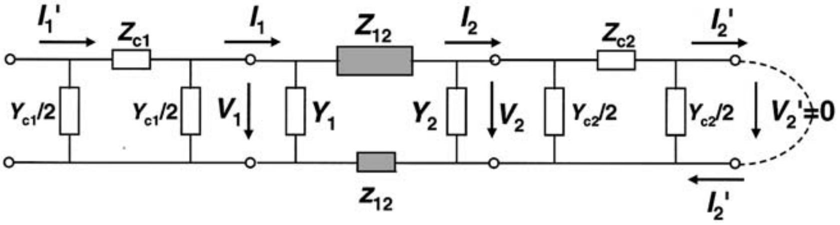


Fig. 14. – Equivalent circuit of a 2-terminal pair impedance with connecting leads.

sum of the main impedance Z_{12} between the terminals and the small impedance z_{12} of the outer. External electromagnetic fields do not influence the impedance significantly. Another advantage of 2-terminal pair impedances is that connecting cables have only little influence on the impedance which can be corrected if the impedances of the cables are known. Z_{C1} together with the admittances Y_{C1}, Y_1 and $1/Z_{12}$ then form a voltage divider at the high terminal which causes a small difference between the input voltage V_1' and the voltage V_1 at the high terminal of Z_{12}. Z_{C2} together with Y_2 and Y_{C2} form a current divider which causes a difference between the output current I_2' and the current I_2 out of the low terminal of Z_{12}.

$$(15) \qquad Z_{12}' = \frac{V_1'}{I_2'} = Z_{12}\left[1 + Z_{\mathrm{C1}}\left(Y_{\mathrm{C1}}/2 + Y_1 + 1/Z_{12}\right)\right] \cdot \left[1 + Z_{\mathrm{C2}}\left(Y_2 + Y_{\mathrm{C2}}/2\right)\right].$$

With coaxial cables 1 m in length, and at an angular frequency of $\omega = 2\pi\,\mathrm{f} = 10^4\,1/\mathrm{s}$, the correction term will be smaller than $1 \cdot 10^{-8}$. A 10 pF capacitance can be measured with an uncertainty of $1 \cdot 10^{-8}$ with a cable 1 m in length which has a capacitance of 100 pF between the inner and the outer conductor.

Combining the concept of shielding with the concept of a 4 terminal impedance, a 4-terminal pair impedance is obtained which is suited for the most accurate measurements (fig. 15). At improved potential definitions, 4-terminal pair impedances are characterised by perfect electrostatic shielding and stable internal leakages to the shield. They are used for resistances below 100 kΩ and capacitances larger than 100 pF where highest accuracy is important. The defining points for the potential across a 4-terminal pair impedance are ports 2 and 4, provided that the current at these ports is zero. The defining point for the current is port 3. As the defining points are not at the four terminal junctions of the inner conductors of Z_{H}', the apparent impedance Z_{H} will differ from Z_{H}'.

$$(16) \qquad Z_{\mathrm{H}} = \frac{V_2}{I_3}\bigg| I_2 = 0, I_4 = 0, V_4 = 0 = Z_{\mathrm{H}}' \frac{1}{(1 + Z_2 Y_2/2) \cdot (1 + Z_3 Y_3/2)}.$$

This is referred to as the lead correction for the 4-terminal pair impedance. This correction accounts for the voltage divider at terminal 2 (Z_2, $Y_2/2$) and the current divider at terminal 3 (Z_3, $Y_3/2$).

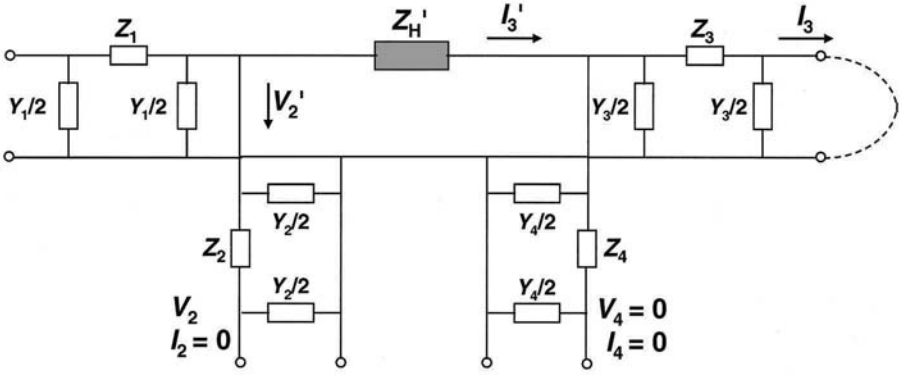


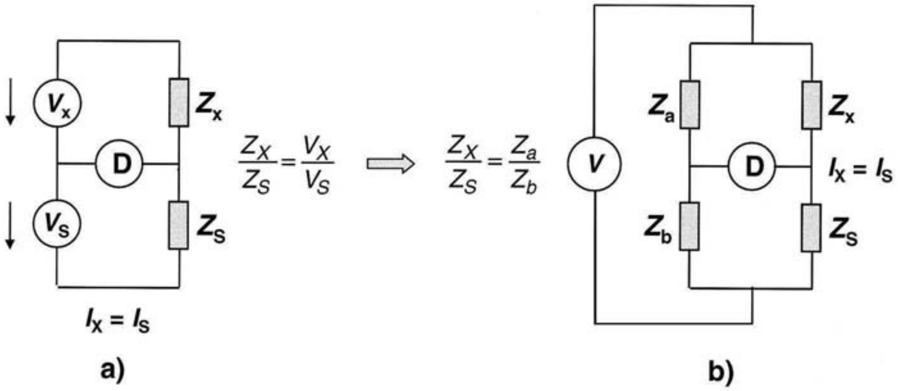
Fig. 15. – Equivalent circuit of a 4-terminal pair impedance.

8. – Bridge and ratio techniques

There are common concepts which reoccur in many impedance bridges. Understanding of these concepts can help to mentally translate an intimidating and confusing web of networks into just a simple bridge with a few wires. Bridge and ratio techniques are used to compare impedances of the same kind or of different kinds [14]. Impedances are rarely measured by precise voltage and current measurements. Usually, a bridge will be used to scale from a known impedance to an unknown impedance. Bridges convert the accurate voltage measurement into the ratio of two voltages or two known impedances, respectively. This has other advantages: one single detector instead of two, a null detector instead of a linearly calibrated detector and a low source output impedance and, thus, lower noise. The *Wheatstone* bridge (fig. 16b) is a variation of a potentiometer (fig. 16a) which requires no voltage or current measurements, but only impedance ratios. This bridge is insensitive to the accuracy or the stability of the voltage supply. Interchanging the ratio impedances Z_a and Z_b allows for accurate 1 : 1 measurements.

The *Kelvin* double bridge (fig. 17a) is a 4-terminal version of the *Wheatstone* bridge in which the ratio is still provided by two 2-terminal impedances Z_a and Z_b. The *Kelvin* Network with the impedances K_a and K_b combines the two low potential leads of the 4-terminal impedances Z_X and Z_S into a single detection port. K_a and K_b are adjustable impedances. They are adjusted so that $K_\mathrm{a}/K_\mathrm{b} = Z_\mathrm{X}/Z_\mathrm{S}$. At bridge balance, an increase in Z_l (or addition of a small voltage in the connecting leads) does not change the reading of the central null detector. *Kelvin* networks are used at DC and low frequency in many high-accuracy 4-terminal and 4-terminal pair bridges.

The *Warshawski* bridge (fig. 17b) is similar to the *Kelvin* double bridge, but a pair of 4-terminal impedances Z_a and Z_b now determines the bridge ratio. This requires two low-potential *Kelvin* networks to combine the potential leads of Z_a and Z_b and Z_X and Z_S, respectively, into a single detection port for each impedance pair. The two low-potential *Kelvin* networks are adjusted in such a way that an increase in the impedance of the link



V_x
V_S
D
Z_x
Z_S
$I_X = I_S$
$\frac{Z_X}{Z_S} = \frac{V_X}{V_S}$
$\frac{Z_X}{Z_S} = \frac{Z_a}{Z_b}$
V
Z_a
Z_b
$I_X = I_S$
a)
b)

Fig. 16. – Comparison of two like impedances using a potentiometer (a)) or a *Wheatstone* bridge (b)).

between each two impedances does not change the reading of the main null detector. In a similar way, the two high-potential *Kelvin* networks make the bridge insensitive to the impedances in the potential leads of the main impedances. They are adjusted so that the detector does not respond to changes in the connecting leads of the high terminals of these impedances. Like the *Wheatstone* bridge, the bridge is insensitive to the accuracy or stability of the source. *Kelvin* networks are used at DC and low frequency in many high-accuracy 4-terminal bridges for low-value impedances.

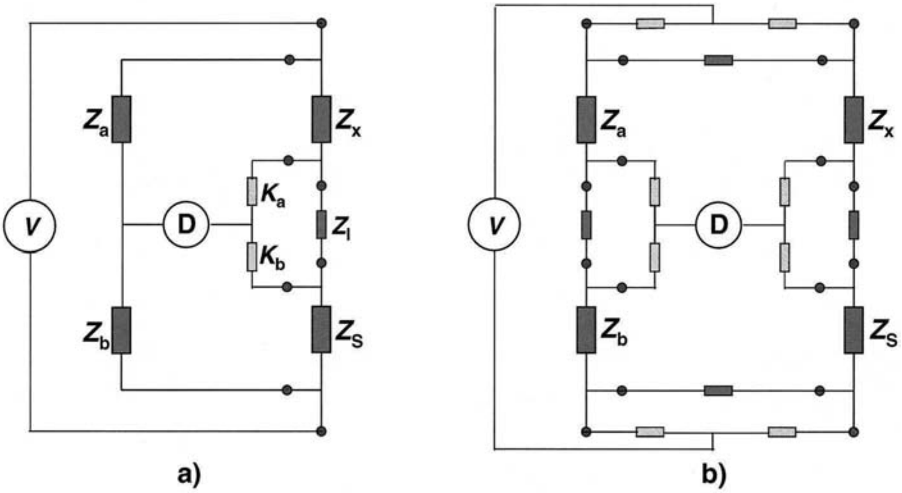


Fig. 17. – Comparison of two like impedances using a *Kelvin* double bridge (a)) or a *Warshawski* bridge (b)).

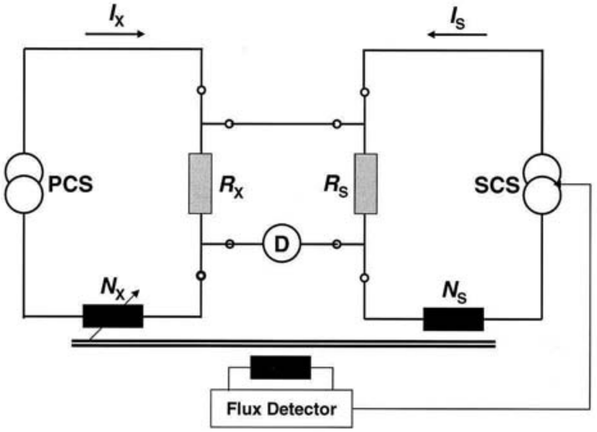


Fig. 18. – Principal circuit of a current comparator.

Current comparator bridges convert the accurate voltage and current measurement into the ratio of two currents or into a turns ratio (fig. 18). DC current comparators are used to compare low- and medium-valued resistances [15]. They consist of two current sources PCS and SCS which supply currents I_{X} and I_{S} to the two resistors. These currents also flow through the windings N_{X} and N_{S} of the comparator. One of the current sources, SCS, is controlled by a flux detector which detects equilibrium of the ampere turns in the two windings and adjusts SCS in such a way that the current ratio is equal to the inverse turns ratio. By variation of N_{X}, the difference in voltage drops across the resistors can be zeroed. Then the ampere turns of the two windings $I_{\mathrm{X}}N_{\mathrm{X}}$ and $I_{\mathrm{S}}N_{\mathrm{S}}$ are equal, and hence the ratio of the two resistances is equal to the turns ratio $R_{\mathrm{X}}/R_{\mathrm{S}} = N_{\mathrm{X}}/N_{\mathrm{S}}$. The DC current comparator bridge is a 4-terminal ratio bridge with good linearity and a resolution which depends on the number of winding turns N_{X}. Commercial bridges with range extender cover a range from $10\,\mu\Omega$ to $10\,\mathrm{k}\Omega$.

A *Hamon* resistor is a combination of n nominally equal resistors which can be connected in series and parallel, thus forming resistance ratios $R_{\mathrm{s}}/R_{\mathrm{p}} = n^2$ (fig. 19). In the series mode, the total resistance is given by the sum of the single elements which is equal to $R_{\mathrm{s}} = n \cdot R(1 + \sum \delta/n)$. δ is the small deviation of the resistor from its nominal value. In the parallel mode, special auxiliary networks ensure that nearly the same current flows through each resistance element. In this way, the total resistance is given by $R_{\mathrm{p}} = R/n \cdot 1/(1 + \sum \delta/n + \sum \delta^2/n)$. When the two equations are combined, a ratio of the series to parallel resistance is obtained which is close to n^2. Deviations from the actual resistances of the elements from their nominal value have a second-order effect.

$$\frac{R_{\mathrm{s}}}{R_{\mathrm{p}}} = n^2 + \frac{\sum \delta^2}{n}\,. \tag{17}$$

Hamon resistors allow resistance ratios with an uncertainty in the order of $1 \cdot 10^{-8}$ to be realised. Commercial devices cover a range from $1\,\Omega$ to $1\,\mathrm{G}\Omega$.

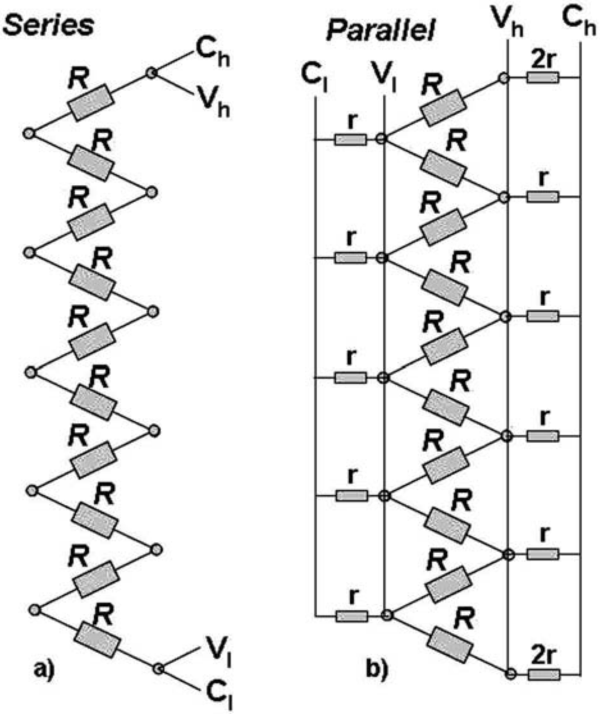


Fig. 19. – *Hamon* resistor in a serial (a)) and parallel (b)) connection.

For the transition from DC bridges to AC bridges (fig. 20), voltage sources are replaced by transformers or inductive voltage dividers (IVD's), resistors by impedances, null current detectors by injection or detection transformers, single wires by coaxial cables with current equalisers, and voltage null detectors by phase sensitive lock-in amplifiers [14]. Balancing of an AC bridge always requires two adjustments, the in-phase

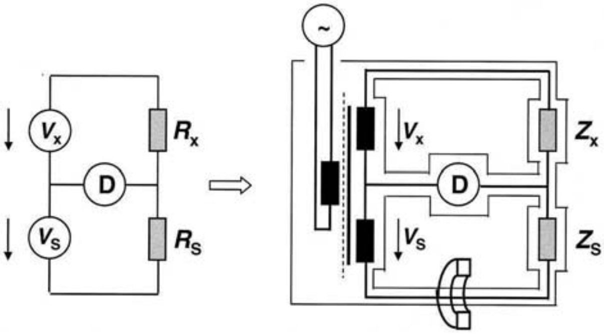


Fig. 20. – Transition from a DC to an AC bridge.

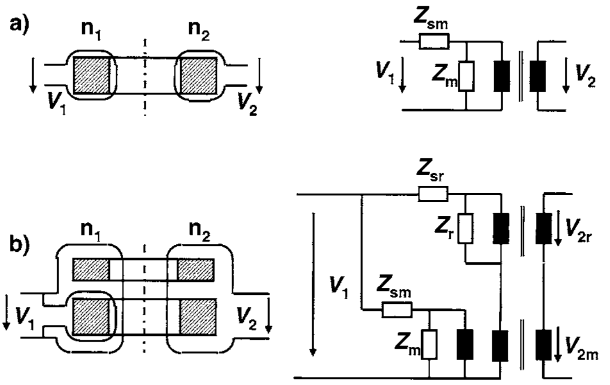


Fig. 21. – Principal drawing and equivalent circuit for a single-stage (a)) and a two-stage (b)) transformer.

and the quadrature balance, because AC quantities are characterised by their real and imaginary part or by their modulus and argument.

Single-stage transformers (fig. 21a) consist of two windings wound around a high-permeability core. If they are operated to generate a voltage ratio, *i.e.* with negligible output current, the voltage ratio is given by the turns ratio reduced by the ratio of the stray impedance of the primary winding and the main impedance of the transformer

$$\frac{V_2}{V_1} = \frac{n_2}{n_1}\left(1 - \frac{Z_{\mathrm{sm}}}{Z_{\mathrm{m}}}\right). \tag{18}$$

With a two-stage transformer (fig. 21b) it can be achieved that the voltage ratio comes closer to the turns ratio by several orders of magnitude. A two-stage transformer consists of two cores, a magnetising core and a ratio core. The magnetising winding is only wound around the magnetising core, while the two ratio windings surround both cores. The deficiency of the flux in the magnetising core is compensated by the much smaller flux in the ratio core which is produced by an electromagnetic force due to the discrepancy between the exact flux of an ideal transformer and the smaller flux provided by the magnetising core. The secondary voltage is the sum of the output voltages of the magnetising and ratio cores $V_{2\mathrm{m}}$ and $V_{2\mathrm{r}}$,

$$\frac{V_2}{V_1} = \frac{V_{2\mathrm{m}}}{V_1} + \frac{V_{2\mathrm{r}}}{V_1} = \frac{n_2}{n_1}\left(1 - \frac{Z_{\mathrm{sm}}}{Z_{\mathrm{m}}}\frac{Z_{\mathrm{sr}}}{Z_{\mathrm{r}}}\right). \tag{19}$$

$V_{2\mathrm{m}}$ is identical with the output voltage of the single-stage transformer while $V_{2\mathrm{r}}$ is this voltage reduced by a term $Z_{\mathrm{sm}}/Z_{\mathrm{m}}$. The stray impedances of a two-stage transformer only have a second-order effect on the voltage ratio which comes close to the turns ratio. With a well-designed transformer, Z_{s}/Z can easily be reduced to below 10^{-3}. Bridge transformers and IVD's can either be one-stage or two-stage devices, depending on the field of application.

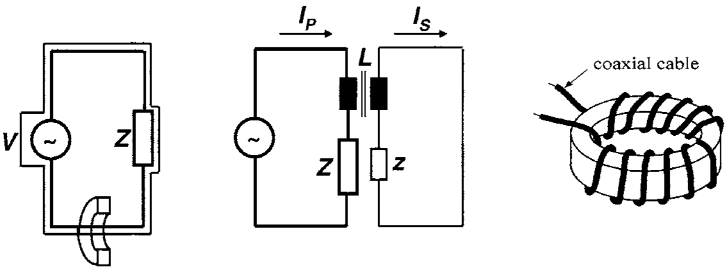


Fig. 22. – Principal drawing, equivalent circuit and realisation for a current equaliser.

Current equalisers are used to equalise currents in the inner and outer conductor of coaxial cables which therewith become magnetically astatic (fig. 22). A current equaliser can easily be established by threading a coaxial cable a few times through a high-permeability toroidal core. To equalise the currents in a network, one current equaliser is required for each mesh of the network. By equalising the currents in the inner and outer conductors, the network becomes insensitive to external electromagnetic fields and, at the same time, does not disturb its environment, because no magnetic field is generated outside the coaxial cables. The function of a current equaliser is that of a 1 : 1 current transformer. The secondary current, *i.e.* the current in the outer of the coaxial cable, is nearly equal and opposite to the primary current, *i.e.* the current in the inner of the coaxial cable. Due to the equaliser, the small impedance of the outer conductor z increases the impedance in the circuit of the inner conductor Z.

In coaxial bridges it is very useful to be able to either introduce a generator of a small voltage ΔV at a point along the inner of a coaxial cable or to detect the vanishing of a current $i = 0$ at a point along the cable. This is, for example, very important for fulfilling the defining conditions of a 4-terminal pair impedance. Both can be achieved by using an injection/detection transformer. Such a transformer is a $n : 1$ transformer whose primary is the n turns on a toroidal magnetic core and whose secondary is the single turn of the inner of a coaxial cable. The screen of the coaxial cable which forms the secondary winding must be gapped to prevent shortening of the primary winding of the transformer.

Figure 23 shows a classical four-arm bridge with the main admittances $Y_1 \ldots Y_4$. The stray admittances of the main components are concentrated in the two source admittances Y_{S} and Y'_{S} and the two detector admittances Y_{D} and Y'_{D}. After balancing of the bridge, points "a" and "b" will be at the same potential, but not necessarily at earth potential. This makes the bridge very sensitive to changes in the detector admittances. This problem can be solved by adjusting the source admittances so that points "a" and "b" are not only at the same potential but also at earth potential. This condition is fulfilled when the ratio of the source admittances equals the ratio of the main admittances. Note that Y_3, Y_4, Y_{S} and Y'_{S} form a bridge which is balanced, if $Y_{\mathrm{S}}/Y'_{\mathrm{S}} = Y_3/Y_4 = Y_1/Y_2$. The main balance is then insensitive to changes of the detector admittances, because

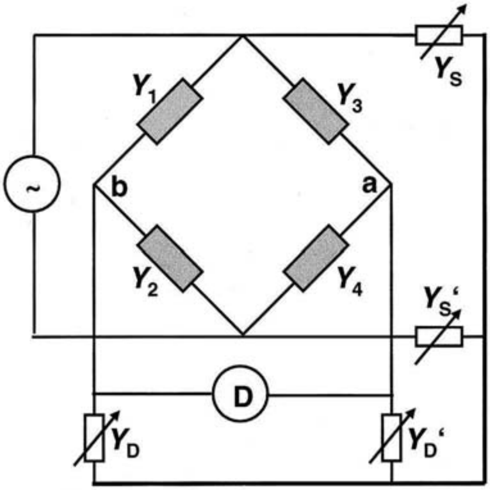


Fig. 23. – General four-arm bridge with source and detector balance.

they are now shortened. In a similar way, an adjustment of the detector admittances can be performed by applying the reciprocity principle, *i.e.* the unchanged behaviour of a bridge when source and detector are interchanged. In this case, Y_2, Y_4, Y_{D} and Y'_{D} form a bridge which is balanced if the condition $Y_{\mathrm{D}}/Y'_{\mathrm{D}} = Y_1/Y_3 = Y_2/Y_4$ is fulfilled. This balance makes the bridge insensitive to changes of the source admittances. Both auxiliary balances gain, so that adjustment of both with an accuracy of $1 \cdot 10^{-3}$ is sufficient to ensure that shunt impedances do not affect the bridge balance by more than $1 \cdot 10^{-6}$.

Figure 24 shows how these principles can be applied to a 1 : 1 ratio bridge to compare 4-terminal pair impedances. The bridge is formed by the main transformer T_2 which is a two-stage IVD and by impedances Z_{H} and Z_{S}. Bridge balance is controlled by the lock-in amplifier Det together with a preamplifier. The 4-terminal pair conditions on the high voltage side are controlled by nulling of the two detection transformers at the $+U$ and $-U$ terminal of T_2 and adjusting the two current sources D_3 and D_5 accordingly. The main balance of the bridge is performed by means of D_1 together with Inj_1 and D_2 together with C_{Q}. The in-phase balance is achieved by injecting a small voltage in addition to one of the voltages of the main bridge divider and therewith altering the ratio of $+U$ and $-U$. The quadrature balance is made by injecting a small current into the detector point of the bridge generated by the voltage source D_2 and the capacitor C_{Q}. A *Kelvin* network is used to fulfil the 4-terminal pair conditions on the low-voltage side of the impedances and to combine the two potential leads into one single detection point. The source balance (*Wagner* balance) is made by injecting a small voltage into the zero tap of the bridge supply by means of D_4 and Inj_4 so that the zero tap of the main bridge divider is at zero potential with no current flowing through this tap as detected by Det_4. The bridge is fully balanced by repeating the main and auxiliary adjustments several times until the detector reads zero for all balances at the same time. After this has been

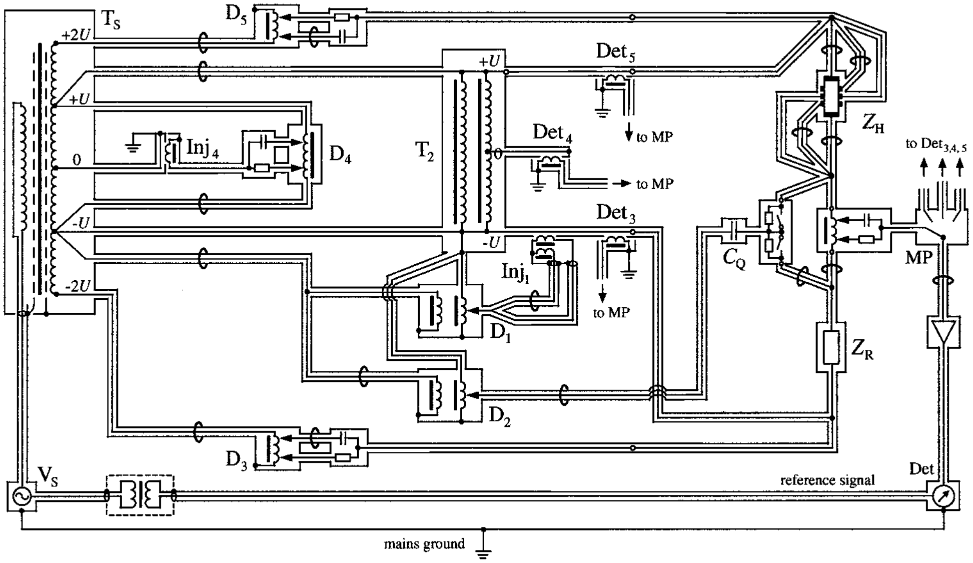


Fig. 24. – Realisation of a 4-terminal pair bridge for a comparison of like impedances.

done, the balance condition is given by

$$Z_{\mathrm{H}} = Z_{\mathrm{R}} / \left(1 + \alpha + j\beta\omega C_{\mathrm{Q}} Z_{\mathrm{R}}\right), \tag{20}$$

with α and β being the settings of dividers D_1 and D_2 for the main bridge balance. For a 1 : 1 ratio bridge, the error of the main bridge divider T_2 can be determined by performing two bridge adjustments, with and without reversal of the windings of T_2.

9. – AC/DC transfer

AC/DC transfer techniques form the interface between DC and AC measuring techniques. They are the prerequisite for accurate voltage, current and power measurements in a wide frequency range. Equivalence of AC and DC quantities is achieved when they produce the same electrical power which leads to a temperature increase in a heating element sensed by means of a thermal converter [16]. Today's electronic instrumentation demands most accurate calibration of AC voltages, currents and power. In a frequency range from 100 Hz to 100 MHz, thermal methods are used, whereas for frequencies from DC to 100 Hz sampling methods are predominating. The voltage range for transfer measurements extends from less than 2 mV to 1000 V with a basic uncertainty of 1 to 10 μV/V at 3 V in a frequency range from 10 Hz to 1 MHz. Current transfer measurements can be performed in a range from 3 mA to 100 A with an uncertainty of 3 μA/A at 10 mA in a frequency range from 10 Hz to 100 kHz. In the overlapping frequency range from 10 Hz to 100 Hz, thermal and sampling methods agree within one part in 10^6.

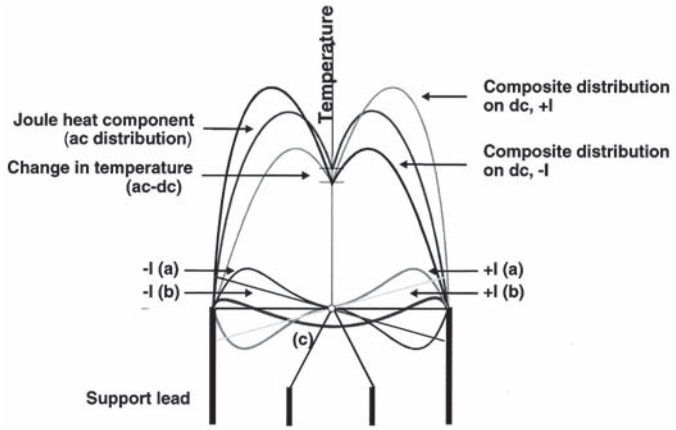

Temperature
Composite distribution on dc, +I
Joule heat component (ac distribution)
Composite distribution on dc, -I
Change in temperature (ac-dc)
-I (a)
-I (b)
+I (a)
+I (b)
(c)
Support lead

Fig. 25. – Thermoelectric effects of a single-junction thermal converter: a) first-order Thomson heat component; b) first-order Peltier heat component; c) second-order Thomson heat component (independent of current direction).

Thermal converters work in accordance with the equivalent heating power principle, *i.e.* the equivalence of the *Joule* heat of a DC and AC current in a heater. The *Joule* heat leads to an increase in temperature at the hot junction of a thermocouple which can be measured by the *Seebeck* effect. The *Seebeck* effect describes a phenomenon where an electrical potential E is generated when a material with a differential thermoelectric force Q is exposed to a temperature gradient $\mathrm{grad}\, T$, $E = Q \cdot \mathrm{grad}\, T$. In practice, the electric current flows through a thin resistive wire generating a *Joule* heat which causes an increase in temperature (T_h). This temperature is very sensitively measured by means of a thermocouple made of the materials A and B: $V = \alpha(T_\mathrm{h} - T_\mathrm{c})$, T_c is the temperature of the cold junction. The thermoelectric voltage is measured with a nanovoltmeter. It is very important that the connecting leads to the voltmeter are of the same material C or of a material having the same *Seebeck* coefficient and that they are at the same temperature T_c as the low temperature junction of the thermocouple. Otherwise, an unwanted additional thermoelectric voltage will be generated which leads to an erroneous measurement. A single-junction thermal converter consists of a thin heater wire and a centrally arranged thermocouple in an evacuated glass bulb. The thermocouple is insulated from the heater by means of a bead. Due to their small dimensions, single-junction thermal converters are distinguished by a good frequency characteristic. Their sensitivity is, however, much smaller than that of multijunction thermal converters, and their AC/DC transfer error cannot be calculated with the desired low uncertainty.

Thermoelectric effects (fig. 25) appear when different materials are connected or when a temperature gradient exists within a material. These effects cause a change in the temperature distribution along the heater for DC+, DC- and AC and therewith cause an AC/DC transfer error. First-order *Thomson* effect: In addition to the Joule heat, a Thomson heat exists which is proportional to the *Thomson* coefficient σ, the current I

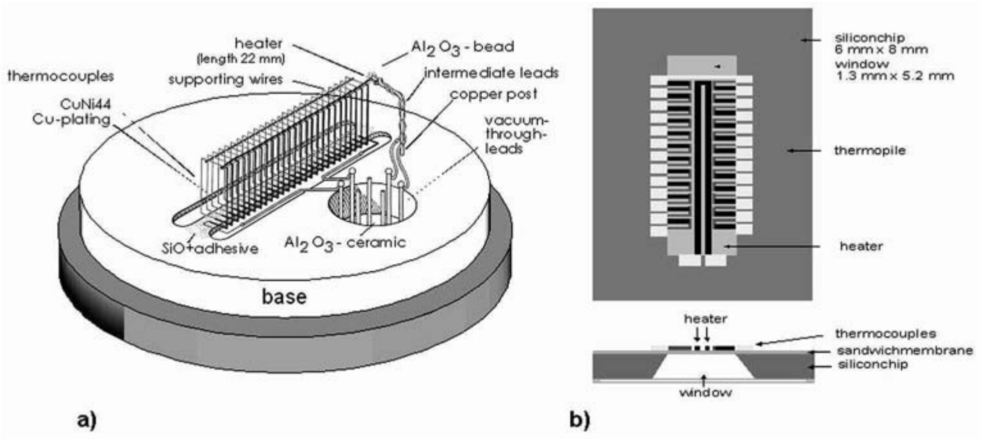


Fig. 26. – Multijunction thermal converters in 3D (a)) and planar (b)) design.

and the temperature difference ΔT, $Q_{\mathrm{Th}} = \pm\sigma \cdot I \cdot \Delta T$. This effect only exists for DC, because due to the thermal inertia the temperature cannot follow for AC. *Peltier* effect: In a steady-state condition, the *Peltier* effect causes a linear temperature gradient along the heater caused by the connection of the heater to the support leads. Note that the heater and the support leads are normally made of different materials! Again this effect only exists for DC. Second-order *Thomson* effect: As the *Thomson* coefficient σ does not only depend on the material but also on the absolute temperature, an additional *Thomson* heat is generated which depends on the temperature. This heat and, therewith, changes in temperature, are independent of the current direction. The three characteristics in the upper part of the graph summarise all these effects and show the temperature distribution along the wire for AC, DC+ and DC-. It results in a temperature change between AC and DC and therewith leads to an AC/DC transfer error. These effects limit the accuracy of a single thermal converter; they can be drastically reduced or even eliminated when multijunction thermal converters are used for which the temperature distribution along the heater is nearly constant.

Work at PTB on AC/DC transfer started with the development of 3D multijunction thermal converters (fig. 26a). The heater is formed by a twisted bifilar wire. It is supported by 60 to 120 thermocouples which are connected in series and allow the temperature difference between the heater and the copper post to be measured with a resolution of $1 \cdot 10^{-7}$. The AC/DC transfer differences are due to reactive components, skin effect and dielectric losses. They were calculated with an uncertainty of $3 \cdot 10^{-7}$ in a frequency range from 10 Hz to 100 kHz. The construction of 3D multijunction thermal converters is most sophisticated and time-consuming as complicated manual work has to be done under the microscope with wires of only 10 to 20 μm in diameter. Due to the construction of MJTC's, first- and second-order thermoeffects do only have negligible influence on the AC/DC transfer error, because the larger number of thermocouples equalises the temperature along the heater and therewith eliminates these effects. Due to the larger output voltage, the temperature difference between the hot and cold junctions can be kept small without loosing sensitivity. The bifilar design of the heater (the current flows in both directions!) together with a good thermal connection between the support leads keeps the *Peltier* effect small.

In planar multijunction thermal converters (fig. 26b), manual manufacture can be replaced by photolithography, thermal evaporation and sputtering techniques [17]. A silicon chip acts as a heat sink. In the middle, a window is opened by anisotropic etching so that only a thin membrane is left. To avoid mechanical stress which would destroy the membrane, it is designed as a sandwich of Si_3N_4-SiO_2-Si_3N_4. It exhibits poor thermal conductivity. The heater and the hot junctions are placed on the membrane, while the cold junctions are deposited on the silicon chip. The converter chip is pasted on a ceramic carrier, and the electrical connections are made by bonding. A disadvantage is the small time constant (30 ms) of the heater-thermocouple-membrane system which is characterised by an increase of the AC/DC transfer error at frequencies below about 100 Hz. To increase this time constant, a silicon obelisk can be arranged below the heater (not shown). It is also formed by anisotropic etching and increases the time constant τ (1.3 s instead of 30 ms). In this way, the transfer error can be decreased for low frequencies. At the same time, thermoelectric effects become negligible, because the temperature along the heater is equalised.

The AC/DC transfer is based on the equality of heating power at AC and DC, monitored by equal output voltages E_{DC} and E_{AC} of the converter. For voltage transfer, the power at AC and DC is given by

$$V_{\mathrm{AC}}^2 \frac{\operatorname{Re}\{Z_{\mathrm{H}}\}}{|Z_{\mathrm{H}}|^2} = V_{\mathrm{DC}}^2 \frac{1}{R_{\mathrm{H}}}\,, \tag{21}$$

resulting in an AC/DC transfer error of

$$\delta_V = \left.\frac{V_{\mathrm{AC}} - V_{\mathrm{DC}}}{V_{\mathrm{DC}}}\right|_{E_{\mathrm{AC}}=E_{\mathrm{DC}}} = \frac{|Z_{\mathrm{H}}|}{\sqrt{R_{\mathrm{H}} \cdot \operatorname{Re}\{Z_{\mathrm{H}}\}}} - 1. \tag{22}$$

For current transfer, the power at AC and DC is given by

$$I_{\mathrm{AC}}^2 \operatorname{Re}\{Z_{\mathrm{H}}\} = I_{\mathrm{DC}}^2 R_{\mathrm{H}}, \tag{23}$$

resulting in an AC/DC transfer error of

$$\delta_I = \left.\frac{I_{\mathrm{AC}} - I_{\mathrm{DC}}}{I_{\mathrm{DC}}}\right|_{E_{\mathrm{AC}}=E_{\mathrm{DC}}} = \sqrt{\frac{R_{\mathrm{H}}}{\operatorname{Re}\{Z_{\mathrm{H}}\}}} - 1. \tag{24}$$

Reasons for transfer errors are:

1) The modulus and the in-phase component of the heater resistance at AC ($|Z_{\mathrm{H}}|$ and $\mathrm{Re}\{Z_{\mathrm{H}}\}$) differ from the value at DC (R_{H}).

2) The temperature along the heater differs for AC and DC due to secondary thermoelectric effects.

3) The characteristic of the thermal converter does not exactly follow a quadratic characteristic. This effect increases with decreasing frequency, because the temperature of the heater can then follow the instantaneous value of the heating power.

Transfer differences of a thermal converter strongly increase towards low frequencies (fig. 27a). The reason for this difference is the time constant of the converter. As soon as the time for half a period of the input signal comes into the order of the time constant of the converter, the temperature of the heater can follow the Joule heat, and the output voltage starts to oscillate with twice the frequency of the input signal. Due to non-linearities between input and output voltage, the transfer error increases with decreasing frequency, because the maximum temperature of the heater increases as well. An obelisk below the heater shifts these transfer errors to even lower frequencies, because it increases the thermal time constant (1.3 s instead of 30 ms). The increase in the transfer error at low frequencies with higher heater resistances can be explained by a non-uniform distribution of the temperature across the heater cross-section, because the thickness of the heater decreases with higher resistances and the temperature cannot reach equilibrium. The capacitance between different parts of the heater, the self-inductance of the heater, skin effect, dielectric losses, and the capacitance and mutual inductance between the heater and the thermocouples influence the high frequency behaviour of a PMJTC (fig. 27b). There are two opposing effects which influence the transfer error at high frequencies: the skin effect and the conductance due to capacitance losses between the heater and the hot junctions. The AC/DC transfer difference caused by the skin effect is inversely proportional to the heater resistance, *i.e.* it increases with decreasing heater resistance. The AC/DC transfer difference caused by the conductance decreases with increasing heater resistance. The transfer errors measured at higher frequencies clearly show these dependences. PMJTC's on a quartz membrane improve the high-frequency behaviour and, at higher frequencies, are characterised by transfer errors which are one order of magnitude smaller than those for thermal converters on a silicon sandwich membrane.

The 2-channel-method (fig. 28a) is used to compare an unknown thermal converter with a reference converter with known AC/DC transfer error. The heaters of the two thermal converters are connected in parallel. It is important to connect the housing of the converter and its input and output circuits to ground to protect the insulation between the heater and the thermocouples against damage. If we suppose that the AC voltage is the same for both thermal converters and that the two DC voltages are nearly equal, the difference $\delta_{\mathrm{U}} - \delta_{\mathrm{S}}$ is given by $(U_{0\mathrm{U}} - U_{0\mathrm{S}})/U_0$. As this method is very sensitive to instabilities of the input voltages, the differential measurement method (fig. 28b) is used

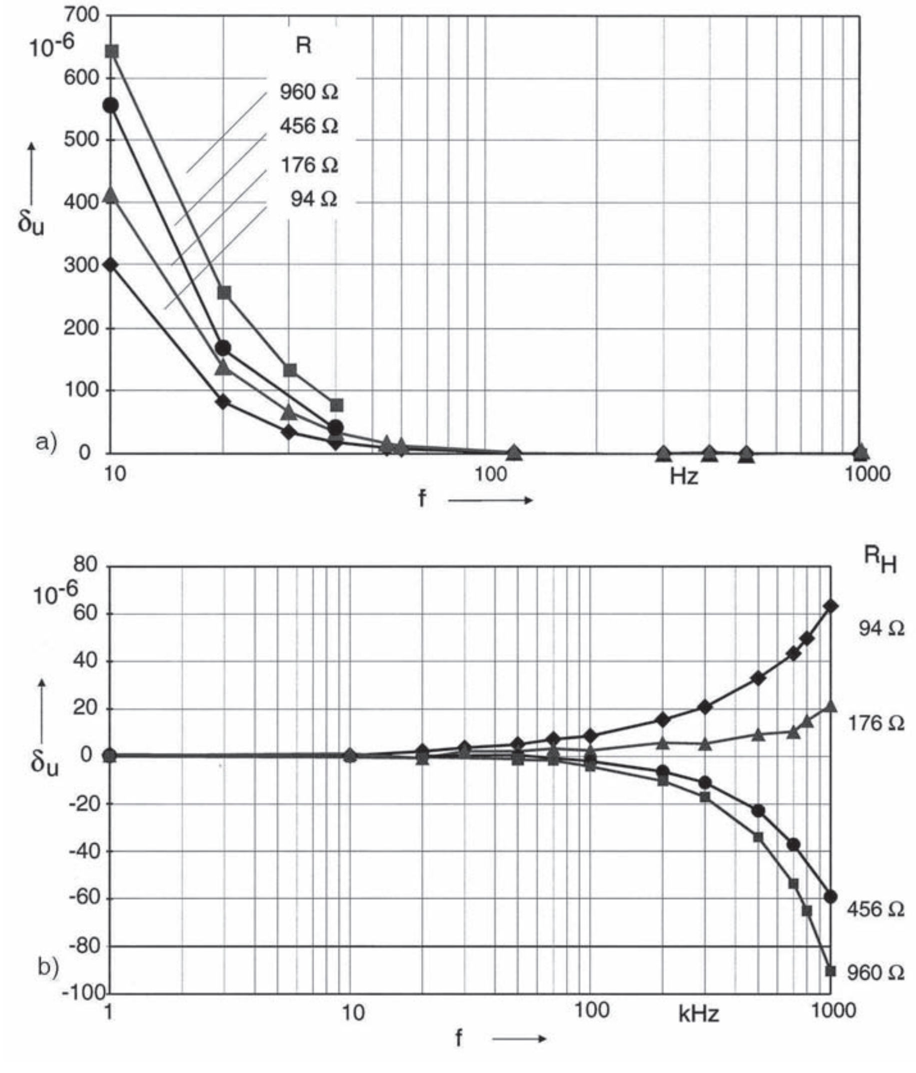


Fig. 27. – a) AC/DC transfer error at low frequencies; b) AC/DC transfer error at high frequencies.

for most precision measurements of a DUT against a standard. It is insensitive to the drift of the input voltages. Necessary prerequisites for this method are with $V_\mathrm{a} = k \cdot V_\mathrm{e}^n$:

1) The exponents for the relation between input and output voltage agree within 10^{-3}.

2) The proportional factor k_U and k_S can be adjusted to the same value by means of a load resistor in one of the output circuits.

3) The time constants of the two converters are nearly equal.

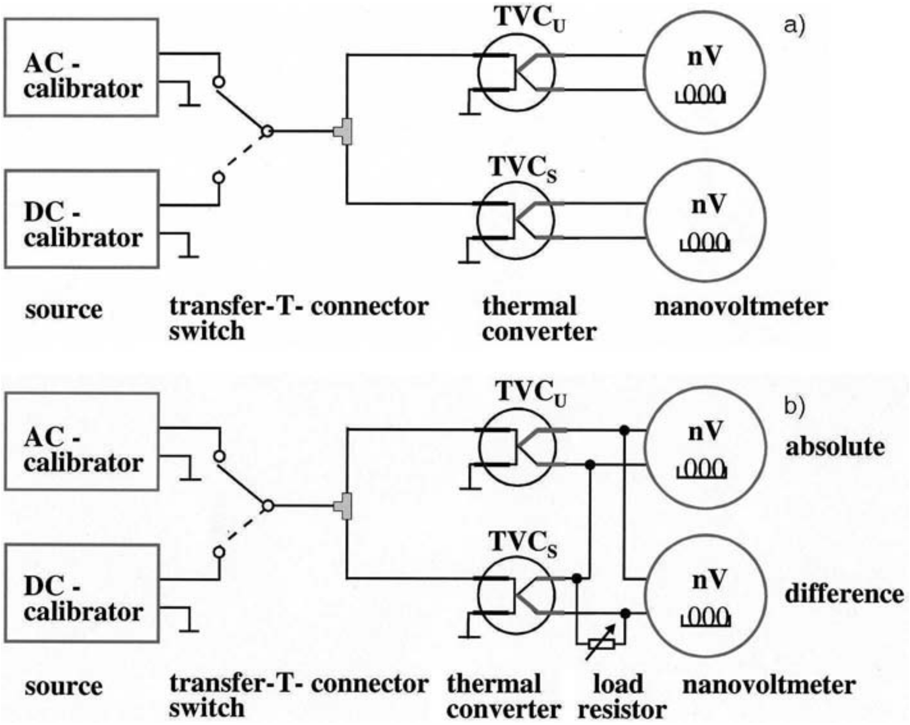


Fig. 28. – a) 2-channel-method for comparing thermal converters; b) differential measurement method for comparing thermal converters.

On these conditions, the difference $\delta_U - \delta_S$ is given by the differences of the output voltages at AC and DC related to n_S times the output voltage at DC $(\Delta U_{af} - \Delta U_{a0})/n_s U_{a0}$. The measurements are performed under computer control. The computer equally adjusts the time intervals between DC and AC measurements. With a PMJTC directly connected to a voltage source, a range from 100 mV to 3 V can be covered with PMJTS's having different heater resistances. Above 3 V, a series connection of a range resistor and a PMJTC must be used. For a voltage step-up it is always necessary to have a certain overlap between the different PMJTC's and range resistors. In this way, an AC voltage scale ranging from 100 mV to 1000 V can be established. For voltages below 100 mV, voltage dividers or micro-potentiometers are used which allow the voltage scale to be scaled down to 2 mV. Even smaller voltages in the μV range are presently under development.

For a comparison of two thermal current converters (fig. 29), both converters must be connected in series which causes a common mode voltage for the upper converter. By means of a sophisticated guard technique it can be guaranteed that exactly the same current is fed to both converters. This technique consists in a symmetrical arrangement of the DUT (TC$_X$) and the standard (TC$_N$). The upper converter is connected to the

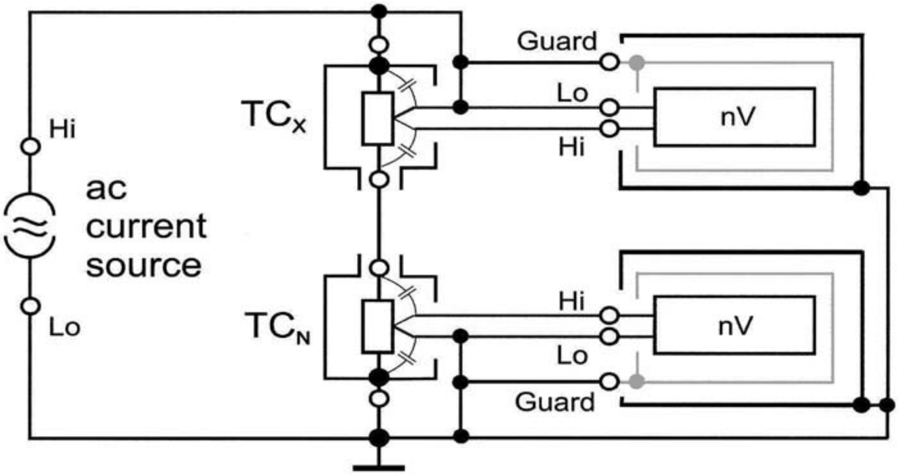


Fig. 29. – Comparison of two thermal current converters.

circuit topside down. Parasitic capacitances between the heater and the thermal elements are therefore constant and independent of the converter position, and both converters can be exchanged without changing the transfer difference. With a PMJTC directly connected to a current source, a range from 3 mA to 10 mA can be covered with PMJTS's of different heater resistances. Above 10 mA, a combination of a shunt together with a PMJTC must be used. For a current step-up, it is always necessary to have a certain overlap between the different PMJTC's and shunts. In this way, an AC current scale ranging from 3 mA to 20 A can be established. As PMJTC's on a quartz substrate have a higher current capability, the current step-up can be made with a smaller number of steps and, therewith, with a smaller uncertainty. Even higher currents of up to 100 A are presently under development. The frequency range extends from 10 Hz to 100 kHz and will be extended to 1 MHz.

10. – Power and energy measurements

Power and energy measurements will gain in importance due to the limited resources and renewable energy sources which require special measuring techniques. Ensuring a consumer-oriented reliable and sustainable supply of electric power in view of the increasingly scarce and expensive resources will be one of the future challenges of society. Metrology must contribute to the solution of this problem by providing enhanced measuring capabilities for the quality and efficiency of electrical power, and by monitoring and protecting power apparatus. In the Federal Republic of Germany, electric energy of approximately 50 000 000 000 € per year is consumed. This is equivalent to an energy consumption of 480 000 000 000 kWh per year. More than 95% of all electricity meters used in Germany for invoicing are electromechanical induction meters. 42 million of them are in use to measure the private consumption in households. About 650 thousand transformer-connected electronic meters measure the industrial consumption. Electrical power measurements —and other areas of metrology as well— are hierarchically structured. To serve our customers in industry and in the private sector, an unbroken chain of measurements from the national standard to the meters used by our customers must be established. To calibrate electronic and electromechanical meters with the desired accuracy, the standards kept at the test centres should be about one order of magnitude better than the meters under test and this is the case with the working standards at the NMI's. This means that power measurements are required with an uncertainty of a few parts in 10^6. These primary standards have to be compared with the primary standards of other NMI's to guarantee world-wide uniformity of energy measurements. Power measurements are traceable to the units of voltage, resistance and time which are realised by macroscopic quantum effects (*Josephson* and quantum *Hall* effects) and atomic clocks.

Let us consider an AC generator which generates a sinussoidal output voltage $u(t)$ and supplies a current $i(t)$ to a complex load. The real power dissipated in this load is the time integral of the instantaneous power which the product of the instantaneous voltage and current values:

$$P = \frac{1}{T}\int_0^T u(t)\cdot i(t)\mathrm{d}t = \frac{1}{T}\int_0^T p(t)\mathrm{d}t = U\cdot I\cdot\cos\varphi, \tag{25}$$

with U and I being the effective voltage and current values and φ the phase shift between them. The apparent power $S = U\cdot I$ and the reactive power $Q = U\cdot I\cdot\sin\varphi$ can be calculated in a similar way. The electrical energy is the time integral of the instantaneous power:

$$W = \int_0^T p(t)\mathrm{d}t = P\cdot t. \tag{26}$$

It equals $P\cdot t$ for a time interval t, provided the power is kept constant over the integration time.

The effective value of a voltage can be measured by sampling the sine-shaped signal $u(t)$ with period T using an integrating sampling voltmeter with t_i being the integration time of the voltmeter and t_s the sampling interval. n is the number of samples per period. Only if the following conditions are fulfilled, can the rms value of the voltage be calculated with the desired small uncertainties from the samples U_i:

1. the sampled signal must be sine-shaped and periodic (period T),

2. the distortion of the signal must be negligibly small,

3. following the Nyquist criterion, the number n of samples per period must be > 2, and

4. the number of samples n must be an integer number, *i.e.* $n\cdot t_\mathrm{s} = T$.

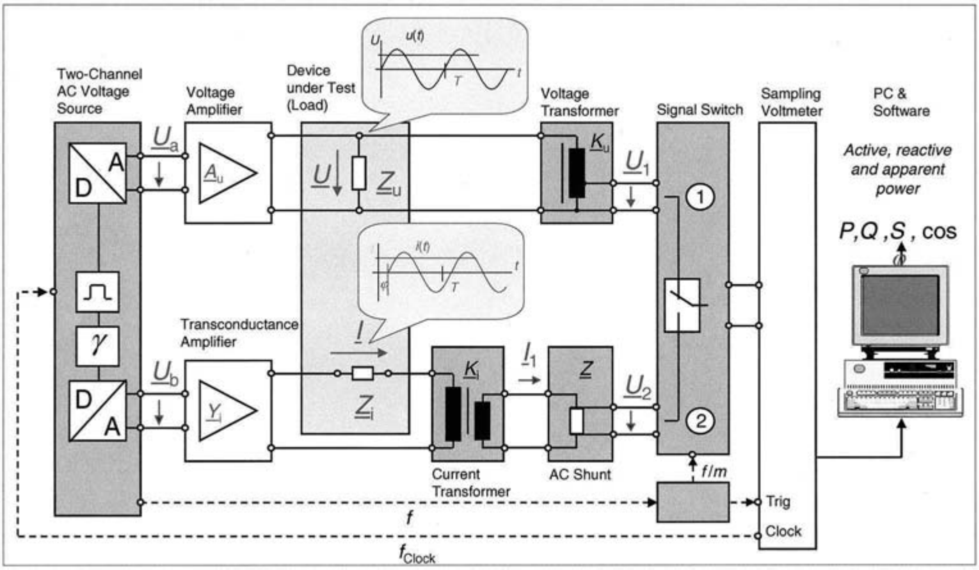


Fig. 30. – Calibration system for active, reactive and apparent electric power.

In this case, the effective value can be calculated by

$$U_{\mathrm{rms}} = \operatorname{Sinc}\alpha \cdot \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} u_i^2}\,, \tag{27}$$

with

$$\operatorname{Sinc}\alpha = \frac{\sin\alpha}{\alpha} = \frac{\sin(\pi t_i/t_{\mathrm{s}})}{\pi t_i/t_{\mathrm{s}}}\,. \tag{28}$$

The PTB calibration system for active, reactive and apparent electric power (fig. 30) is based on digitally synthesised AC voltages, the synchronous sampling of two AC signals with one sampling voltmeter and on the determination of their complex AC voltage ratio by means of discrete *Fourier* transform [18]. The two-channel AC voltage source generates two sinusoidal voltages $\underline{U}_{\mathrm{a}}$ and $\underline{U}_{\mathrm{b}}$ which, after amplification and conversion into a current, are fed into the DUT ($\underline{U}$ and $\underline{I}$). Their frequency f is derived from the clock signal provided by the sampling voltmeter. By means of a voltage transformer and a current transformer with an AC shunt, voltage $\underline{U}$ and current $\underline{I}$ are converted and scaled-down to voltages $\underline{U}_1$ and $\underline{U}_2$ which are sampled by the sampling voltmeter. A PC assumes the sampled values and performs the *Fourier* transform. The voltages $\underline{U}_1$ and $\underline{U}_2$ are sampled alternately depending on the setting of the signal switch. The measuring cycle starts with sampling $\underline{U}_1$ over 8 periods, only 5 periods being used for the *Fourier* transform to prevent distortions due to the switching of the signals. $\underline{U}_2$ is sampled in an identical way. As the two voltages are sampled subsequently, the short-term stability of the double voltage source is of great importance. To calculate the real, apparent and reactive power from the voltage $\underline{U}$ and current $\underline{I}$ applied to the DUT, $\underline{U}$ must derived from $\underline{U}_1$ and the ratio of the voltage divider, $\underline{U} = \underline{U}_1 \cdot \underline{K}_\mathrm{U}$. $\underline{I}$ can be obtained from $\underline{U}_2$, the shunt impedance $\underline{Z}$ and the ratio of the current transformer, $\underline{I} = \underline{I}_1 \cdot \underline{K}_\mathrm{I} = \underline{U}_2 / \underline{Z} \cdot \underline{K}_\mathrm{I}$. The phase shift between $\underline{U}$ and $\underline{I}$ is the arctangent of the quotient of the imaginary and real part of the complex ratio $\underline{I}/\underline{U}$, given by the quotient $\underline{U}_2/\underline{U}_1$, the quotient of the transformer ratios $\underline{K}_\mathrm{i}/\underline{K}_\mathrm{u}$ and the inverse shunt impedance $1/\underline{Z}$. The ratio $\underline{U}_2/\underline{U}_1$ is determined by discrete *Fourier* transform. Within the basic range $U = 120\,\mathrm{V}$, $I = 5\,\mathrm{A}$, $\cos\varphi = 0\ldots 1$ (active and reactive power), active, apparent and reactive power can be measured with an uncertainty below 5 parts in 10^6 and a probability of 95% ($k = 2$). This uncertainty doubles if the voltage and current ranges are extended from 60 V to 240 V (test points 40% to 120%) and from 0.1 A to 10 A (test points 40% to 120%) which corresponds to a range of apparent power from 1 VA to 3600 VA.

11. – International comparisons

Globalisation also asks for a harmonisation of metrology. International comparisons guarantee the uniformity of measurements and therewith make a valuable contribution to this process. In the globalisation era, metrology must follow the economic development, *i.e.* for a NMI it is no longer sufficient to keep the national standards with the desired accuracy, but it must also compare them with the standards of other NMI's on a regional and international level. To do this, CIPM introduced a Multilateral Recognition Arrangement (MRA) for the mutual recognition of national measurement standards and of the calibration and measurement certificates issued by National Metrology Institutes (NMI's) which provides the necessary framework for these activities [19]. International comparisons have a long tradition in the *Consultative Committees* (CC's) of the CIPM and the *Regional Metrology Organisations* (RMO's). They gained in importance with the introduction of the MRA of the CIPM. Key comparisons form the technical basis of the MRA. Their role is to test the principal techniques in each field, to provide data for calculating the degree of equivalence of national standards and therewith give mutual confidence in the measurement capabilities of the participating NMI's. Key comparisons are carried out by the BIPM or the CC's of the CIPM (so-called CIPM key comparisons) or the RMO's (so-called RMO key comparisons). The organisations active in organising comparisons are APMP, COOMET, EUROMET and SIM.

A close-meshed network of comparisons covers the field of AC/DC transfer for voltage and current in different ranges and at different frequencies. A comparison always starts with a CCEM comparison which is supplemented by RMO comparisons of the same kind. As a few laboratories participate in both, the CCEM and the corresponding RMO comparison, there is always a close link between these comparisons which allows the results to be compared directly. CCEM-K6.a is the basic comparison for AC/DC transfer. It was performed at 3 V in a frequency range from 1 kHz to 1 MHz with utmost accuracy. In total, 45 laboratories participated in the CCEM and the corresponding RMO com-

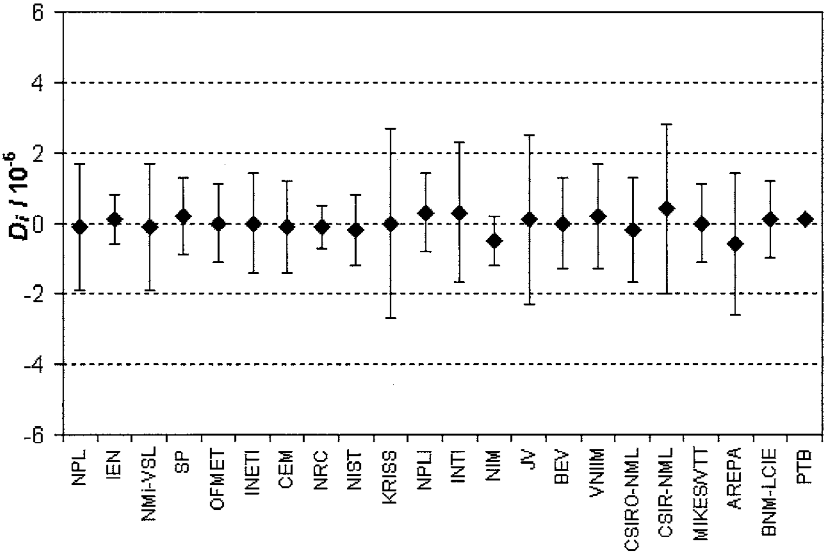


Fig. 31. – Results of the CCEM key comparison CCEM-K6.a for AC/DC transfer.

parisons. Figure 31 shows the results of the CCEM comparison at 1 kHz, which are very satisfying, and furnish the proof that the participants are able to make AC/DC transfer measurements with utmost accuracy. The results of the 22 participants agree with each other within less than 1 part in 10^6 which is well within their stated uncertainty demands.

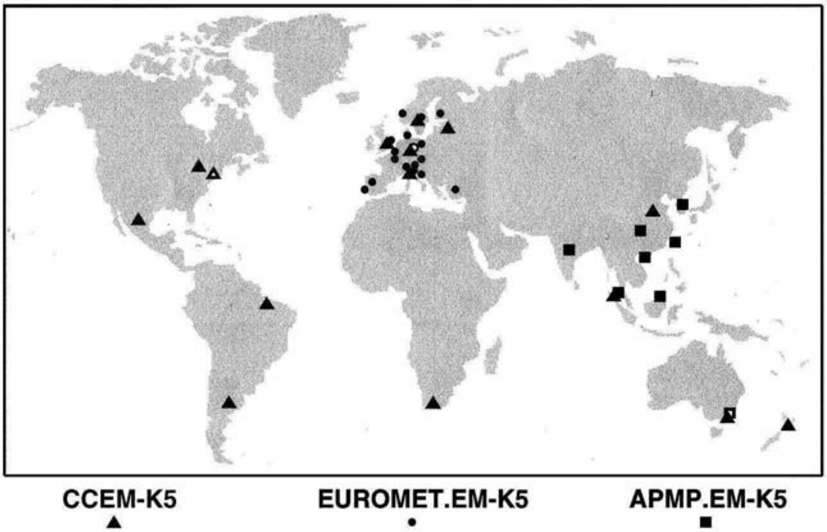


Fig. 32. – World-wide network of comparisons for electric power.

CCEM-K5, a key comparison on AC power, is another example. Together with the RMO key comparisons, four comparisons were organised in different regions. The comparison was made at 120 V, 5 A and at a frequency between 50 Hz and 60 Hz for power factors 1 to zero (leading and lagging). More than 50 laboratories participated in these comparisons. Again all participants agree well within their stated uncertainties. Together with the RMO key comparisons, the CCEM key comparisons form a world-wide network of comparisons as can be shown, for example, for the K5 comparisons on AC power (fig. 32). The white circles denote the pilot laboratories (NIST, NMIA and PTB), two circles at the same place mean that this laboratory acted as linking laboratory for two comparisons (Australia and Singapore for APMP.K5 and UK, Sweden, Italy and Germany for EUROMET.EM-K5). This map impressively shows the international role which metrology plays today.

REFERENCES

[1] *The International System of Units*, 8th edition (International Bureau of Weights and Measures) 2006.

[2] Vigoureux P., *Metrologia*, **1** (1965) 3.

[3] Kibble B. P., *At. Masses Fundam. Constants*, **5** (1975) 545.

[4] Steiner R., Williams E., Newell D. and Liu R., *Metrologia*, **42** (2005) 431.

[5] Sienknecht V. and Funck T., *IEEE Trans. Instrum. Meas.*, **34** (1985) 195.

[6] Clothier W. K., *Metrologia*, **1** (1965) 36.

[7] Jeffery A. M., Elmquist R., Lee L. H., Shields J. Q. and Dziuba R. F., *IEEE Trans. Instrum. Meas.*, **46** (1997) 264.

[8] Josephson B. D., *Phys. Lett.*, **1** (1962) 251.

[9] Pöpel R., *Metrologia*, **29** (1992) 153.

[10] Taylor B. N. and Witt T. J., *Metrologia*, **26** (1989) 47.

[11] v. Klitzing K., Dorda G. and Pepper M., *Phys. Rev. Lett.*, **45** (1980) 494.

[12] Jeckelmann B. and Jeanneret B., *Rep. Prog. Phys.*, **64** (2001) 1603.

[13] Keller M. W., in *Recent Advances in Metrology and Fundamental Constants, Proceedings of the International School of Physics "Enrico Fermi", Course CXLVI*, edited by Quinn T. J., Leschiutta S. and Tavella P. (IOS Press, Amsterdam) 2001, p. 291.

[14] Kibble B. P. and Rayner G. H., *Coaxial AC Bridges* (Adam Hilger Ltd, Bristol) 1984.

[15] Macmartin M. P. and Kusters N. L., *IEEE Trans. Instrum. Meas.*, **15** (1966) 212.

[16] Klonz M., *IEEE Trans. Instrum. Meas.*, **36** (1987) 320.

[17] Laiz H., Klonz M., Kessler E. and Spiegel Th., *IEEE Trans. Instrum. Meas.*, **50** (2001) 333.

[18] Ramm G., Moser H. and Braun A., *IEEE Trans. Instrum. Meas.*, **48** (1999) 422.

[19] *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes* (International Bureau of Weights and Measures) 1999.

This page intentionally left blank

The application of the Josephson and quantum Hall effects in electrical metrology

B. Jeckelmann and B. Jeanneret

Swiss Federal Office of Metrology (METAS) - Lindenweg 50, 3003 Bern-Wabern, Switzerland

1. – Introduction

During the last century, the International System of Units —the SI— has evolved from an artefact-based system to a system mainly based on fundamental constants and atomic processes. The modern units have major advantages over their artefact counterparts: they do not depend on any external parameters like the ambient conditions and, most important, they do not drift with time. In addition they can be simultaneously realized in laboratories all over the world which strongly simplifies and improves the traceability of any measurements to the primary standards.

With the discovery of the Josephson and the quantum Hall effects, two electrical quantum standards became available. As an important consequence, the worldwide consistency in the realization and maintenance of the electrical units and the electrical measurements based on them has improved hundredfold in the last decade. The two quantum effects will certainly also play a major role in the next modernization of the SI when the last base unit still based on an artefact, the kilogram, will be linked to fundamental constants.

In 1962, Brian Josephson published a theoretical study on transport phenomena in weakly coupled superconductors [1]. The prediction of quantized voltage steps in such systems, called Josephson junctions, was experimentally confirmed by Shapiro [2]. When a Josephson junction is exposed to an electromagnetic radiation of frequency f, its current-voltage characteristic exhibits precisely quantized voltage steps (see fig. 1) described by the relation $V_n = nf/K_\mathrm{J}$ where n is the step number. K_J is the Josephson

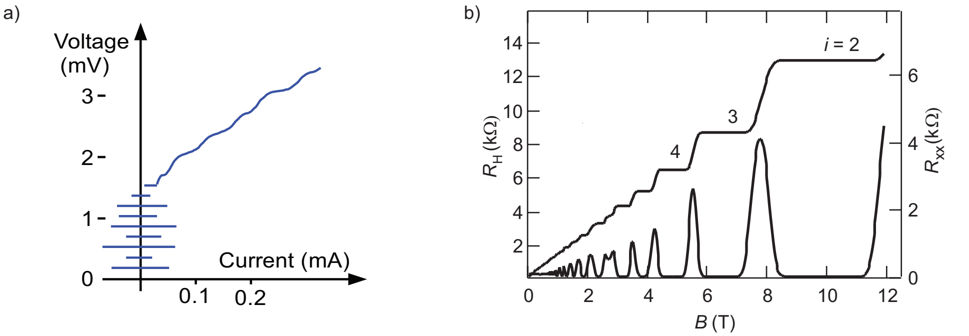


Fig. 1. – a) Current-voltage characteristic of a weakly damped Josephson junction. b) Measurement of the Hall resistance R_{H} and of the longitudinal resistance R_{xx} for a GaAs heterostructure at a temperature of 0.3 K.

constant which —according to the present theoretical and experimental evidence— is given by

$$K_{\mathrm{J}} = 2e/h, \tag{1}$$

where e is the elementary charge and h the Planck constant.

The quantum Hall effect (QHE) was discovered in 1980, when Klaus von Klitzing was investigating the transport properties of a Si-MOSFET device at very low temperature and high magnetic field in Grenoble [3]. The discovery, which was totally unanticipated by the physics community, relied on the existence of a two-dimensional electron gas (2DEG) in a semiconducting device. The great technological progress that followed the invention of the transistor led to the realization of the first 2DEG in semiconducting devices in the middle of the sixties. When a QHE device (an heterostructure or a MOSFET) is placed at low temperature in a strong magnetic flux B perpendicular to the plane of conduction, regions appear in the Hall voltage *vs.* magnetic flux curve for a heterostructure (or in the Hall voltage *vs.* gate voltage in case of a MOSFET) where the Hall voltage is constant over a certain range of B (or over a certain range of gate voltage). On these plateaus, the Hall resistance R_{H} is quantized and described by $R_{\mathrm{H}} = R_{\mathrm{K}}/i$. R_{K} is the von Klitzing constant and i is the plateau number. The longitudinal voltage along the probe displays a markedly oscillatory behaviour (Shubnikov-de-Haas effect). The plateaus in the Hall voltage thereby fall together with extended minima in the longitudinal voltage (see fig. 1). As the temperature is lowered, the voltage in these minima becomes so small as to be unmeasurable, and consequently, as absolute zero is approached, the current flow through the probe shows zero dissipation. There is an impressive amount of evidence supporting the relation

$$R_{\mathrm{K}} = h/e^2. \tag{2}$$

In the next section, the implication of the discoveries of the two quantum effects will be summarized. The following sections will be dedicated to a short review of the application of the Josephson and quantum Hall effect in electrical metrology.

2. – The conventional system of electrical units

The Josephson and quantum Hall effects can be used to realize very reproducible voltage and resistance values which, to our knowledge, depend only on natural constants. To be used as practical standards, the value of the Josephson and von Klitzing constants have to be known in SI units. In the SI, the electrical units are defined in terms of the mechanical base units metre, kilogram and second through the definition of the ampere and the assumption that electrical power and mechanical power are equivalent. To put the concept of the electrical units in practice, it is sufficient to realize two electrical units in terms of the metre, kilogram and second. At present, the ohm and the watt are the two chosen units, since they are the most accurately determined.

2·1. *The determination of R_K and K_J.* – To measure the von Klitzing constant, the quantized Hall resistance (QHR) has to be compared to a resistance standard whose value is known in SI units. In practice, the unit ohm is realized by means of a calculable cross capacitor based on an electrostatics theorem discovered in 1956 by Thompson and Lampard [4]. When the theorem is correctly put into practice, the cross capacitance depends only on the capacitor length. Using a.c. bridge techniques, the capacitance of the calculable capacitor is scaled to a value which can be compared to the resistance of an a.c. resistor using a quadrature bridge. After proper scaling, this a.c. resistor is compared to another a.c. resistor which has a small and calculable a.c./d.c. difference. D.c. techniques are finally applied to link the calculable resistor to the QHR.

Despite the long and complicated measurement chain, an accuracy of a few parts in 10^8 is reached using this method [5-7].

There is an important consequence of the QHE in the field of fundamental constants which should also be addressed here. The von Klitzing constant R_K is related with the fine-structure constant through the simple relation

$$R_\mathrm{K} = \frac{h}{e^2} = \frac{\mu_0 c}{2\alpha}\,. \tag{3}$$

In the SI, the permeability of vacuum μ_0 and the speed of light c are fixed quantities with $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{NA}^{-2}$ and $c = 299\,792\,458\,\mathrm{m\,s}^{-1}$. The fine-structure constant can thus be used to determine R_K and test possible corrections to the QHR. Conversely, if R_K is assumed to be identical to $i \cdot R_\mathrm{H}(i)$, the QHE opens up an additional route to the determination of α which does not depend on QED calculations. In fig. 2, all the results are shown which contributed to the least-square adjustment of α, as given in the 2002 set of fundamental physical constants recommended by the CODATA task group [8].

At present, the most accurate value for α is derived from the anomalous magnetic moment a_e of the electron measured using single electrons or positrons stored in a Penning

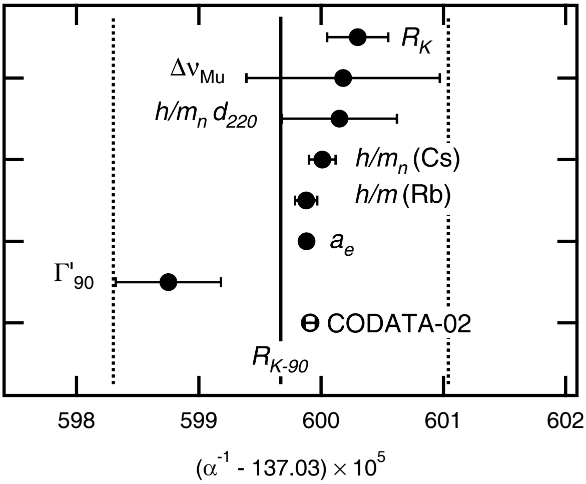


Fig. 2. – Values for the fine-structure constant taken into account in the 2002 adjustment of fundamental constants [8]. The vertical lines indicate the value corresponding to $R_{\text{K-90}}$ and its uncertainty. Γ'_{90} is the value from the measurement of the gyromagnetic ratio of the shielded proton; $\Delta\nu_{\text{Mu}}$ is related to the muonium ground-state hyperfine splitting, a_{e} to the anomalous magnetic moment of the electron. h/m_x is the ratio of the Planck constant to various atomic masses.

trap at 4.2 K and exposed to a magnetic flux [9]. A relative experimental uncertainty of 3.7×10^{-9} has been reached so far [8]. A value for the fine-structure constant can be obtained from the experimental value of a_{e} by comparing it to the theoretical value which can be, up to some insignificant correction terms due to electroweak and hadronic interactions, expressed in the framework of quantum electrodynamics as a power series in α. The most important terms in the series can be calculated analytically, but for some of the higher-order terms extensive numerical calculations are necessary [10]. The uncertainty of the theoretical calculation of a_{e} is estimated to be 1 part in 10^9 [8].

The second most important result taken into account in the calculation of the actual value for α comes from the realization of R_{K} through the calculable capacitor assuming $R_{\text{H}}(i = 1) = R_{\text{K}}$. As the comparison shows, R_{K} and the a_{e} derived value for α agree only fairly within the experimental uncertainty.

The Josephson constant K_{J} can be determined by comparing the Josephson voltage to a voltage standard known in terms of the SI unit volt. The volt can be realized directly in an electromechanical experiment where an electrostatic force arising from a voltage is counterbalanced with a known gravitational force. The accuracy of these experiments (see [8] for a review) is limited to approximately $0.6\,\mu\text{V/V}$.

A more accurate route to K_{J} is the watt balance experiment [11] in which electrical and mechanical power are compared. If the electrical power is measured in terms of the Josephson voltage and the quantized Hall resistance, the product $K_{\text{J}}^2 R_{\text{K}}$ is determined in the experiment. The most accurate result so far was obtained at the National Institute of Standards and Technology (NIST) [12] with an uncertainty of 5 parts in 10^8 for the product $K_{\mathrm{J}}^2 R_{\mathrm{K}}$.

2·2. *Conventional values for R_{K} and K_{J}.* – The best realizations of the volt and the ohm in the SI are about two orders of magnitude less accurate than the reproducibility of quantum standards based on the Josephson and the quantum Hall effects. Two electrical units realized in terms of the non-electrical SI units metre, kilogram and second are needed to make the other electrical units measurable in the SI. With the QHE and the Josephson effect, two fundamentally stable standards are available and thus it was realized that the worldwide consistency of electrical measurements could be improved by defining conventional values for R_{K} and K_{J}. The Comité Consultatif d'Électricité (CCE) was asked to recommend such values based on the data available. All the values for R_{K} and K_{J} available by June 1988 in SI units were analysed and the following conventional values were proposed [13]:

$$\begin{aligned} R_{\mathrm{K\text{-}90}} &= 25812.807\,\Omega\,,\\ K_{\mathrm{J\text{-}90}} &= 483597.9\,\mathrm{GHz/V}. \end{aligned}$$

Relative uncertainties with respect to the SI of 2×10^{-7} and 4×10^{-7}, respectively, were assigned to the two values. The conventional values were accepted by all member states of the Metre Convention and became effective as of January 1, 1990. Due to further progress in the experiments, the assigned uncertainty for R_{K} with respect to the ohm was reduced in 2000 by a factor of two to 1×10^{-7}.

In the case of $R_{\mathrm{K\text{-}90}}$, the value chosen was essentially the mean of the most accurate direct measurements of R_{K} based on the calculable capacitor and the value from the calculation of the fine-structure constant based on the anomalous magnetic moment of the electron [13]. In the most recent least-square adjustment of fundamental constants carried out by the CODATA Task Group on Fundamental Constants [8], a value of $R_{\mathrm{K}} = 25812.807449\,\Omega$ with a relative uncertainty of 3.3 parts in 10^9 was evaluated. This new value is in good agreement with the conventional value, $R_{\mathrm{K\text{-}90}}$. Figure 2 shows the results that were taken into account in the calculation of the new R_{K} value and consequently the new recommended value for α.

In the case of $K_{\mathrm{J\text{-}90}}$, the value chosen was dominated by the watt balance result obtained at the National Physical Laboratory (NPL) [14] and the value of R_{K}. In the CODATA 2002 adjustment, a value of $K_{\mathrm{J}} = 483597.879\,\mathrm{GHz/V}$ with a relative uncertainty of 8.5 parts in 10^8 was evaluated. Again, this is in a very good agreement with the conventional value $K_{\mathrm{J\text{-}90}}$.

3. – The Josephson voltage standard

The Josephson array voltage standards development started with the discovery of the Josephson effect in 1962. Nowadays numerous Josephson voltage standards are in use around the world in national, industrial and military standard laboratories. These standards can reach a voltage of 10 V with an uncertainty which is typically smaller than 1 part in 10^9. The development, design and operation of the Josephson voltage standard has been the subject of many detailed review papers [15-22]. The present chapter is rather a short and basic introduction to the subject, including the new development related to the application of the Josephson effect in a.c. voltage metrology.

3˙1. *Theoretical background of the Josephson effect.* – In 1962, Josephson [1] predicted several effects associated with the tunnelling of Cooper pairs in a junction consisting of two superconducting electrodes separated by a thin insulating barrier. In particular, when such an ideal junction is connected to an external source, the current flow through the junction is described by the two equations

$$I = I_{\rm c} \sin\varphi\,, \tag{4}$$

$$V = \frac{\hbar}{2e}\frac{{\rm d}\varphi}{{\rm d}t} = \frac{h}{2e} f_{\rm J}\,, \tag{5}$$

where φ denotes the phase difference between the two macroscopic wave functions of the superconducting electrodes, $I_{\rm c}$ is the critical current of the junction and $\hbar = h/2\pi$. The first equation (d.c. Josephson effect) implies that a current can flow without a d.c. voltage across the junction as long as the current is smaller than the critical current. If the critical current is exceeded, a voltage appears across the junction which gives rise to an alternating current of frequency $f_{\rm J}$ (a.c. Josephson effect). Conversely, irradiation of the junction with microwaves of frequency f produces steps of constant voltage V_n due to the phase locking of the Josephson oscillator to the external frequency

$$V_n = n\frac{h}{2e} f, \tag{6}$$

where n is the step number. These voltage steps observed for the first time in 1963 by Shapiro [2] form the basis of the quantum voltage standard.

In a real Josephson junction, the ideal junction is always shunted by its own capacitance C and resistance R. The dynamic of such a junction is often investigated by the so called RCSJ model [23, 24]. The second Josephson equation is modified to take into account the current flow in the resistance and the capacitor. The equation describing the circuit when the junction is biased by both a d.c. current I_0 and an a.c. current $I_{\rm rf}$ of frequency $f = \frac{\omega}{2\pi}$ is

$$\frac{\hbar C}{2e}\frac{{\rm d}^2\varphi}{{\rm d}t^2} + \frac{\hbar}{2eR}\frac{{\rm d}\varphi}{{\rm d}t} + I_{\rm c}\sin\varphi = I_0 + I_{\rm rf}\sin(\omega t). \tag{7}$$

This model properly describes the behaviour of the junction when the current is uniformly distributed over the junction area: $I_{\rm c} = wlJ_{\rm c}$, where w is the junction width, l is the junction length (in the direction of the current) and $J_{\rm c}$ is the critical current density. The dynamic of the junction is thus described by a strongly non-linear second-order differential equation. Such non-linear systems are prone to show chaotic behaviour

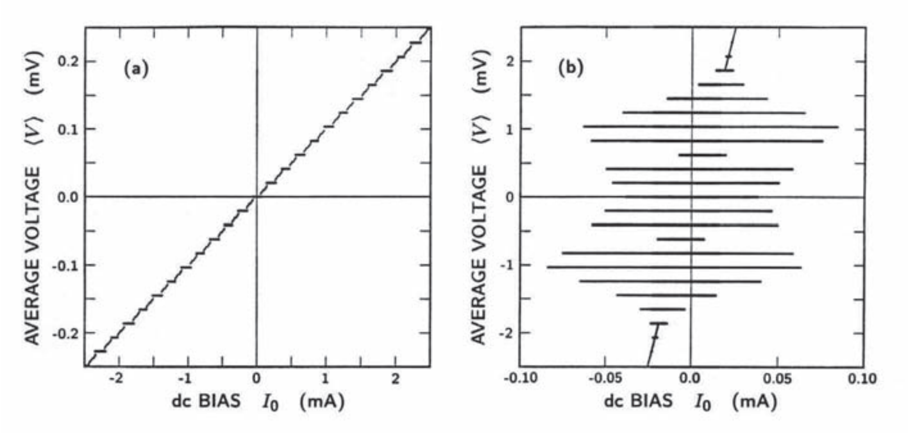


Fig. 3. – Simulated current-voltage curve computed using the Stewart-McCumber model in the limit (a) $\beta_\mathrm{c} \leq 1$ and (b) $\beta_\mathrm{c} \gg 1$ (after [25]).

(see [25] for a review) which must be avoided for metrological applications by a careful optimization of the junction parameters.

For small angle, eq. (5) becomes $V = (\hbar/2eI_\mathrm{c})\mathrm{d}I/\mathrm{d}t = L_\mathrm{J}\mathrm{d}I/\mathrm{d}t$, where $L_\mathrm{J} = \hbar/2eI_\mathrm{c}$ is the kinetic inductance of the ideal junction. In this case, the Stewart-McCumber model is a LCR resonator circuit with a resonance frequency $\omega_\mathrm{p} = (L_\mathrm{J}C)^{1/2} = (2eI_\mathrm{c}/hC)^{1/2}$ called the plasma frequency. A fundamental parameter of the junction is the McCumber parameter β_c defined as the square of the quality factor of the LCR resonator $Q = R(L/C)^{1/2}$:

$$\beta_\mathrm{c} = \frac{2e}{h} I_\mathrm{c} R^2 C. \tag{8}$$

In the limit $\beta_\mathrm{c} \gg 1$ the junction is underdamped and shows an hysteretic I-V curve (see fig. 3b); such junctions are used in conventional Josephson voltage standards. In the opposite limit $\beta_\mathrm{c} \leq 1$, the junction is overdamped and its I-V curve is single valued (see fig. 3a); such junctions are at the heart of the newly developed programmable voltage standards.

When the junction is phase locked to the microwave current, the supercurrent is forced to oscillate at the frequency f (or any of its higher harmonics nf). This synchronization of the junction to the external current generates voltage steps V_n (given by eq. (6)) in the I-V curve. These steps occur over a range of d.c. current ΔI_n (step width) given by the n-th–order Bessel function J_n:

$$\Delta I_n = 2I_\mathrm{c}|J_n(2eV_\mathrm{rf}/hf)|, \tag{9}$$

where V_rf denotes the amplitude of the radio frequency voltage across the junction.

The accuracy of the voltage-frequency relation was tested in different types of junctions and arrays [26-30]. These highly precise and accurate experiments were based on a method using a SQUID magnetometer [31]. The most precise measurement to date has been obtained by comparing two overdamped Josephson junctions. The difference in voltage when the junctions were biased on the same voltage step was found to be smaller than 3×10^{-19} [28].

3˙2. *Conventional Josephson voltage standard.* – In the early days, the voltage standard consisted of single junctions which provided only small voltages (5–10 mV). Although the stability of the single junction standard already exceeded the stability of the primary Weston cell standard, comparing the Weston cell to the Josephson standard required a precise voltage divider that was difficult to calibrate with the required accuracy. Therefore, attempts were made to increase the Josephson voltage output by connecting several junctions in series. The most ambitious project [32] used 20 junctions in series to produce a voltage of 100 mV with an uncertainty of a few parts in 10^9. A number of 20 individually adjustable current sources was needed to ensure that each junction remained on the appropriate voltage step. The difficulty of the tuning procedure brought this approach to an end.

Finally, the multiple bias problem was solved using a suggestion made by Levinsen [33] in 1977. Levinsen showed that a junction with a large McCumber parameter ($\beta_{\mathrm{c}} > 100$) can generate an hysteretic I-V curve with voltage steps that cross the zero current axis, hence their name of zero-crossing steps (see fig. 3b). The lack of stable regions between the first few steps shows that the voltage of the junction must be quantized, at least for small current bias.

After the problems of junction stability and microwave power distribution were solved, the first large array based on the Levinsen idea was fabricated [34], leading to the first practical 1 V Josephson standard in 1985 [35, 36]. Improvements in the superconductive integrated-circuit technology allowed the fabrication of the first 10 V array in 1987 [37]. This array consisted of 14 484 junctions that generated about 150 000 quantized voltage steps spanning the range between −10 V and 10 V. The 10 V Josephson voltage standard was then implemented in many National Metrology Institutes (NMI). The accuracy of theses standards is determined by international comparisons between the transportable Josephson system of the Bureau International des Poids et Mesures (BIPM) and those of the NMI. Typically, the difference between two quantum standards is less than 1 part in 10^9 at a voltage of 10 V. The best comparisons, however, have uncertainties on the order of a few parts in 10^{11} [38].

In the next paragraphs, the Josephson standard will be described in more details.

3˙2.1. Junctions and arrays design. Nowadays, all the SIS junctions fabricated for application in voltage metrology are based on Nb/Al_2O_3/Nb structures (see fig. 4). Developed during the ’80 s, this technology has several advantages:

- The sputtering of the sandwich forming the junction can be performed without braking the vacuum. This ensures very clean interfaces and allows fabrication of an extremely thin and homogeneous junction barrier.

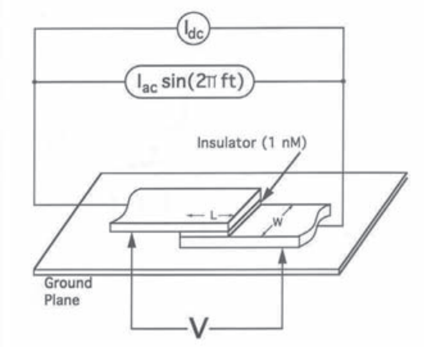


Fig. 4. – Schema of a typical SIS junction used in an large array (after [19]).

- Using Nb, the junctions are mechanically and chemically stable. This was not the case with the lead junctions used earlier. As a result, no aging of the Josephson arrays is observed.

- As the critical temperature of Nb is 9 K, the circuit can be operated in liquid He at a temperature of 4.2 K. At a temperature of half the critical temperature, all the superconducting parameters have approached their $T = 0$ value.

The most important condition for accurate measurements using a Josephson voltage standard is the stability of the phase lock between the microwave current and the Josephson oscillator. This phase lock must be strong enough to prevent the array from frequently jumping from one voltage step to another during the duration of a calibration. On the basis of the McCumber model, Kautz analyzed how the various junction parameters influence the stability of the phase lock with regard to chaos, thermal noise and uniformity of the current distribution (see [16, 25] for a review). Four conditions are required for a stable operation of the voltage standard:

1) The junction length l must be small enough so that the flux created by the microwave current over the junction's surface is much less than the flux quantum $\phi_0 = h/2e$.

2) Both the junction width w and length l must be small enough so that the lowest resonant cavity mode of the junction is greater than f.

3) To avoid chaotic behaviour, the plasma frequency must satisfy the relation $f_\mathrm{p} < f/3$. Since $f_\mathrm{p} \sim J_\mathrm{c}^{0.5}$, the critical current density is limited to $J_{\mathrm{c\ max}} = (f/3)^2(\pi h C_\mathrm{s}/e)$, where C_s is the specific capacitance of the junction $C_\mathrm{s} = C/wl$. Together with the limitation of the first and second condition, the critical current is therefore limited to $I_{\mathrm{c\ max}} = w_{\max} l_{\max} J_{\mathrm{c\ max}}$ which in turn limits the maximum step width to $\Delta I_{n\ \max} = 2I_{\mathrm{c\ max}} |J_n(2eV_\mathrm{rf}/hf)|_{\max}$.

4) The critical current should be as large as possible to prevent noise-induced step transitions, in other words, the coupling energy of the junction $E_\mathrm{J} = \hbar I_\mathrm{c}/2e$ must be larger than the thermal fluctuations kT.

Table I. – *Junction design parameters (after* [19]*).*

Junction material	Nb/Al_2O_3
Critical current density J_c	20 A/cm^2
Junction length l	18 μm
Junction width w	30 μm
Critical current I_c	110 μA
Plasma frequency f_p	20 GHz
Lowest resonant cavity mode	175 GHz
Microwave frequency f	75 GHz
Specific capacitance C_s	5 μF/cm^2

The conditions three and four are clearly antinomic. Therefore, the stability of the array is caught in a region of the parameter space between instabilities due to thermal noise or chaos. However, an optimized design can lead to an excellent stability which is sufficient for most of the d.c. calibration work. As an example, the set of parameters given in table I for a typical 10 V array ensures a stability that can reach several hours under appropriate conditions.

For a 10 V array, the 20 208 junctions form a series array as shown in the schematic of fig. 5. The microwave power is collected by a finline antenna, split 16 ways, and injected into 16 segments, forming each a stripline of 1263 junctions. The most important consideration in the design of the array is that each junction must receive the same microwave power in order to develop the largest possible zero-crossing steps. The maximum number of junctions per segment is limited by the attenuation of the stripline. Microwave reflection at the end of each stripline is suppressed by an optimized resistive load. To meet the appropriate packaging density, the striplines are folded taking into account that the microwave bend radius has a minimum value of three times the stripline width. The d.c. voltage appears across superconducting pads at the edge of the chip.

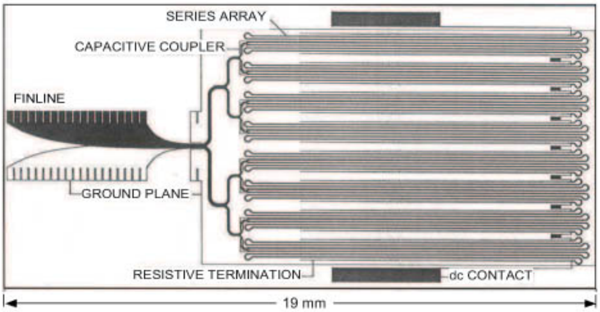


Fig. 5. – Schema of a typical 10 V NIST array (after [19]). This design is the result of a joint NIST/PTB effort (see [18, 21]).

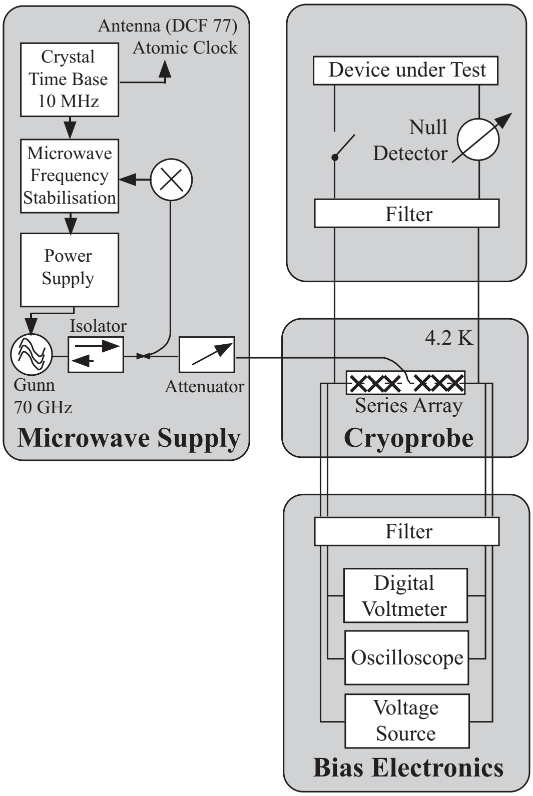


Fig. 6. – Block diagram of a typical Josephson voltage standard (after [21]).

3˙2.2. Measurement system. A block diagram of a typical Josephson voltage standard is shown in fig. 6 (after [21]). The array is mounted in a magnetically shielded cryoprobe fitted with a WR-12 waveguide and three pairs of heavily filtered wires. The cryoprobe is immersed in liquid helium at 4.2 K. The microwave power is provided by a Gunn diode which operates at a frequency of 70 to 90 GHz. The Gunn must have enough power to deliver around 15 mW at the chip finline for a 10 V array. An attenuator allows adjustment of the power to the array and a directional coupler diverts parts of the power to the phase locking counter which establishes the phase lock to the external frequency reference (most of the time a Cs clock or a GPS receiver).

A low-frequency triangle wave generator is used to trace the *I*-*V* curve of the array on the oscilloscope's screen. This allows the measurement of the critical current, the check of the *I*-*V* curve, and the visualization of the steps in order to optimize the power settings and to control whether all the junctions are on a quantized step. The tuning of the power is the most critical parameter adjustment (see [15] on how to proceed). A voltage source connected to the array through a variable resistor allows the selection and the stabilization of the desired voltage step.

The device under test (DUT) is connected to the third pairs of wires, most of the time through a switch or a scanner which allows the reversal of the polarity of the DUT and the measurement of the voltage difference between the array and the DUT with a null detector.

Although the voltage appearing at the Josephson array is, in principle, exact, the accuracy of a real Josephson voltage standard is limited by a large number of uncertainties. A list of all the identified sources of uncertainty is given below:

1) Reference frequency offset and noise

2) Leakage current in the measurement loop

3) Detector gain error

4) Detector bias current

5) Nanovoltmeter offset, input impedance, non-linearity and noise

6) Uncorrected thermal voltages

7) Rectification of the reference frequency current

8) Electromagnetic interference

9) Sloped steps (bias-dependent voltages).

In the above list, only the uncertainties 1) and 2) depend on the voltage being measured. This observation allowed Hamilton to develop a powerful method to collectively evaluate the uncertainties 3) to 8) by a sequence of short circuit measurements [39]. Therefore, the uncertainty budget of the Josephson voltage standard has finally only three components. For the 10 V METAS system, using a HP3458A DVM as the null detector, the uncertainty components have the following typical values: 0.7 nV for the frequency, 1.0 nV for the leakage current and 5.0 nV for the repeatability (uncertainty 3) to 8)). The combined standard uncertainty of the system is thus 5.1 parts in 10^{10}. This uncertainty can be reduced to a few parts in 10^{11} by using an analog nanovoltmeter [38].

3˙2.3. Application in d.c. voltage metrology. The most important application of the conventional Josephson voltage standard is the calibration of Zener-diode–based d.c. reference standards. Zener standards are convenient transportable voltage standards that are used to maintain the traceability chain to the primary Josephson standard [40] at 1.018 V and 10 V. The stability of the 10 V output of a Zener is around 10^{-6} per year. By carefully controlling the environmental conditions and by modelling temporal drift, output voltages can be predictable over periods of several weeks to within a few parts in 10^8. Ultimately, the uncertainty of the output voltage of a Zener standard is limited by a $1/f$ noise floor having a value comprised between two and ten parts in 10^9 [41, 42]. Nevertheless, using great care, standard uncertainties on the order of a few part in 10^8 have been achieved using Zeners as travelling standards in international comparisons

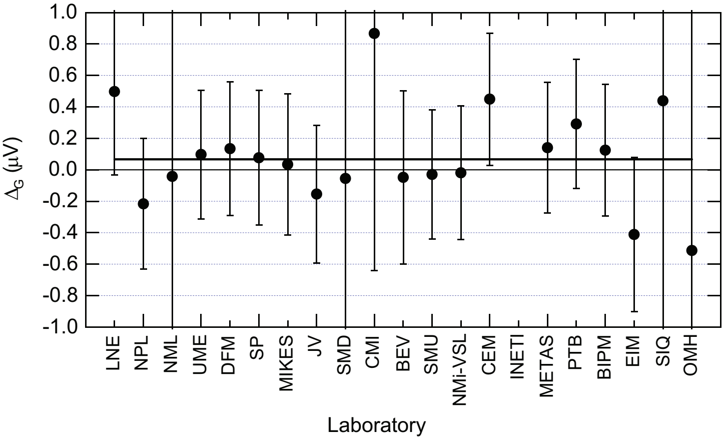


Fig. 7. – Result of the EUROMET 429 comparison. Δ_G is the value given by each participant —relative to 10 V— for a group of four Zener standards. The solid line is the reference value. See the BIPM database for more details.

(see [43, 44] and references therein). As an example, the results of an EUROMET comparison are presented in fig. 7. A group of four Zener standards carefully characterized by the BIPM was sent to the participants. The data represent the result given by each participant for the mean value of the group. The overall agreement is excellent, most of the results agree with the comparison uncertainty.

Another important application of the conventional Josephson voltage standard is the calibration and linearity measurement of high-precision digital voltmeters. An example of such a measurement is given in fig. 8a. A HP3458A DVM was used to read the output voltage of the Josephson standard for voltages ranging between -10 V and 10 V by step of 1 V. The gain of this instrument exceeds 1 by 0.9×10^{-6}. The linearity of the instrument which is given by the standard deviation of residuals from the fit is shown in fig. 8b. The linearity is 350 nV or in relative unit 3.5×10^{-8} which is really outstanding. During the development of this instrument, a Josephson voltage standard was used to characterize the linearity of the analog-to-digital converters. Clearly, such a linearity would have been impossible to achieve without the use of a Josephson voltage standard [45] in the development phase of the instrument.

3˙3. *Programmable voltage standards.* – The major difficulty when operating SIS Josephson junctions arrays is their inherent weak stability. This lack of stability prevents them to rapidly and reliably switch between different target values of the voltage, limiting their application to d.c. calibration. These characteristics stem from the hysteretic nature of the I-V curve of SIS junctions. A way to avoid this drawback is to use non-hysteretic junctions obtained by shunting the junctions with a resistor, either externally or internally by using SNS (Superconductors-Normal metal-Superconductors) or SINIS (Superconductors-Isolator-Normal metal-Isolator-Superconductors) junctions.

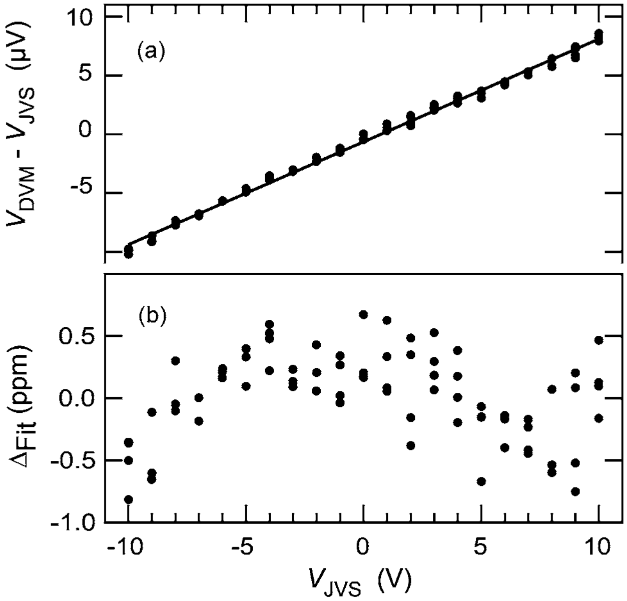


Fig. 8. – Measurement of the linearity of a commercial high-resolution DMM (HP 3458) in its 10 V range. Data taken at the Swiss Federal Office of Metrology (METAS).

In this case the *I*-*V* curve is single valued, providing an intrinsic stability. The price to pay for this stability is a far more complex bias electronics, since each junction must be individually biased on its specific voltage step. This approach was successfully demonstrated by Hamilton *et al.* [46] who fabricated a D/A converter with an array of 2048 SIS junctions externally shunted by 1 Ω resistors. This work has initiated a new area of development in the voltage metrology: the programmable voltage standard. Two different approaches have been considered so far:

In the first approach the array is divided in a binary sequence, each of the arrays' segment being independently biased. This configuration forms a fast d.c. programmable source by biasing the appropriate segments of the array. This source also allows the generation of a.c. waveforms with a predictable r.m.s. value by rapidly switching the voltage of the array segments. However, the transients occurring during the switching phase limit both the accuracy and the operation frequency.

In the second approach, the transient's problem is solved by using the array as a pulse quantizer. Trains of pulses —synchronized to the microwave frequency— are launched on the array. Theses pulses are quantized by the array since the time integral of the voltage is quantized in flux units (see eq. (5)). By carefully clocking the pulse train sent to the array, any waveform of predictable r.m.s. value can be generated.

The two different techniques mentioned above will be described in more detail in the following paragraphs.

3·3.1. Binary arrays. To develop a programmable voltage standard, the most important characteristics of the array are a good voltage stability as well as a fast selection of the voltage level desired. Due to the hysteretic I-V curve of SIS junctions, such goals are clearly impossible to achieve with conventional arrays. One way to reduce the McCumber parameter β_c to obtain non-hysteretic junctions is to use SNS junctions. In this case, the voltage steps are no longer metastable and can uniquely be selected by an appropriate choice of the bias current. The high critical current of such junctions provides a greater immunity to thermal and electrical noise. Their low characteristic voltage $V_c = I_c R$ leads to lower operating frequencies which is an advantage for the cost of the microwave electronics but a disadvantage for the maximum achievable voltage per junction. The small capacitance of the SNS junctions also helps to maintain $\beta_c < 1$.

In 1997, a 1 V binary array [47] was developed using $2\,\mu\text{m} \times 2\,\mu\text{m}$ Nb-PdAu-Nb junctions [48, 49] having a critical current of $I_c \sim 8\,\text{mA}$ and a resistance of $R \sim 3\,\text{m}\Omega$, giving a characteristic voltage of $V_c \sim 24\,\mu\text{V}$. The binary array consists of 32 768 junctions divided into 13 segments. The number of junctions N in each of the 13 segments is: 128, 128, 256, 512, 1024, 2048, and seven cells with 4096 junctions each (see the schematic of fig. 9). When the array is exposed to a 16 GHz microwave drive, each segment, biased with a current on the order of typically 15 mA, develops a constant voltage step at a value of $V = Nf/K_J$, where f is the microwave frequency. The steps are typically between 1 mA and 4 mA wide. Therefore, by selecting the appropriate frequency, segment configuration and bias current sign, it is possible to generate any voltage in the range from -1.1 V to -50 mV and from 50 mV to 1.1 V. The time necessary to commute to a different voltage value is limited by the switching time of the bias electronics. The accuracy of such a programmable array was checked by performing a comparison with a traditional Josephson standard [50]. The agreement between the two systems was $(1.4 \pm 3.5) \times 10^{-10}$ at 1 V. This agreement was confirmed in a later study [51].

The next challenge with these SNS arrays is to increase the maximum output voltage. Since only the first step can be used and since the operation frequency is around 16 GHz, the maximum voltage per junction is a factor ~ 25 smaller than with a traditional array. This means that a 10 V array will need around 400 000 junctions! The strategy is to use stacked junctions to increase the integration density.

The first kind of stacked arrays is based on Nb-MoSi$_2$-Nb junctions [52]. With such a technology, it is possible to fabricate double and even triple staked arrays [53] which can actually deliver voltages up to 2.6 V and 3.9 V, respectively [54, 55].

Another type of junctions for stacked arrays is based on the NbN technology and the multilayer for a double stack has the following composition: NbN/TiN$_x$/NbN/TiN$_x$/NbN. One of the main interests of this system is its high critical temperature ($T_c \sim 16\,\text{K}$) which allows the operation of the array using a 10 K cryocooler, avoiding the costly use of liquid helium. In 2005, a 1 V array was fabricated [56] with this technology. This array showed perfect operation at 4.2 K. A comparison with a traditional SIS array showed an agreement better than 1 nV at a level of 1 V. Further improvement, including the double-stack technology, allowed to fabricate an array with 307 200 fully operational junctions working at a temperature of 10.2 K. Preliminary

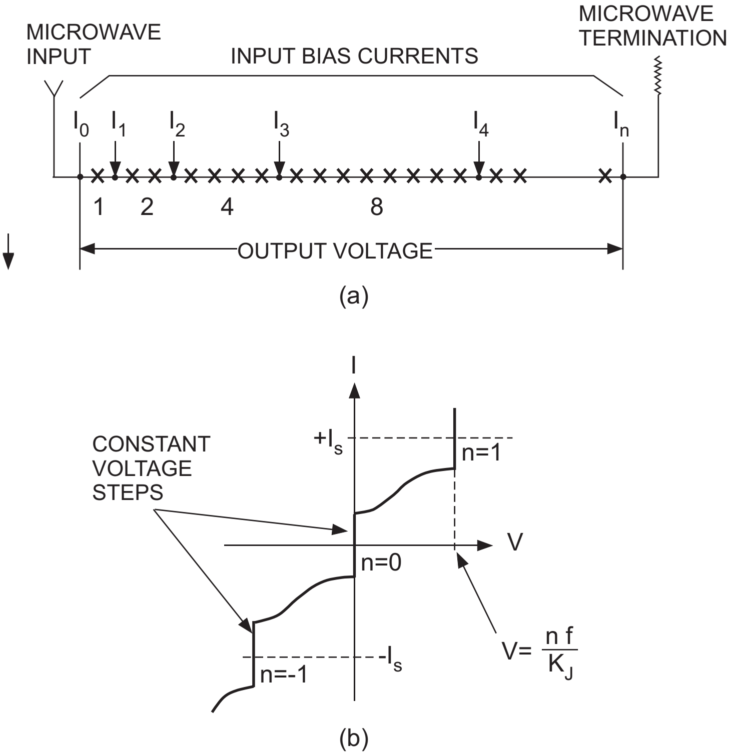


Fig. 9. – a) Schematic design of a programmable voltage standard based on a binary divided array. b) The *I*-*V* curve of a single junction with the microwave power set to equalize the $n = 0$ and $n = \pm 1$ steps.

measurement showed a voltage step at 10 V, the width of this step being larger than 1 mA [57, 58]. This circuit is certainly the most complicated superconducting circuit fabricated ever!

Another very interesting and successful design for binary arrays is based on the so-called SINIS junctions consisting of a multilayer of Nb/Al/AlO$_x$/Al/AlO$_x$/Al/Nb [59,60]. Such junctions have a critical current of $I_{\mathrm{c}} \sim 1.5\,\mathrm{mA}$ and a resistance of $R \sim 100\,\mathrm{m}\Omega$, giving a characteristic voltage of $V_{\mathrm{c}} \sim 150\,\mu\mathrm{V}$. These parameters lead to a McCumber parameter around 1 meaning that the junctions have a non-hysteretic intrinsically stable *I*-*V* curve. The high value of V_{c} leads to an operating frequency of 70 GHz. The fabrication of the first large 1 V array succeeded in 1999 [61,62] and an improved design was developed in 2002 [63]. The arrays contain 8192 junctions binary divided into 14 bits. The least significant bit consists of a single junction giving a resolution of 150 μV for the entire array. The first precision measurement performed was a comparison with a traditional array [61]. The comparison showed that the voltage steps are flat to a resolution of better than 1 nV over a current range of $\sim 200\,\mu$A. At the 1 V level, a good

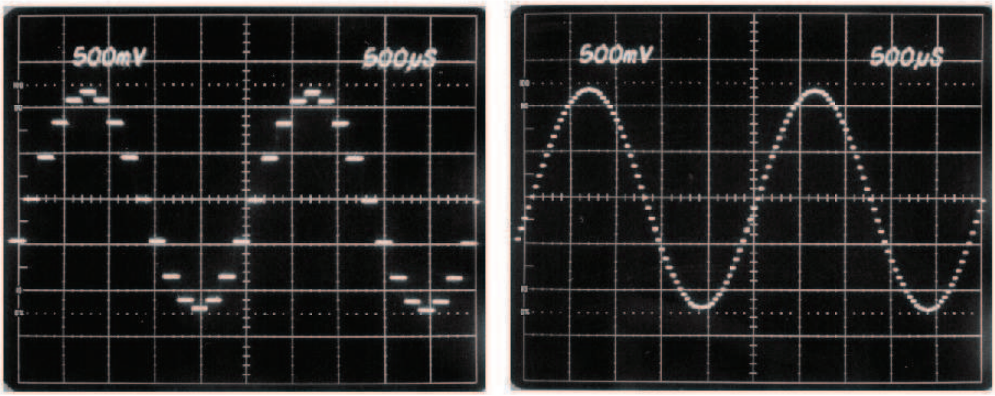


Fig. 10. – A sine wave synthesized with a 1 V binary SINIS array at a frequency of 400 Hz. Left side: 16 samples; right side: 64 samples (after [75]).

agrement with the SIS array was found with a measurement uncertainty of 2×10^{-10}. Later on, arrays were sent to different metrology institutes for precision measurements and the agreement mentioned previously was confirmed by all the participants [64, 51].

When the first SINIS array was measured, it was found that the microwave power required to phase lock the array was very low, around $100\,\mu$W for a 7000 junctions array, which corresponds to a reduction factor of around 40 in comparison with a traditional SIS array. This observation immediately suggested the possibility to extend the maximal voltage output to 10 V. Indeed, the first 10 V arrays were fabricated in 2000 [65,66]; they consist of 69 120 junctions integrated in 64 parallel branches with 1080 junctions each. The I-V characteristic revealed a voltage step —200 μA wide— at a voltage level of 10 V. The fabrication of arrays of that size is at the limit of the present day technology.

The first realization of a binary array was achieved using externally shunted junctions [46]. This approach is being further pursued and 1 V arrays were fabricated in 2001 using a frequency-dependent damping of the junctions [67, 68]. This type of arrays was compared to traditional SIS standards and the agreement was found to be better than 1 nV at a level of 1 V [64, 51].

Since these binary arrays are much easier to operate than the traditional system due to their inherent stability and their fast voltage selection capability, they immediately found applications in many different areas of importance for high-precision electrical calibration: DVM linearity measurements [47, 50], Zener standard calibrations [50], fast reversed d.c. measurements of thermal converters [69, 70], potentiometric systems [71], quantum voltmeters [72] and watt balance experiments [12, 73].

Originally, the binary arrays were developed to synthesize any waveforms with a calculable r.m.s. value [46] by rapidly switching between the different voltage steps available (see fig. 10 for an example). However, it was soon realized that the transients occurring during the switching of the bias electronics constitute a major limiting factor [74]. The array voltage during the step transition is undefined and this, of course, introduces errors in the computed r.m.s. value of the signal. Nevertheless a bias electronics with a rise time of 250 ns between steps with virtually no overshooting or ringing has been designed. Using a binary 1 V SINIS array biased with this high-speed source, it was possible to synthesize sine waves with an uncertainty smaller than 1 part in 10^7 at frequencies below 200 Hz [75]. This uncertainty was obtained by performing a comparison with a multi-junction thermal converter. In the near future, 10 V SINIS arrays will be available and faster bias electronics will be developed (a model with 25 ns rise time is already in use) expanding the voltage and frequency range of the waveforms. This result shows that binary arrays can already be useful in a.c.-d.c. measurement at low frequencies (primarily in the sub-kHz regime) and at voltages below 10 V, exactly the range where the performances of thermal converters decline. Finally, this a.c. SINIS 1 V source was also used to characterize dynamic parameters of a high resolution analog-to-digital converter used in high end digital multimeters [76].

3˙3.2. Pulse-driven arrays. As briefly mentioned in the previous paragraph, the major limiting factors using binary arrays to synthesize a.c. waveforms are the transient voltages that occur during the step transition. In 1996, a different approach based on so-called pulse driven arrays was proposed as a solution to this problem [77] for high-frequency applications. The basic principle is to use the Josephson junction as a pulse quantizer since the time integral of the voltage across the junction is quantized in multiples of the flux quantum $\phi_0 = \frac{h}{2e} = K_{\mathrm{J}}^{-1}$. Therefore, if a pulse train of frequency f_{p} is launched through an array of N junctions, an average voltage $V = nNK_{\mathrm{J}}^{-1}f$ will appear across the array, where n is the number of flux quantum crossing the junction for each applied current pulse [78, 79].

This principle is illustrated in the schematic of fig. 11 for a bipolar a.c. waveform source (after [80]). First, the array is biased with a sinusoidal microwave current $I_{\mathrm{a.c.}}$ of frequency f (upper part of fig. 11b), leading to the I-V characteristic shown in fig. 11a. Then a sequence of pulses $I_{\mathrm{d.c.}}$ (middle trace of fig. 11b), synchronized to the sinusoidal drive, is applied to the array. The sequence of the quantized pulses appearing across the junctions is shown in the bottom part of fig. 11b. If $I_{\mathrm{d.c.}} = +I_0$ for exactly one period of $I_{\mathrm{a.c.}}$ then a single positive polarity voltage pulse occurs across the array. Of course, a negative polarity voltage pulse appears if $I_{\mathrm{d.c.}} = -I_0$ during exactly one period. Time-averaged voltages at values between these two extrema are generated using periodic sequences of pulses of appropriate number and polarity. An example of a six bit 011101-bit pattern is shown in the middle trace of fig. 11b. In case this pattern is continuously repeated, a d.c. voltage of $Nf/3K_{\mathrm{J}}$ will be generated across the array. In a similar way, time-dependent voltages are generated by repeating complex bit patterns: an example of a sine wave is shown in fig. 12a.

The block diagram describing the Josephson a.c. waveform source is shown in fig. 12b. The modulator is a computer algorithm that digitizes the input waveform $S(t)$ (frequency f_1) and creates a digital code S_i of length N_{s} at a sampling frequency $f_{\mathrm{s}} = N_{\mathrm{s}}f_1$. For repetitive waveforms, the code is calculated only once and stored in the circulating memory of the digital code generator. The digital code generator recreates this two-level code as a bipolar output voltage in real time $S_{\mathrm{D}}(t)$ by clocking its memory at

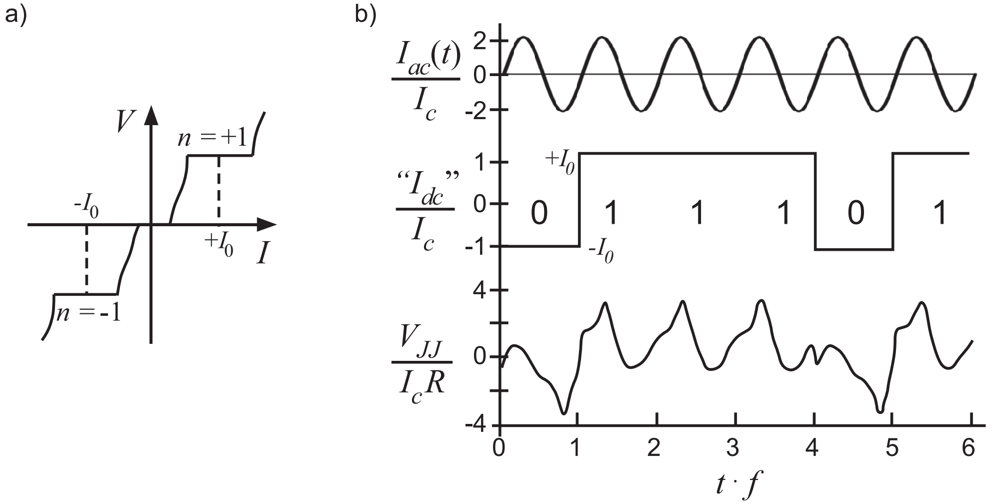


Fig. 11. – Principle of a bipolar a.c. waveform voltage source (after [80]). a) The I-V curve a non-hysteretic junction biased with a sinusoidal current at frequency f. b) Top trace: sinusoidal current drive at frequency f; middle trace: sequence of the current pulses applied to the junction; bottom trace: bipolar quantized voltage pulses appearing across the junction.

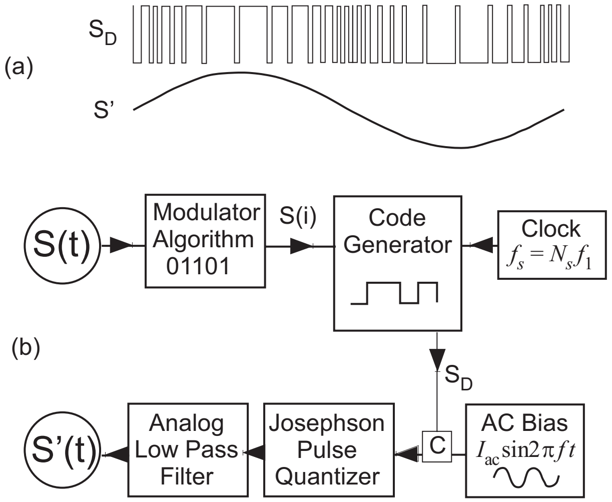


Fig. 12. – a) Two-level high-speed code S_{D} representing the synthesized sine wave S'. b) Block diagram of the bipolar Josephson voltage source (after [80]). C is a directional coupler.

the sampling frequency. The two-level high-speed code is combined with the sinusoidal drive (at frequency f) to bias the Josephson junction array. The quantized signal $S_\mathrm{J}(t)$ is then low-pass filtered to remove the unwanted quantization noise from the spectrum, leaving the desired waveform $S'(t) = S(t)$. A knowledge of the digital code, the sampling frequency and the number of junctions of the array is sufficient to exactly compute the spectrum and the r.m.s. value of the output signal $S'(t)$.

The original design of the pulse-driven source was based on a unipolar source [81, 78] able to generate voltages of a few mV. The introduction of the bipolar source allowed to extend the output voltage by a factor of 6 [80]. Advances in circuit design and fabrication enabled further refinements of the a.c. voltage source [82-84] which can actually synthesize waveforms with 240 mV peak voltage. Waveforms at both 3.3 kHz and 33 kHz were synthesized at this voltage with a harmonic distortion below 93 dBc (dB below the fundamental) [85]. Such a voltage level allows measurements with metrological accuracy (typically better than 1 part in 10^6) by comparing the a.c. source with multi-junction thermal converters [85].

Further increase of the output voltage requires a new approach based on lumped arrays [86, 22]. In a lumped array, all the junctions are placed within a quarter of a wavelength of the highest drive frequency. The idea is to fabricate a $50\,\Omega$ array to match the transmission line impedance. In this way most of the power will not be wasted in the termination resistor as in the distributed arrays. Such an array would allow to increase the output voltage by a factor of 8, bringing the a.c. pulse source close to 1 V. To reach this goal, the challenge is to fabricate a lumped array with 13 500 junctions spaced 120 nm apart! Today the junction spacing in the SNS pulsed arrays is around $7\,\mu$m. Therefore, a significant improvement is needed. There exist different schemes to reach such a large integration density. One of them is based on the multiple stack junctions briefly described in the previous paragraph [52, 53].

This newly developed a.c. Josephson voltage source will certainly find a large application field in many areas of metrology in the near future. It is already in use to measure absolute temperatures with Johnson noise thermometry [87, 88]. In these experiments, the Josephson array is used as a calculable quantized noise source. The power of this source is compared to the voltage power across the resistor whose temperature has to be measured. The absolute temperature of the resistor was recently measured with an accuracy of $150\,\mu$K/K [89], a very promising result indeed.

4. – The quantum Hall resistance standard

The quantum Hall effect is used in many National Metrology Institutes as an invariant reference for resistance measurements. Over 30 QHE systems are in operation worldwide and comparisons have demonstrated that resistance artifacts can be calibrated with a reproducibility on the order of 1 part in 10^9. Many aspects of the metrological application of the QHE have been covered in a comprehensive review article published recently [90]. The present section is a summary of this paper.

4·1. *Basic principles.* – The classical quantum Hall effect discovered in 1879 [91] is the starting point for the explanation of the QHE. A current I is flowing through a sheet of conducting material perpendicular to a magnetic flux B_z. As a consequence of the Lorentz force acting on the charge carriers, a Hall voltage perpendicular to the current and the magnetic flux is measured. The voltage depends on the carrier density n_s and the thickness of the conductive sheet. In the two-dimensional case, the Hall voltage U_H is independent of the geometrical dimensions

$$U_\mathrm{H} = \frac{B_z \cdot I}{n_s \cdot e}\,. \tag{10}$$

A two-dimensional electron gas (2DEG) can, *e.g.*, be realized at the semiconductor insulator interface of a Si-MOSFET (Metal Oxide Field Effect Transistor) or in a GaAs-AlGaAs heterostructure cooled down to very low temperatures. In the following description, the 2DEG is assumed to be ideal at zero temperature with no impurities and no electron-electron interactions. Due to the magnetic field the carriers perform cyclotron motions with angular frequency

$$\omega_\mathrm{c} = \frac{eB_z}{m^*}\,, \tag{11}$$

where m^* is the effective electron mass ($m^* = 0.068m_\mathrm{e}$ in GaAs, $0.19m_\mathrm{e}$ in Si, respectively). In the quantum-mechanical description, the Schrödinger equation leads to a shifted harmonic oscillator solution where the energy eigenvalues are given by (see fig. 13)

$$E_n = \hbar\omega_\mathrm{c}\left(n + \frac{1}{2}\right), \quad n = 0, 1, 2, 3 \ldots . \tag{12}$$

The spin splitting is ignored in this expression. The magnetic length $l = \sqrt{\hbar/(eB_z)}$ is the fundamental length scale of the problem. Every electron occupies the area πl^2. Taking the boundary conditions into account (length L and width w of the 2DEG), it can be shown that the orbital degeneracy is given by $N = Lw/(2\pi l^2)$. Inserting the expression for the magnetic length, the density of states n_B becomes the number of flux quanta within the area of the sample

$$n_B = \frac{1}{2\pi l^2} = \frac{eB_z}{h}\,. \tag{13}$$

The filling factor i is defined as the carrier density divided by the density of states

$$i = \frac{n_s}{n_B}\,. \tag{14}$$

The Hall expression (eq. (10)) in the 2D case shows that a quantized Hall resistance $R_\mathrm{H} = U_\mathrm{H}/I$ is observed when i Landau levels are fully occupied

$$R_\mathrm{H} = \frac{B}{i \cdot Ne} = \frac{h}{e^2 i}, \quad i = 1, 2, 3 \ldots . \tag{15}$$

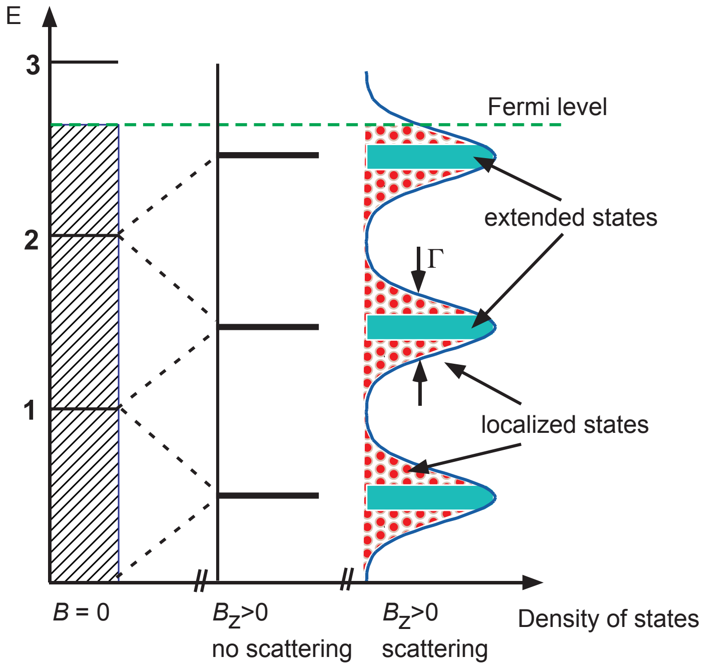


Fig. 13. – Landau quantization (spin interaction neglected) of a 2DEG in zero magnetic induction and when $B = B_z$.

Under this condition the longitudinal conductivity σ_{xx} in the direction of the current flow becomes zero ($\sigma_{xx} = 0$). The reason is that in the absence of scattering and with no electric field in the x-direction, the free electrons uniformly drift in the y-direction with no net velocity component along the x-axis. As a consequence of the relations between resistivity and conductivity tensor components in a two dimensional system,

$$(16) \qquad \rho_{xx} = \frac{\sigma_{xx}}{\left(\sigma_{xx}^2 + \sigma_{xy}^2\right)}, \qquad \rho_{xy} = \frac{-\sigma_{xy}}{\left(\sigma_{xx}^2 + \sigma_{xy}^2\right)},$$

the longitudinal resistivity vanishes at the same time ($\rho_{xx} = 0$).

The model just described does not explain the occurrence of wide plateaus for the quantized Hall resistance as observed in experiment (see fig. 1b). However, in real devices, due to disorder and scattering, the orbital degeneracy in the Landau levels is lifted. As a consequence, the Landau levels are broadened into bands. The bandwidth Γ should thereby necessarily remain smaller than the level spacing which sets an upper limit to the disorder allowed in a Hall device.

As a consequence of disorder, two different kinds of electronic states are formed: localized and extended states. The existence of the plateaus can be explained in this picture by the presence of the localized states in the tails of the distribution above and below each Landau level center (see fig. 13). When the Fermi level E_{F} resides within

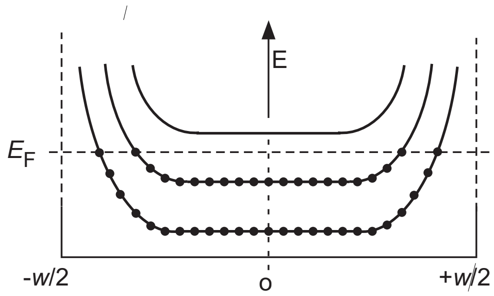


Fig. 14. – Energy spectrum of a 2DEG in a magnetic field with an infinite confining potential at the edges of the sample. States below the Fermi energy are occupied (solid dots). The edge channels are located at the intersection of the Landau levels with the Fermi energy.

the localized states, the current through the device is carried by the extended states. Due to the large separation between these states from empty states, no scattering takes place and the current flow is dissipationless ($\rho_{xx} = 0$). When E_{F} moves through the extended states, ρ_{xx} becomes non-zero and the transition from one plateau to the next occurs. The surprising fact is that —despite the localized states do not carry current— the resistance on a plateau is a universal quantity which corresponds to $h/(ie^2)$. Many attempts have been made to explain this experimental finding [92-94].

There have been alternative approaches to describe the physics of the QHE. A broad review can be found in [95]. Most of these models describe ideal systems at zero temperature. Real experiments, however, are carried out with samples of finite size at non-zero temperatures. For the current injection, source and drain contacts are needed which short out the Hall voltage at both ends of a device. Therefore, electrons enter and exit at diagonally opposite corners of the device and the source drain resistance equals the Hall resistance. As a consequence, an electrical power of $R_{\mathrm{H}}I^2$ is dissipated in the contacts. All these non-ideal features are difficult to model. Therefore, a complete quantitative theory which predicts, *e.g.*, deviations from the exact quantization under non-ideal conditions is still missing.

4˙2. *Edge state model.* – Most of the theoretical models put forward to explain the QHE, including the localization theory presented above and the elegant topological argument of Laughlin [96], are considering ideal systems with specific boundary conditions. Although these models provide clear and detailed insight in the physics of the 2DEG, resistance measurements performed in real samples rely on finite-size devices with imperfect electrical contacts to the 2DEG. Therefore, the question of whether these models properly describe the experimental situation, especially in high precision measurements, was quite open in the early days of the QHE.

At the edges of a real sample the confining potential produces an upward bending of the Landau levels (see fig. 14). For each Landau level intercepting the Fermi energy a one-dimensional edge channel is formed. Classically this corresponds to the trajectories of an electron moving along the edge of the device in a magnetic field (skipping orbit). As a consequence, there exist extended states at the Fermi energy near the sample boundaries.

Soon after the discovery of the QHE, Halperin recognized [97] the importance of theses edge channels in the transport properties of the 2DEG. In combination with the Landauer formalism [98, 99] for transport, the edge state approach proved to be very efficient to understand electrical transport at high field. In the following, the approach adopted by Büttiker [100] will be very briefly summarized, although some pioneering work was done by Středa *et al.* [101] and by Jain *et al.* [102]. For additional information, excellent review papers have been published on the subject [103, 104].

In the Landauer formalism of transport, the current is taken as the driving force and the electric field can be obtained by calculating the charge distribution due to the current flow. Using transmission and reflection probabilities, the current is given as a function of the electrochemical potential at the contacts. For a single edge state k located between two electron reservoirs at electrochemical potential μ_1 and μ_2, the current fed by the contact in the absence of scattering is

$$I = ev_k D(E)(\mu_1 - \mu_2) = \frac{e}{h}\Delta\mu, \tag{17}$$

where v_k is the drift velocity of the electron which is proportional to the slope of the Landau level and therefore has an opposite sign on each side of the device. The density of states $D(E)$ is given by $D(E) = 2\pi\hbar v_k$ in a one-dimensional channel [100]. The voltage drop V between the reservoirs is $eV = \Delta\mu$ and the two-terminal resistance of the edge state is $R = h/e^2$. For N channels, one obtains

$$R = \frac{h}{e^2}\frac{1}{N}\,. \tag{18}$$

When elastic scattering takes place along the edge channels with a transmission probability T across the disordered region, the two-terminal resistance becomes

$$R = \frac{h}{e^2}\frac{1}{NT}\,. \tag{19}$$

The situation of a localized impurity scattering in an edge channel is schematically depicted in fig. 15. In a very intuitive way, the figure shows that the magnetic field suppresses backscattering of the electrons over a distance larger than the cyclotron orbit. This suppression of backscattering, which allows dissipationless current to flow along the edges, is the fundamental property responsible for the occurrence of the quantum Hall effect. As a consequence, $T = 1$ and the resistance is again given by eq. (18).

The power of the edge channel approach lies in the possibility of studying the role of the contacts which are fundamental in precision measurements as already noticed in 1992 by Büttiker [104].

The situation of a real Hall bar is depicted in fig. 16. The contacts are each characterized by transmission T_i and reflection R_i coefficients. As the contacts are separated by a

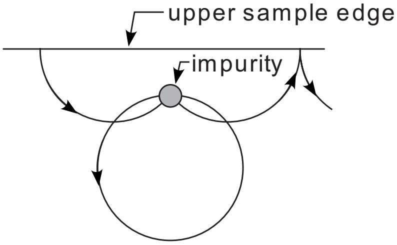


Fig. 15. – Quasiclassical skipping orbits along the upper edge of the sample in the presence of a localized impurity. In a high magnetic field, backscattering over distances large compared to the cyclotron radius is suppressed. (After [100].)

distance larger than the inelastic scattering length, $T_i = 1$ and $R_i = 0$ for all the contacts whether they are ideal or not. This stresses the importance of inelastic scattering, which equilibrates the edge states, in establishing exact quantization. The robustness of the quantum Hall effect with regard to the quality of the contacts stems from this particular property of the electron scattering in high field. In this case, the electrochemical potentials are related by $\mu_1 = \mu_3 = \mu_4$ and $\mu_2 = \mu_5 = \mu_6$. These conditions lead to $I = Ne(\mu_2 - \mu_1)$ and $V_{\mathrm{H}} = V_{56} = V_{34} = \Delta\mu$, yielding $R_{\mathrm{H}} = h/Ne^2$ and $R_{xx} = 0$, as expected.

This edge state model allows a realistic description of the electronic transport in high magnetic field as long as the difference in electrochemical potentials $\Delta\mu$ is small compared to the cyclotron energy $\hbar\omega_{\mathrm{c}}$. For high current densities, however, the current

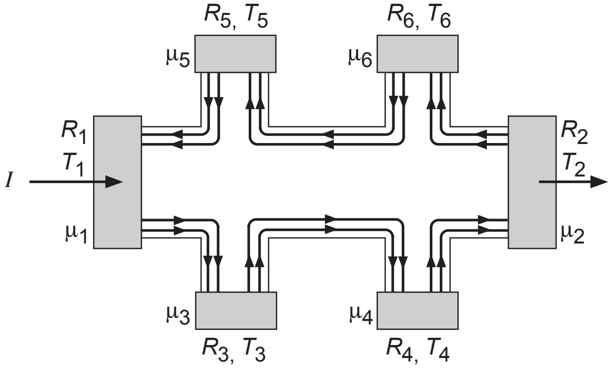


Fig. 16. – Schematic picture of a Hall bar with six ohmic contacts separated by a distance larger than the inelastic scattering length. The filling factor corresponds to the second Hall plateau. The contacts are characterized by reflection R_i and transmission T_i coefficients. Each contact is at an electrochemical potential μ_i. The d.c. transport current I flows between contact 1 (source) and 2 (drain). (After [100].)

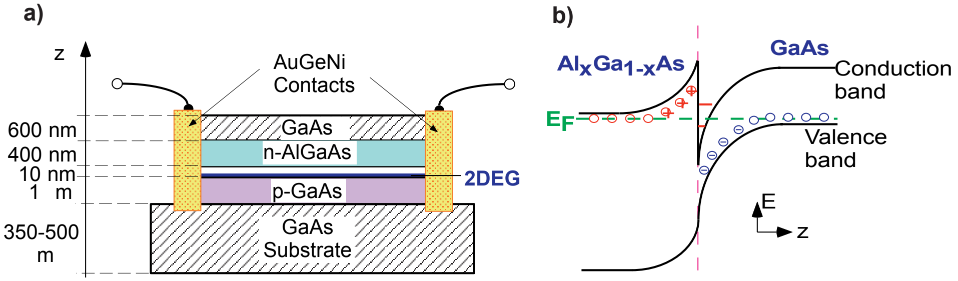


Fig. 17. – GaAs heterostructure. (a) Cross-section; (b) schematic energy diagram.

flows mainly in the bulk of the 2DEG and an extension of the edge state model is needed to explain the QHE in this regime.

The edge state picture has been successfully used to explain many experimental results which have been carefully reviewed in [105].

In the one-electron picture adopted above, the edge channels formed at the intersection of the Landau levels with the Fermi energy are like metallic wires running along the sample boundary and their spatial extension is comparable to the magnetic length (about 10 nm at 10 T). However, this description does not include the screening that takes place near the sample boundary. This screening at high magnetic fields forces the channels into compressible strips separated by incompressible regions [106]. In a quantitative study, Chklovskii *et al.* [107] calculated the width of the edge channels to be on the order of 1 μm on the second Hall plateau. This width is two orders of magnitude larger than in the one-electron picture and agrees fairly well with experimental measurements [108-110].

4˙3. *The two-dimensional electron gas.* – In real samples, the two-dimensional electron gas is located in the inversion layer found in various semiconductor devices. Inversion layers are formed at the interface between a semiconductor and an insulator (like the Si-MOSFET) or at the interface between two semiconductors, one of them acting as the insulator (like the AlGaAs/GaAs heterostructures). The quantum Hall effect was discovered in a Si-MOSFET. However, particularly in metrology, mostly AlGaAs/GaAs heterostructures are used.

In the AlGaAs system (see fig. 17), the GaAs is the semiconductor (energy gap $E_g = 1.5\,\text{eV}$) and the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ ($x \approx 0.3$), which has a wider gap ($E_g = 2.2\,\text{eV}$), plays the role of the insulator. Using the molecular beam epitaxy technique (MBE, see [111]), it is possible to fabricate interfaces with an atomic regularity given the close lattice match between the two materials. The $\text{Al}_x\text{Ga}_{1-x}\text{As}$ material is deliberately n doped to populate the bottom of its conduction band. From there electrons will migrate and populate the holes located at the top of the valence band of the GaAs (which is lightly p doped). Most of them, however, will fill the bottom of the conduction band of the GaAs. The positive charge left on the donors gives rise to an electric field which attracts the electrons towards the interface and, in a similar way to the Si-MOSFET, bends the valence and conduction bands. The transfer of electrons continues until the dipolar layer composed of the positive donors and the negative inversion layer is strong enough. This dipolar layer produces a discontinuity of the potential and finally aligns the Fermi energy of the two materials. The electronic density in the inversions layer is determined by the density of the dopant which is fixed for each sample, in contrast to the Si-MOSFET where it can be varied with the gate voltage.

A technique called modulation doping [112] consists in growing an additional layer of $\approx 10\,\mathrm{nm}$ of undoped $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ at the interface. The idea is to separate the charge carriers from the ionized impurities so that carriers can attain a mobility not affected by impurity scattering. At present, the mobility of the 2DEG can reach values as high as $200\,\mathrm{T}^{-1}$.

In precision measurements, the quality of the electrical contacts to the 2DEG is a critical issue. First, this quality must be such that the measurements are not affected by the contact resistance and second, the contacts have to be reliable because the quantum Hall resistors which are routinely used over periods of years must withstand numerous thermal cycles between room and cryogenic temperatures.

In the early days, contacts to GaAs devices were made by alloying In or Sn through the heterostructures. This method provided low contact resistance. However, its reliability was not suitable for metrology since the contacts deteriorate with time due to diffusion processes. Therefore, the technique used in optoelectronics devices to contact bulk GaAs was modified to take into account the presence of the AlGaAs layer and the modulation doping technique. The result is to sequentially evaporate an alloy of AuGeNi [113]. Contacts produced in this way regularly have a contact resistance below $R_\mathrm{c} = 100\,\mathrm{m}\Omega$. This contact resistance does not depend on the current (as long as the current stays below its breakdown value) and does not depend on the plateau index, as long as $\rho_{xx} \approx 0$. In addition, samples with such contacts have been very intensely used as resistance standards without showing any time deterioration over a period of 10 years.

4˙4. *Measurement techniques.* – In order to test the universality of the QHR or to use it for metrological applications, accurate resistance bridges have to be available. The best ratio accuracy and the lowest random uncertainty are attained with the cryogenic current comparator (CCC) proposed and first realized by Harvey in 1972 [114]. The principle of the method is shown in fig. 18(a). When a current-carrying wire is passed through a superconducting tube, a shielding current is induced on the surface of the tube such that a zero magnetic flux density is maintained in the interior of the superconductor (Meissner effect). The shielding current runs in the same direction as the initial current on the outside of the tube. The current density is uniform over the whole surface and thus independent of the geometrical position of the wire inside the tube. This principle is put in practice in a CCC as illustrated in fig. 18(b). In an arrangement introduced in [115], the superconducting tube is bent to a torus with overlapping ends like a snake swallowing its tail. The overlapping ends are electrically insulated, the length of the overlap has to be > 2 turns to keep the end effects on an acceptable level. Several windings, *e.g.* N_p and N_s with currents I_p and I_s, respectively, are placed inside the torus. The magnetic flux

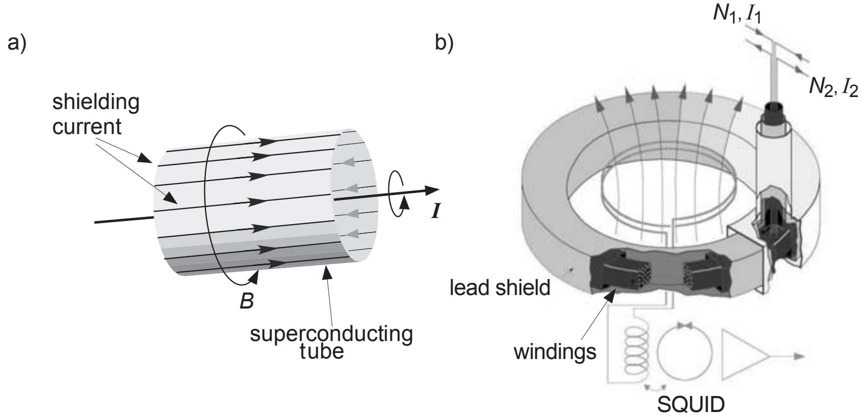


Fig. 18. – The cryogenic current comparator. (a) Illustration of the principle: a shielding current equal to I is induced on the external surface of the superconducting tube. (b) Set-up of the ratio coils: The windings of the current comparator form a toroidal coil which is enclosed in a superconducting shield. The shield overlaps itself like a snake swallowing its tail. The ampere-turns balance is sensed by measuring the magnetic flux in the pick-up coil using a SQUID.

created by the shielding current on the torus is proportional to $N_\mathrm{p}I_\mathrm{p} + N_\mathrm{s}I_\mathrm{s}$. This flux is sensed by a superconducting quantum interference device (SQUID) through a pick-up coil placed in the flux.

With a CCC, current ratios $I_\mathrm{s}/I_\mathrm{p} = N_\mathrm{p}/N_\mathrm{s}$ with a relative accuracy of 10^{-12} can be realized. An experimental check of the ratio accuracy is accomplished if the windings have a binary build-up. By measuring the SQUID signal of two windings with an equal number of turns put in an anti-series configuration, the error of the 1:1 ratio can be determined. In a binary build up, every winding with $2j$ turns can be compared directly with the combination $1 + \sum_{i=0}^{j-1} 2^i$, $j = 1, 2, 3\ldots$ of windings already tested.

The CCC bridge arrangement is schematically shown in fig. 19. A stabilized voltage source steers the primary and the secondary current sources. The ratio $N_\mathrm{p}/N_\mathrm{s}$ of the windings is set as close as possible to the nominal ratio of the two resistors $R_\mathrm{p}/R_\mathrm{s}$ to be measured. The output voltage of the SQUID system regulates the secondary current source in a closed feedback loop. The feedback assures that $N_\mathrm{p}I_\mathrm{p} = N_\mathrm{s}I_\mathrm{s}$. The detector, usually a battery-operated nanovoltmeter, indicates the difference between the resistance ratio and the winding ratio. The detector can be balanced with the help of an additional divider circuit composed of the trim coil N_t, a variable resistance R_l and a fixed high value resistor R_h. The ratio to be measured is then given by

$$(20) \qquad \frac{R_\mathrm{p}}{R_\mathrm{s}} = \frac{N_\mathrm{p}}{N_\mathrm{s}} \frac{1}{(1+d)} \frac{1}{\left(1+\frac{V_m}{V}\right)}\,; \quad d = \frac{N_\mathrm{t}}{N_\mathrm{s}} \frac{R_\mathrm{l}}{(R_\mathrm{l}+R_\mathrm{h})}\,,$$

where V_m is the detector reading ($\simeq 0$) and V the voltage drop across the resistors. The

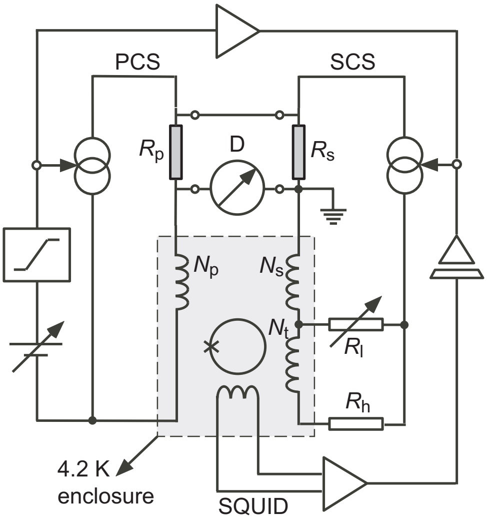


Fig. 19. – Schematic circuit diagram for a cryogenic current comparator bridge. The feedback signal from the SQUID accurately adjusts the current ratio between primary (left) and secondary (right) circuit to the ratio given by the windings N_p and N_s. The divider circuit composed of the trim coil, the adjustable resistor R_l and R_h is used to balance the detector D.

resolution of the bridge is mainly given by three factors: The SQUID noise, the thermal noise of the resistors and the detector noise.

The r.m.s. value of the current noise per turn in a d.c. SQUID-based CCC is around 10^{-10} A in a bandwidth of 0.01 Hz to 1 Hz ($1/f$ corner of the SQUID at 0.3 Hz) [116]. This is about a factor of ten above the optimum value. The current noise of the SQUID detection system, as given above, transforms through the resistance R_s to a voltage noise seen by the detector (typically $0.5\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$ for $R_\mathrm{s} = 100\,\Omega$ and $N_\mathrm{s} = 16$). The white thermal noise of a resistance R at a temperature T in a frequency bandwidth b is given by the Nyquist formula $V_{n\text{-th}} = \sqrt{4k_\mathrm{B}T\ R\ b}$, where k_B is the Boltzmann constant. The thermal noise is often the limiting factor for resistance measurements above $10\,\mathrm{k}\Omega$. The third important noise component originating form the detector itself is often the dominating part for resistances below $100\,\Omega$. Above this value, the best nanovoltmeters usually stay below the corresponding thermal noise of the source resistance at room temperature.

Typical parameters for a comparison of the QHR for $i = 2$ ($R_\mathrm{H}(2) = 12.9\,\mathrm{k}\Omega$) against a $100\,\Omega$ standard are: $N_\mathrm{p} = 2065$, $N_\mathrm{s} = 16$ and $I_\mathrm{p} = 50\,\mu\mathrm{A}$. For this configuration, the total r.m.s. voltage noise amounts to $7\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$. It is dominated by the detector noise because R_p is at 1 K. According to this figure, a type-A relative uncertainty of $1\,\mathrm{n}\Omega/\Omega$ is expected within a measurement time of 2 min. In reality, a slightly worse performance is achieved because of $1/f$ noise components (fluctuations of thermal voltages, detector and SQUID).

Today, the lowest uncertainties in resistance comparisons for $1\,\Omega \le R \le 100\,\mathrm{k}\Omega$ are obtained using CCC bridges (see, *e.g.*, [117-120]).

Cryogenic current comparators working at d.c. are limited to a great extent by very low frequency noise of the $1/f$ type (null detector, SQUID) and by thermal effects (slowly varying thermal voltages, Peltier effect). To overcome these problems, a.c. CCCs working at a frequency close to 1 Hz were developed [121,122]. As it turns out, the CCC coil has not to be specially designed to work at low frequency. However, quite some complexity is added to the bridge set-up to allow in-phase and quadrature current ratio matching and to avoid errors caused by stray capacitive effects.

4˙5. *Universality of the quantised Hall resistance.* – An incomplete quantization of a plateau due to high current through the device or due to increased temperature leads to a finite ρ_{xx}. A linear relationship exists between the deviation in the measured Hall resistance from the expected value and ρ_{xx}. Finite longitudinal voltages can also be measured as a result of non-ideal contacts. The question is whether the extrapolated value $R_{\mathrm{H}}(i, \rho_{xx} \to 0)$ is the same irrespective of the device geometry, material and fabrication process, the carrier mobility and density, the plateau number or other factors. Due to the absence of quantitative theoretical models, this question has essentially been approached experimentally.

Already in 1987 an experimental study [123] has shown that the QHRs observed in four different GaAs devices were in agreement at the level of 5×10^{-9}. The search for possible differences between the QHR realized in a GaAs heterostructure and a Si-MOSFET, respectively, was of special interest. In a direct comparison, Hartland *et al.* [124] found that the difference between the QHR in the two device types was smaller than 3.5 parts in 10^{10}. However, at about the same time, several other groups [125-127] reported anomalous values of the QHR measured in a particular Si-MOSFET device. The authors claimed to see differences in R_{H} up to several parts in 10^7 despite the absence of any measured dissipation within the experimental resolution. Subsequently, a theoretical model [128] was presented which explains such deviations by the presence of short-range elastic scatterers located at the edges.

A more recent experimental study [129] included devices from the same wafer as those for which the deviant data were obtained. In this case, an agreement between Si-MOSFET and GaAs was found at the level of the experimental uncertainty of 2.3 parts in 10^{10}. The study demonstrated that, due to edge effects, the longitudinal voltage measured along one side of a device can be quite different from the value on the other side. The measurement of zero dissipation at one device side only, therefore, does not guarantee zero longitudinal voltage values on both sides which is a prerequisite to measure a fully quantized value of R_{H}.

In the same work [129], it was also shown that the extrapolated Hall resistance value $R_{\mathrm{H}}(i = 2, 4, \rho_{xx} \to 0)$ does not depend on the device mobility $(13\,\mathrm{T}^{-1} \le \mu \le 135\,\mathrm{T}^{-1})$ and the fabrication process (MBE or MOCVD) within 3 parts in 10^{10}.

As for the plateau number i, the results confirm that in GaAs devices no dependence on this quantum number can be seen

$$\overline{\frac{i \cdot R_H\,(i)}{2 \cdot R_H\,(2)}} = 1 - (1.2 \pm 2.9) \cdot 10^{-10}, \quad i = 1, 3, 4, 6, 8. \tag{21}$$

Among the large number of theoretical papers on the QHE, a few address the question of size effects, including the width dependence of the QHR. Although based on different approaches [130-133], the majority of these models find that the relative variation of $R_{\mathrm{H}}(i)$ should scale like the inverse square of the device width w, more precisely

$$\frac{\Delta R_{\mathrm{H}}\,(i)}{R_{\mathrm{H}}\,(i)} = \alpha \left(\frac{l}{w}\right)^2 , \tag{22}$$

where $\Delta R_{\mathrm{H}}(i) = R_{\mathrm{H}}(i, w) - R_{\mathrm{H}}(i, w = \infty)$, l is the magnetic length and α is the parameter reflecting the strength of the size effect. In measurements carried out using GaAs Hall bars of widths varying from $10\,\mu$m to $1000\,\mu$m [134], no size effect was observed within the experimental uncertainty of 1 part in 10^9. The values for the parameter α are $(1.8 \pm 1.8)10^{-3}$ and $(0.7 \pm 5.0)10^{-3}$ for the $i = 2$ and $i = 4$ plateau, respectively.

These results demonstrate that possible size effects are totally negligible for the sample sizes presently used in metrology.

4˙5.1. Contact effects. Another important issue is the influence of non-ideal contacts to the 2DEG on the QHR. It is well known [135] that in the adiabatic regime (*i.e.*, small devices of high mobility, distance between contacts $\leq 100\,\mu$m) and for small enough currents (linear regime), large deviations from the QHR are caused by imperfect voltage contacts. The effects can be explained in the framework of the Landauer-Büttiker formalism (see subsect. 4˙2). However, in the metrological application of the QHE where much higher currents are passed through a device of macroscopic dimensions, the pure edge-state description is no longer appropriate and the consequence of bad contacts is less clear.

The quality of the contacts is characterized by their resistance R_{c} which is measured in the QHE regime as follows. The voltage drop across the contact j to be characterized and the next contact situated at the same Hall potential is measured while passing a current through contact j and one of the current contacts. If the sample is well quantized ρ_{xx} can be neglected and $R_{\mathrm{c}j}$ is obtained directly. The resistance of a good AuGeNi contact is usually well below $1\,\Omega$, provided the device is cooled slowly from room temperature down to the working temperature of $< 2\,$K.

The resistance of the contact region can also be varied in a controlled way by using a gate placed over the probe, *i.e.* the narrow arm which links the contact pad to the main channel of the device in the usual Hall bar geometry. Applying a voltage to the gate partially depletes the 2DEG under the gate. In metrological applications of the QHE when standard ungated Hall bar devices are used, a similar local reduction of the carrier concentration in the narrow voltage probes can be caused accidentally by cooling a device too fast, by passing a current above the critical current through the potential probe or even by leaving the device in the cold for several days. The original contact properties are restored by cycling the device through room temperature or by illuminating the device at low temperature with infrared radiation [136].

The influence of non-ideal voltage contacts on the QHR was extensively studied by Jeckelmann *et al.* [129,137]. It was shown that deviations $\Delta R_\mathrm{H}/R_\mathrm{H}$ of up to 1 part in 10^6 can occur as a consequence of R_c values in the kΩ range. At the same time, a corresponding positive or negative longitudinal voltage V_{xx} is measured along the side of the device where the bad contacts are connected to. A zero V_{xx} is measured on the opposite side if the corresponding contacts have a low resistance. There is no simple relation between ΔR_H and R_c. The data show, however, that the maximum deviation ΔR_H at a given R_c is proportional to $R_\mathrm{c}/R_\mathrm{H}$. The deviations become more pronounced when going to higher filling factors. In addition they are inversely proportional to the current and they decay exponentially with increasing temperature. The data finally demonstrate that deviations in the QHR above the experimental resolution of $0.5\,\mathrm{n}\Omega/\Omega$ are to be expected if the resistance of the potential contacts exceeds $100\,\Omega$ when measuring on the $i = 2$ plateau and $10\,\Omega$ in the case of $i = 4$. These limits apply for a temperature of 0.3 K and a current above $10\,\mu$A and it is assumed that the resistance of the current contacts is in the mΩ range.

A model based on the Büttiker formalism for contacts [138] allows an estimate to be made on the upper limit for the deviation of the four-terminal resistance as a function of the contact resistance. The magnitude of the effects is similar to the experimental findings of [137]. On the other hand, the model, as it only considers pure edge state transport in a uniform device, does not explain the detailed pattern of the observations.

4˙6. *The use of the QHR as a standard of resistance.* – Since January 1, 1990, most major National Metrology Institutes are using the QHE to realize a representation of the SI unit ohm on the basis of the conventional value $R_{\mathrm{K}\text{-}90}$. As shown in previous section, the value of R_K is independent of the experimental conditions as long as the QHE device is fully quantized. Temperature, current or contact effects may cause deviations from the correct value. Most important, however, test measurements can reveal whether the device is in a proper state or not. This means that the value of the QHR can be made as reproducible as today's measurement techniques allow without making reference to an external standard. These are the criteria a standard has to fulfil to be accepted as primary standard.

To guarantee the accuracy and reproducibility of the QHR standard, the QHE device, the measurement system and the procedures have to meet a number of strict requirements. A group of experts under the auspices of the Comité Consultatif d'Électricité has put together the "Technical Guidelines for Reliable Measurements of the quantized Hall Resistance" [139] which, when correctly applied in practice, assure correct QHR measurements.

The resistance bridges of the type briefly introduced in subsect. **4˙4** are used to calibrate traditional room temperature resistance standards in terms of the QHR. As an example, fig. 20 shows the measurements carried out at METAS to determine the drift

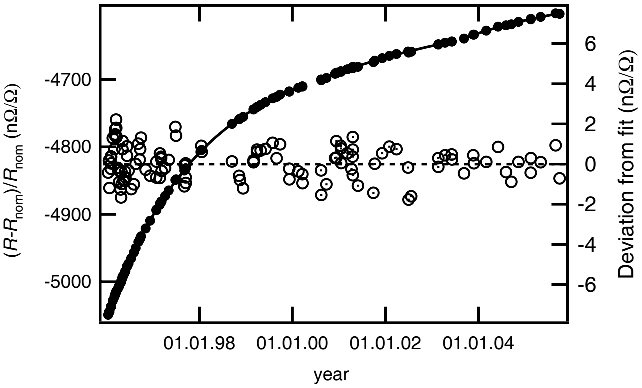


Fig. 20. – Tracking of a 100 Ω standard resistor measured in terms of $R_{\mathrm{K\text{-}90}}$. The open circles (right scale) indicate the deviations of the measured data points to the fit function. Data taken at the Swiss Federal Office of Metrology (METAS).

behaviour of a temperature stabilized wire-wound 100 Ω resistor. The standard is kept under constant ambient conditions. As the results show, its resistance can be described with high accuracy by a smooth fitting function, which makes it usable as a transfer standard at the level of 1 nΩ/Ω.

To check the worldwide consistency of the QHR measurements at the highest accuracy level, the BIPM has started in 1993 to perform on-site comparisons of resistance ratio measurements using a transportable QHE standard and resistance bridge. The results of the bilateral comparisons (see, *e.g.*, [140]) are made public by the BIPM in a database which is accessible by internet (`www.bipm.org`). The comparison results obtained so far are shown in fig. 21. The agreement between each laboratory and the BIPM for the $R_{\mathrm{H}}(2)/100\,\Omega$ is on the order of one part in 10^9 which is well within the combined standard uncertainty of the comparisons.

4·6.1. Quantum Hall array resistance standards. For the practical application of the QHE as a resistance standard, it is desirable to have quantized resistance values covering a wide range of quantum numbers. In reality, however, it turns out that the device characteristics are usually such that only the plateaus two and four are well quantized under normal operational conditions. The question thus arises whether a combination of several QHE devices in a series or parallel configuration may yield the practical resistance value needed in a specific application.

When several QHE devices are put in series or parallel in a network, the resistances of the contacts and the connecting wires have to be taken into account when the overall resistance of the network is determined. Ricketts and Kemeny [141] first described the electrical behaviour of a QHE device in terms of an equivalent circuit. Based on this model, Delahaye [142] has shown that the contact effects can be drastically reduced if the number of links between neighboring devices is increased. This is illustrated in fig. 22.

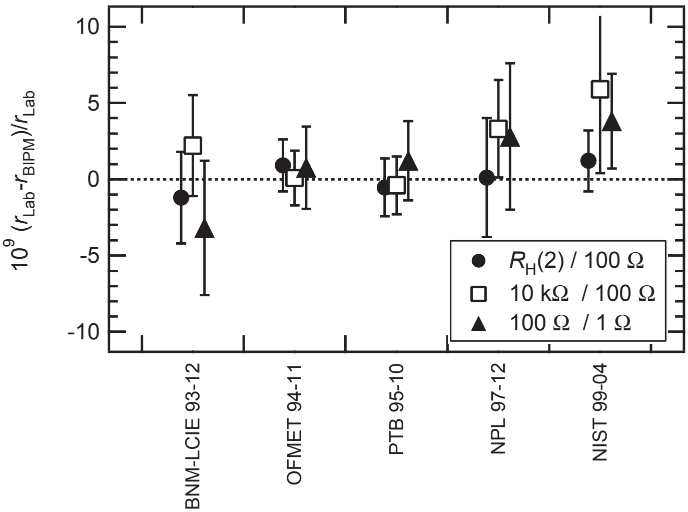


Fig. 21. – Results of on-site comparisons of three different resistance ratios using the Bureau International des Poids et Mesures (BIPM) transportable QHE system.

When two Hall bars are connected in series, the resistance $R_{\rm AB'}$ —where A and B′ denote the potential terminals— is given by $R_{\rm AB'} = 2R_{\rm H}(1+\delta_{\rm c})$, where $\delta_{\rm c} \approx R_{\rm c}/R_{\rm H}$ is the error introduced by the contact resistance $R_{\rm c}$. If in addition, the two potential terminals A′ and B are connected, the error term is reduced to $\delta_{\rm c} \approx (R_{\rm c}/R_{\rm H})^2$. The multiple series arrangement thus allows a reduction of the influence of link resistances to very small levels.

Poirier *et al.* [143] have developed quantum Hall resistance standards with nominal values in the range from $R_{\rm K}/200$ to $50R_{\rm K}$ ($i = 2$ plateau). Using triple series (respec-

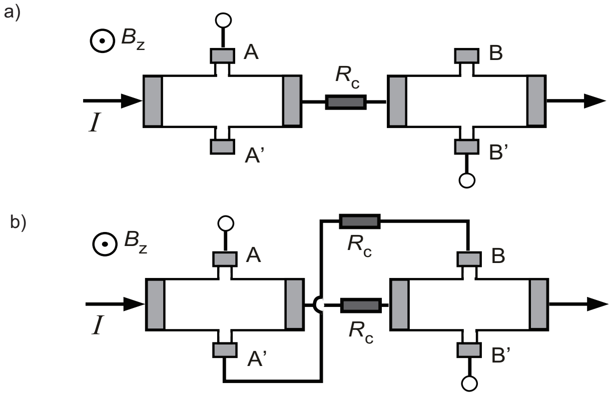


Fig. 22. – Illustration of the multiple series connection scheme.

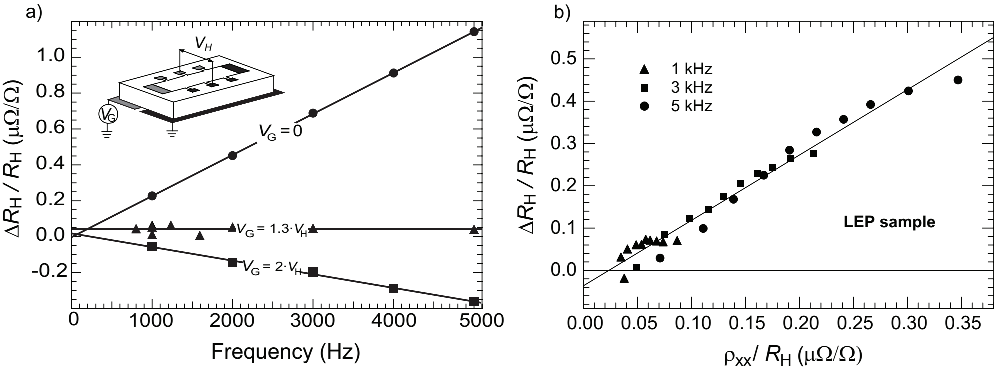


Fig. 23. – a) Relative deviation of the Hall resistance R_{H} measured at the centre of the $i = 2$ plateau as a function of the gate voltage applied to one part of a split backgate for a GaAs sample (after [153]). b) Relative deviation of the Hall resistance as a function of the longitudinal voltage. The current was varied between $10\,\mu$A to $100\,\mu$A for each frequency range (after [154]).

tively, parallel) connection schemes, the contact effects could be reduced to a level below 5 parts in 10^9.

4˙7. *A.c. measurements of the QHR.* – The success of the QHE as a d.c. standard of resistance has stimulated research into the characteristics of the QHR at alternating current. The aim is to use the QHR as an impedance standard up to frequencies in the kHz range [144]. If two QHR standards could, *e.g.*, directly be integrated into a quadrature bridge, the measurement chain linking resistance and capacitance would be simplified considerably. From the theoretical point of view, the picture emerging from the literature was not clear (see [90] for a review) and precision measurements of the QHR were needed to understand its behaviour under a.c. transport conditions. The measurements carried out at several metrology institutes [145-150] revealed a systematic dependence of the Hall resistance with current and frequency. These dependencies are related with losses either within the QHE device itself or losses due to current leaking to or from the sample's surroundings. These losses can be compensated by the use of external gates [151-153] as illustrated in fig. 23a).

The a.c. QHR can only be used as a primary standard, if it can be shown that a universal value of the Hall resistance $R_{\mathrm{H}}(f) = R_{\mathrm{K}}/i$ can be established without the need of additional external standards. In the d.c. case, the absence of a longitudinal voltage drop along both sides of the QHE device ($\rho_{xx} = 0$) is the necessary condition for the perfect quantization of the Hall resistance. In the a.c. case the following strategies were investigated.

Delahaye *et al.* [151] developed a phenomenological model explaining the principal a.c. losses in a QHE device and the resulting current and frequency coefficients of the QHR. They showed that the current coefficient evaluated at fixed frequency can be zeroed by a proper adjustment of the gate voltage applied to a split gate placed below the QHE device. Using this method, the QHR observed in several GaAs heterostructure devices, after proper adjustment of the gate, were compared to a calculable resistance standard. The measured frequency coefficients of the QHR did not exceed ±2 parts in 10^8 per kilohertz.

Another method consists in measuring the a.c. dissipation along the QHE device. Recently, it was shown [155] that the deviation from the perfect quantization of the QHR, ΔR_{H} is proportional to the a.c. dissipation ρ_{xx}

$$\Delta R_{\mathrm{H}} = s' \rho_{xx}. \tag{23}$$

A similar linear relationship was already observed for the temperature dependence in the case of d.c. measurements [156]. As shown by Jeanneret *et al.* [154], the relation can be explained in the a.c. case if the displacement current flowing in the system is properly taken into account. The split gate technique can be used to zero the longitudinal a.c. dissipation [157] and thus the frequency-dependent deviation form the correct value of the QHR. Alternatively, the correct value can be found by extrapolating the measured values to zero dissipation [155]. In this case, no back gate is used and the capacitive currents leaking to or from the surroundings of the QHE device are kept as small as possible. This approach has the advantage that the physical processes taking place in the QHE device itself can be separated from the effects caused by the experimental set-up.

Although major progress has been made in recent years, more systematic studies need to be performed to ensure the universal character of the a.c. QHR. In addition, a deeper understanding is required to describe the physical processes giving rise to a.c. dissipation in the 2DEG.

5. – Conclusions

The discoveries of the Josephson and the quantum Hall effects have started a revolution in electrical metrology. Since 1990, the two effects allow the representation of the electrical units volt and ohm based on natural constants. There is overwhelming experimental evidence that the Josephson voltage and the quantized Hall resistance are universal quantities. The electrical quantum effects are used worldwide as invariant references in electrical calibrations and have improved the reproducibility of calibration results by up to two orders of magnitude during the last two decades. Nevertheless, the definition of the conventional values $K_{\text{J-90}}$ and $R_{\text{K-90}}$ is unsatisfactory on the long term as the consistency of the SI system as a whole still depends on the difficult experiments which link mechanical and electrical units. In addition, the unit of mass, which is one of the base mechanical units, is the last remaining artefact in the SI. The kilogram is defined as the mass of the international prototype of the kilogram, made of platinum-iridium and kept at the BIPM under special conditions. One of the major disadvantages of this definition is the fact that the kilogram is subject to possible changes in time. As a consequence, the electrical units which all depend on the kilogram may also drift with time. Impressive efforts are presently undertaken to find a proper replacement of the present kilogram definition based on fundamental constants (see [158] for a review).

Recently [159] the redefinition of the kilogram, ampere, kelvin and mole linked to fixed values of fundamental constants has been proposed. In this new SI, the kilogram would be defined in terms of the Planck constant h and the ampere in terms of the elementary charge e. As a consequence, the Josephson constant and the von Klitzing constant would become exactly known, thereby allowing the Josephson and quantum Hall effects to be a direct realization of the electrical SI units with no uncertainty contribution from the uncertainty of K_{J} and/or R_{K}.

* * *

It is a real pleasure to thank R. Behr, S. Benz, C. Hamilton, D. Kautz and J. Kohlmann for providing their figures and for their critical reading of sect. **2**.

REFERENCES

[1] Josephson B. D., Possible new effects in superconductive tunnelling, *Phys. Lett.*, **1** (1962) 251.

[2] Shapiro S., Josephson currents in superconducting tunneling: the effect of microwaves and other observations, *Phys. Rev. Lett.*, **11** (1963) 80.

[3] vonKlitzing K., Dorda G. and Pepper M., New method for high-accuracy determination of the fine structure constant based on quantized Hall resistance, *Phys. Rev. Lett.*, **45** (1980) 494.

[4] Thompson A. M. and Lampard D. G., A new theorem in electrostatics and its application to calculable standards of capacitance, *Nature (London)*, **177** (1956) 888.

[5] Hartland A., Jones R., Kibble B. and Legg D., The relationship between the SI ohm, the ohm at NPL, and the quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **36** (1987) 208.

[6] Jeffery A., Elmquist R., Lee L., Shields J. and Dziuba R., NIST comparison of the quantized Hall resistance and the realization of the SI ohm through the calculable capacitor, *IEEE Trans. Instrum. Meas.*, **46** (1997) 264.

[7] Small G., Rickets B., Coogan P., Pritchard B. and Sovierzoski M., A new determination of the quantized Hall resistance in terms of the NML calculable cross capacitor, *Metrologia*, **34** (1997) 241.

[8] Mohr P. J. and Taylor B. N., CODATA recommended values of fundamental physical constants: 2002, *Rev. Mod. Phys.*, **77** (2005) 1.

[9] van Dyck R. S., Schwinberg P. B. and Dehmelt H. G., *The electron* (Kluwer Academic, Netherlands) 1991, pp. 239–293.

[10] Kinoshita T., The fine structure constant, *Rep. Prog. Phys.*, **59** (1996) 1459.

[11] Eichenberger A., Jeckelmann B. and Richard P., Tracing Planck's constant to the kilogram by electromechanical methods, *Metrologia*, **40** (2003) 356.

[12] Steiner R. L., Williams E., Newell D. B. and Liu R., Towards an electronic kilogram: an improved measurement of the Planck constant and electron mass, *Metrologia*, **42** (2005) 431.

[13] Taylor B. N. and Witt T. J., New international electric reference standards based on the Josephson and quantum Hall effects, *Metrologia*, **26** (1989) 47.

[14] Kibble B. P., Robinson I. A. and Belliss J. H., A realization of the SI watt by the NPL moving-coil balance, *Metrologia*, **27** (1990) 173.

[15] Hamilton C. A., Burroughs C. J. and Chieh K., Operation of NIST Josephson array voltage standard, *J. Res. Natl. Inst. Stand. Technol.*, **95** (1990) 219.

[16] Kautz R. L., *Design and operation of series array Jospehson voltage standards*, in *Metrology at the frontier of Physics and Technology*, edited by Crovini L. and Quinn T. J. (North-Holland, Amsterdam) 1992, pp. 259–296.

[17] Pöpel R., The Josephson effect and voltage standards, *Metrologia*, **29** (1992) 153.

[18] Niemeyer J., Josephson voltage standards, *Handbook of Appl. Supercond.*, *Applications*, Vol. **2** (IOP, Bristol) 1998, pp. 1813–1834.

[19] Hamilton C. A., Josephson voltage standards, *Rev. Sci. Instrum.*, **71** (2000) 3611.

[20] Behr R., Müller F. and Kohlmann J., *Josephson junction arrays for voltage standards*, in *Studies of Josephson Junction Arrays II: Studies of High Temperature Superconductors*, edited by Narlikar A. V., Vol. **40** (Nova Science Publ., Huntington) 2002, pp. 155–184.

[21] Kohlmann J., Behr R. and Funk T., Josephson voltage standard, *Meas. Sci. Technol.*, **14** (2003) 1216.

[22] Benz S. P. and Hamilton C., Application of the Josephson effect to voltage metrology, *Proc. IEEE*, **92** (2004) 1617.

[23] Stewart W. C., Current-voltage characteristic of Josephson junctions, *Appl. Phys. Lett.*, **12** (1968) 277.

[24] McCumber D. E., Effect of ac impedance on dc voltage current characteristic of superconductor weak link junctions, *J. Appl. Phys.*, **39** (1968) 3113.

[25] Kautz R. L., Noise, chaos and the Josephson voltage standard, *Rep. Prog. Phys.*, **59** (1996) 935.

[26] Tsai J. S., Jain A. K. and Lukens J. E., High-precision test of the universality of the Josephson voltage-frequency relation, *Phys. Rev. Lett.*, **51** (1983) 316.

[27] Niemeyer J., Grimm L., Hamilton C. A. and Steiner R. L., High precision measurement of a possible resistive slope of Josephson array voltage steps, *IEEE Electron Device Lett.*, **EDL7** (1986) 44.

[28] Jain A. K., Lukens J. E. and Tsai J. S., Test for relativistic gravitational effects on charged particles, *Phys. Rev. Lett.*, **58** (1987) 1165.

[29] Kautz R. L. and Lloyd F. L., Precision of series-array Josephson voltage standards, *Appl. Phys. Lett.*, **51** (1987) 2043.

[30] Krasnopolin I. Y., Behr R. and Niemeyer J., High precise comparison of Nb/Al/AlOx/Al/AlOx/Al/Nb Josephson junction arrays using a SQUID as a null detector, *Supercond. Sci. Technol.*, **15** (2002) 1034.

[31] Clarke J., Experimental comparison of the Josephson voltage-frequency relation in different superconductors, *Phys. Rev. Lett.*, **21** (1968) 1566.

[32] Endo T., Koyanagi M. and Nakamura A., High-accuracy Josephson potentiometer, *IEEE Trans. Instrum. Meas.*, **IM-32** (1983) 267.

[33] Levinsen M. T., Chiao R. Y., Feldman M. J. and Tucker B. A., An inverse ac Josephson effect voltage standard, *Appl. Phys. Lett.*, **31** (1977) 776.

[34] Niemeyer J., Hinken J. H. and Kautz R. L., Microwave-induced constant-voltage steps at one volt from a series array of Josephson junctions, *Appl. Phys. Lett.*, **45** (1984) 478.

[35] Hamilton C. A., Kautz R. L., Steiner R. L. and Lloyd F., A practical Josephson voltage standard at 1 V, *IEEE Electron Device Lett.*, **EDL-6** (1985) 623.

[36] Niemeyer J., Grimm L., Meier W., Hinken J. H. and Vollmer E., Stable Josephon reference voltages between 0.1 and 1.3 V for high precision voltage standards, *Appl. Phys. Lett.*, **47** (1985) 1222.

[37] Lloyd F., Hamilton C. A., Beall J., Go D., Ono R. H. and Harris R. E., A Josephson array voltage standard at 10 V, *IEEE Electron Device Lett.*, **EDL-8** (1987) 449.

[38] Reymann D., Witt T. J., Eklund G., Pajander H., Nilsson H., Behr R., Funk T. and Müller F., A three-way, on site comparison of the 10 V Josephson voltage standards of the PTB, the SP, and the BIPM, *IEEE Trans. Instrum. Meas.*, **48** (1999) 257.

[39] Hamilton C. A. and Tang Y. H., Evaluating the uncertainty of Josephson voltage standards, *Metrologia*, **36** (1999) 53.

[40] Witt T. J., Maintenance and dissemination of voltage standards by Zener-diode-based instruments, *IEE Proc.-Sci. Meas. Technol.*, **149** (2002) 305.

[41] Witt T. J., Low-frequency spectral analysis of dc nanovoltmeters and voltage reference standards, *IEEE Trans. Instrum. Meas.*, **46** (1997) 318.

[42] Witt T. J. and Reymann D., Using power spectra and Allan variances to characterize the noise of Zener-diode voltage standards, *IEE Proc.-Sci. Meas. Technol.*, **147** (2000) 177.

[43] Wang C. M. and Hamilton C. A., The fourth interlaboratory comparison of 10 V Josephson voltage standards in North America, *Metrologia*, **35** (1998) 33.

[44] Reymann D., Witt T. J., Vrabcek P., Tang Y., Hamilton C. A., Katkov A. S., Jeanneret B. and Power O., Recent developments in BIPM voltage standard comparisons, *IEEE Trans. Instrum. Meas.*, **50** (2001) 206.

[45] Giem J. I., Sub-ppm linearity testing of a DMM using a Josephson junction array, *IEEE Trans. Instrum. Meas.*, **40** (1991) 329.

[46] Hamilton C. A., Burroughs C. J. and Kautz R., Josephson D/A converter with fundamental accuracy, *IEEE Trans. Instrum. Meas.*, **44** (1995) 223.

[47] Benz S. P., Hamilton C. A., Burroughs C. J., Harvey T. E. and Christian L. A., Stable 1 Volt programmable voltage standard, *Appl. Phys. Lett.*, **71** (1997) 1866.

[48] Benz S. P., Superconductor-normal metal-superconductor junctions for programmable voltage standard, *Appl. Phys. Lett.*, **67** (1995) 2714.

[49] Benz S. P. and Burroughs C. J., Constant-voltage steps in arrays of Nb-PdAu-Nb Josephson junctions, *IEEE Trans. Appl. Supercond.*, **7** (1997) 2434.

[50] Jeanneret B., Rüfenacht A. and Burroughs C. J., High precision comparison between SNS and SIS Josephson voltage standards, *IEEE Trans. Instrum. Meas.*, **50** (2001) 188.

[51] Lo-Hive J., Djordjevic S., Cancela P., Piquemal F., Behr R., Burroughs C. and Seppä H., Characterization of binary Josephson series arrays of different types at BNM-LNE and comparisons with conventional SIS arrays, *IEEE Trans. Instrum. Meas.*, **52** (2003) 516.

[52] Chong Y., Dresselhaus P. and Benz S., Electrical properties of Nb-MoSi2-Nb Josephson junctions, *Appl. Phys. Lett.*, **86** (2005) 2505.

[53] Chong Y., Burroughs C., Dresselhaus P., Hadacek N., Yamamori H. and Benz S., Practical high-resolution programmable Josephson voltage standard using double- and triple-stacked MoSi2-barrier junctions, *IEEE Trans. Appl. Supercond.*, **15** (2005) 461.

[54] Chong Y., Burroughs C., Dresselhaus P., Hadacek N., Yamamori H. and Benz S., 2.6 V high-resolution programmable Josephon voltage standard circuit using double-stacked MoSi2-barrier junctions, *IEEE Trans. Instrum. Meas.*, **54** (2005) 616.

[55] Chong Y., Burroughs C., Dresselhaus P., Hadacek N., Yamamori H. and Benz S., Practical high resolution programmable Josephson voltage standards using double- and triple-stacked MoSi2-barrier junctions, *IEEE Trans. Appl. Supercond.*, **15** (2005) 461.

[56] Urano C., Murayama Y., Iwasa A., Shoji A., Yamamori H. and Ishizaki M., A precise evaluation of NbN-based 1 V programmable voltage standard arrays, *IEEE Trans. Instrum. Meas.*, **54** (2005) 645.

[57] Ishizaki M., Yamamori H., Shoji A., Dresselhaus P. and Benz S., Programmable Josephson voltage standard circuits using arrays of NbN/TiN/NbN/TiN/NbN double-junction stacks operated at 10 K, *IEEE Trans. Instrum. Meas.*, **54** (2005) 620.

[58] Yamamori H., Ishizaki M., Shoji A., Dresselhaus P. and Benz S., 10 V programmable Josephson voltage standard circuits using NbN/TiNx/NbN/TiNx/NbN double-junction stacks, *Appl. Phys. Lett.*, **88** (2006) 2503.

[59] Schulze H., Behr R., Müller F. and Niemeyer J., Nb/Al/AlOx/Al/Nb Josephson junctions for programmable voltage standard, *Appl. Phys. Lett.*, **73** (1998) 996.

[60] Müller F., Schulze H., Behr R., Kohlmann J. and Niemeyer J., The Nb-Al technology at PTB -a common base for different types of Josephson voltage standards, *Physica C*, **354** (2001) 66.

[61] Behr R., Schulze H., Müller F., Kohlmann J. and Niemeyer J., Josephson arrays at 70 GHz for conventional and programmable voltage standards, *IEEE Trans. Instrum. Meas.*, **48** (1999) 270.

[62] Schulze H., Müller F., Behr R., Kohlmann J., Niemeyer J. and Balashov D., SINIS Josephson junctions for programmable Josephson voltage standard circuits, *IEEE Trans. Appl. Supercond.*, **9** (1999) 4241.

[63] Kieler O., Behr R., Müller F., Schulze H., Kohlmann J. and Niemeyer J., Improved 1 V programmable Josephson voltage standard using SINIS junctions, *Physica C*, **372-376** (2002) 309.

[64] Behr R. *et al.*, Analysis of different measurement setups for a programmable Josephson voltage standard, *IEEE Trans. Instrum. Meas.*, **52** (2003) 524.

[65] Schulze H., Behr R., Kohlmann J. and Niemeyer J., Design and fabrication of 10 V SINIS Josephson array for programmable voltage standards, *Supercond. Sci. Technol.*, **13** (2000) 1293.

[66] Kohlmann J., Schulze H., Behr R., Müller F. and Niemeyer J., 10 V SINIS Josephson junction series arrays for programmable voltage standards, *IEEE Trans. Instrum. Meas.*, **50** (2001) 192.

[67] Hassel J., Seppä H., Grönberg L. and Suni I., SIS junctions with frequency dependent damping for a programmable Josephson voltage standard, *IEEE Trans. Instrum. Meas.*, **50** (2001) 195.

[68] Hassel J., Seppä H., Grönberg L. and Suni I., Optimization of a Josephson voltage array based on frequency dependently damped superconductor-insulator-superconductor junctions, *Rev. Sci. Instrum.*, **74** (2003) 3510.

[69] Burroughs C., Benz S., Hamilton C. A., Harvey T., Kinard J. R., Lipe T. E. and Sasaki H., Thermoelectric transfer difference of thermal converters measured with a Josephsn source, *IEEE Trans. Instrum. Meas.*, **48** (1999) 282.

[70] Funck T., Behr R. and Klonz M., Fast reversed dc measurements on thermal converters using a SINIS Josephson junction array, *IEEE Trans. Instrum. Meas.*, **50** (2001) 322.

[71] Behr R., Funck T., Schumacher B. and Warnecke P., Measuring resistance standards in terms of the quantized Hall resistance with a dual Josephson voltage standard using SINIS Josephson arrays, *IEEE Trans. Instrum. Meas.*, **52** (2003) 521.

[72] Behr R., Grimm L., Funck T., Kohlmann J., Schulze H., Müller F., Schumacher B., Warnecke P. and Niemeyer J., Application of series arrays to a dc quantum voltmeter, *IEEE Trans. Instrum. Meas.*, **50** (2001) 185.

[73] Beer W., Eichenberger A. L., Jeanneret B., Jeckelmann B., Pourzand A. R., Richard P. and Schwarz J. P., Status of the METAS Watt balance experiment, *IEEE Trans. Instrum. Meas.*, **52** (2003) 626.

[74] Hamilton C. A., Burroughs C. J., Benz S. P. and Kinard J. R., AC Josephson voltage standard: progress report, *IEEE Trans. Instrum. Meas.*, **46** (1997) 224.

[75] Behr R., Williams J., Patel P., Janssen T., Funck T. and Klonz M., Synthesis of precision waveforms using a SINIS Josephson junctions array, *IEEE Trans. Instrum. Meas.*, **54** (2005) 612.

[76] Ihlenfeld W. G. K., Mohns E., Behr R., Williams J., Patel P., Ramm G. and Bachmair H., Characterization of a high resolution analog-to-digital converter with a Josephson ac voltage source, *IEEE Trans. Instrum. Meas.*, **54** (2005) 649.

[77] Benz S. P. and Hamilton C. A., A pulse driven programmable Josephson voltage standard, *Appl. Phys. Lett.*, **68** (1996) 3171.

[78] Benz S. P., Hamilton C. A., Burroughs C. J., Harvey T., Christian L. A. and Przybysz J., Pulse driven Josephson digital/analog converter, *IEEE Trans. Appl. Supercond.*, **8** (1998) 42.

[79] Williams J., Janssen T., Palafox L., Humphreys D., Behr R., Kohlmann J. and Müller F., The simulation and measurement of the response of Josephson junctions to optoelectronically generated short pulses, *Supercond. Sci. Technol.*, **17** (2004) 815.

[80] Benz S. P., Hamilton C. A., Burroughs C. J. and Harvey T., Ac and dc bipolar voltage source using quantized pulses, *IEEE Trans. Instrum. Meas.*, **48** (1999) 266.

[81] Benz S. P., Burroughs C. J. and Hamilton C. A., Operation margins for a pulse-driven programmable voltage standard, *IEEE Trans. Appl. Supercond.*, **7** (1997) 2653.

[82] Benz S. P., Burroughs C. J. and Dresselhaus P. D., Low harmonic distorsion in a Josephson arbitrary waveform synthesizer, *Appl. Phys. Lett.*, **77** (2000) 1014.

[83] Benz S. P., Burroughs C. J., Dresselhaus P. D. and Christian L., Ac and dc voltages from a Josephson arbitrary waveform synthesizer, *IEEE Trans. Instrum. Meas.*, **50** (2001) 181.

[84] Benz S. P., Burroughs C. J. and Dresselhaus P. D., Ac coupling technique for Josephson waveform synthesis, *IEEE Trans. Appl. Supercond.*, **11** (2001) 612.

[85] Burroughs C. J., Benz S. P., Dresselhaus P. D. and Chong Y., Precision measurements of ac Josephson voltage standard operating margins, *IEEE Trans. Instrum. Meas.*, **54** (2005) 624.

[86] Benz S. P., Dresselhaus P. D. and Burroughs C. J., Nanotechnology for next generation Josephson voltage standard, *IEEE Trans. Instrum. Meas.*, **50** (2001) 1513.

[87] Nam S., Benz S. P., Dresselhaus P. D., Tew W., White D. and Martinis J., Josephson noise termometry measurements using a quantized voltage noise source for calibration, *IEEE Trans. Instrum. Meas.*, **52** (2003) 550.

[88] Benz S. P., Martinis J., Dresselhaus P. D. and AND S. N., An ac Josephson source for Johnson noise thermometry, *IEEE Trans. Instrum. Meas.*, **52** (2003) 545.

[89] Nam S., Benz S. P., Dresselhaus P. D., Burroughs C. J., Tew W., White D. and Martinis J., Progress on Johnson noise thermometry using a quantum voltage noise source for calibration, *IEEE Trans. Instrum. Meas.*, **54** (2005) 653.

[90] Jeckelmann B. and Jeanneret B., The quantum Hall effect as an electrical resistance standard, *Rep. Prog. Phys.*, **64** (2001) 1603.

[91] Hall E. H., On a new action of the magnet on electric currents, *Am. J. Math.*, **2** (1879) 287.

[92] Prange R., Quantized Hall resistance and the measurement of the fine structure constant, *Phys. Rev. B*, **23** (1981) 4802.

[93] Aoki H., Quantised Hall effect, *Rep. Prog. Phys.*, **50** (1987) 655.

[94] Středa P., Theory of quantized Hall conductivity in two dimensions, *J. Phys. C*, **15** (1982) L717.

[95] Janssen M., Viehweger O., Fastenrath U. and Hajdu J., *Introduction to the theory of the integer quantum Hall effect* (VCH Verlagsgesellschaft, Weinheim) 1994.

[96] Laughlin R., Quantized Hall conductivity in two dimensions, *Phys. Rev. B*, **23** (1981) 5632.

[97] Halperin B., Quantized Hall conductance, current-carrying edge states, and the existence of extended states in a two-dimenisonal disordered potential, *Phys. Rev. B*, **25** (1982) 2185.

[98] Landauer R., Spatial variation of currents and fields due to localized scatterers in metallic conduction, *IBM J. Res. Dev.*, **1** (1957) 223.

[99] Landauer R., Electrical resistance of disordered one-dimensional lattices, *Philos. Mag.*, **21** (1970) 863.

[100] Büttiker M., Absence of backscattering in the quantum Hall effect in multiprobe conductors, *Phys. Rev. B*, **38** (1988) 9375.

[101] Středa P., Kucera J. and MacDonald A., Edge states, transmission matrices and the Hall resistances, *Phys. Rev. Lett.*, **59** (1987) 1973.

[102] Jain J. and Kivelson S., Quantum Hall effect in quasi one-dimensional systems: resistance fluctuations and breakdown, *Phys. Rev. Lett.*, **60** (1988) 1542.

[103] Beenakker C. and van Houten H., *Quantum transport in semiconductor nanostructures*, in *Solid State Physics*, edited by Ehrenreich H. and Turnbull D., Vol. **44** (Academic Press, New York) 1991, pp. 1–228.

[104] Büttiker M., *The quantum Hall effect in open conductors*, in *Semiconductors and Semimetals*, Vol. **35** (Academic Press, San Diego) 1992, pp. 191–277.

[105] Haug R., Edge state transport and its experimental consequences in high magnetic fields, *Semicond. Sci. Technol.*, **8** (1993) 131.

[106] Beenakker C., Edge channels for the fractional qantum Hall effect, *Phys. Rev. Lett.*, **64** (1990) 216.

[107] Chklovskii D., Shklovskii B. and Glazman L., Electrostatics of edge channels, *Phys. Rev. B*, **46** (1992) 4026.

[108] Hwang S., Tsui D. and Shayegan M., Experimental evidence for finite-width edge channels in the integer and fractional quantum Hall effect, *Phys. Rev. B*, **48** (1993) 8161.

[109] Zhitenev N., Haug R., von Klitzing K. and Eberl K., Experimental determination of the dispersion of edge magnetoplasmons confined in edge channels, *Phys. Rev. B*, **49** (1994) 7809.

[110] Takaoka S., Oto K., Kurimoto H., Murase K., Gamo K. and Nishi S., Magnetocapacitance and the edge state of a two-dimensional electron system in the quantum Hall regime, *Phys. Rev. Lett.*, **72** (1994) 3080.

[111] Vossen J. L. and Kern W. (Editors), *Thin films processes II* (Academic Press, San Diego) 1991.

[112] Dingle R., Störmer H. L., Gossard A. C. and Wiegmann W., Electron mobilities in modulation-doped semiconductor heterojunction superlattices, *Appl. Phys. Lett.*, **33** (1978) 665.

[113] Jucknischke D., Bühlmann H.-J., Houdré R., Ilegems M., Py M. A., Jeckelmann B. and Schwitz W., Properties of alloyed AuGeNi-contacts on GaAs/AlGaAs heterostructures, *IEEE Trans. Instrum. Meas.*, **40** (1991) 228.

[114] Harvey I., Precise low temperature dc ratio transformer, *Rev. Sci. Instrum.*, **43** (1972) 1626.

[115] Sullivan D. B. and Dziuba R. F., Low temperature direct current comparators, *Rev. Sci. Instrum.*, **45** (1974) 517.

[116] Sesé J., Camon A., Rillo C. and Rietveld G., Ultimate current resolution of a cryogenic current comparator, *IEEE Trans. Instrum. Meas.*, **48** (1999) 1306.

[117] Williams J. M. and Hartland A., An automated cryogenic current comparator resistance ratio bridge, *IEEE Trans. Instrum. Meas.*, **40** (1991) 267.

[118] Delahaye F. and Bournaud D., Low-noise measurements of the quantized Hall resistance using an improved cryogenic current comparator bridge, *IEEE Trans. Instrum. Meas.*, **40** (1991) 237.

[119] Dziuba R. F. and Elmquist R. E., Improvements in resistance scaling at NIST using cryogenic current comparators, *IEEE Trans. Instrum. Meas.*, **42** (1993) 126.

[120] Jeckelmann B., Fasel W. and Jeanneret B., Improvements in the realisation of the quantized Hall resistance standard at OFMET, *IEEE Trans. Instrum. Meas.*, **44** (1995) 265.

[121] Delahaye F., An AC-bridge for low-frequency measurements of the quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **40** (1991) 883.

[122] Seppä H. and Satrapinski A., Ac resistance bridge based on the cryogenic current comparator, *IEEE Trans. Instrum. Meas.*, **46** (1997) 463.

[123] Delahaye F. and Dominguez D., Precison comparison of quantized Hall resistances, *IEEE Trans. Instrum. Meas.*, **36** (1987) 226.

[124] Hartland A., Jones K., Williams J., Gallagher B. and Galloway T., Direct comparison of the quantized Hall resistance in gallium arsenide and silicon, *Phys. Rev. Lett.*, **66** (1991) 969.

[125] Kawaji S., Nagashima N., Kikuchi N., Wakabayashi J., Ricketts B. W., Yoshihiro K., Kinoshita J., Inagaki K. and Yamanouchi C., Quantized Hall resistance measurements, *IEEE Trans. Instrum. Meas.*, **38** (1989) 270.

[126] vanDegrift C., Yoshihiro K., Cage M., Yu D., Segawa K., Kinoshita J. and Endo T., Anomalously offset quantized Hall plateaus in high-mobility si-MOSFETs, *Surf. Sci.*, **263** (1992) 116.

[127] Yoshihiro K., van Degrift C., Cage M. and Yu D., Anomalous behavior of a quantized Hall plateau in a high-mobility si metal-oxide-semiconductor field-effect transistor, *Phys. Rev. B*, **45** (1992) 14204.

[128] Heinonen O. and Johnson M., Failure of the integer quantum Hall effect without dissipation, *Phys. Rev. B*, **49** (1994) 11230.

[129] Jeckelmann B., Jeanneret B. and Inglis D., High precision measurements of the quantized Hall resistance: Experimental conditions for universality, *Phys. Rev. B*, **55** (1997) 13124.

[130] MacDonald A. and Streda P., Quantized Hall effect and edge currents, *Phys. Rev. B*, **29** (1984) 1616.

[131] Shapiro B., Finite-size corrections in the quantum Hall effect, *J. Phys. C*, **19** (1986) 4709.

[132] Brenig W. and Wysokinski W., Scattering approach to the von Klitzing effect, *Z. Phys. B*, **63** (1986) 149.

[133] Johnston R. and Schweitzer L., An alternative model for the integral quantum Hall effect, *Z. Phys. B*, **72** (1988) 217.

[134] Jeanneret B., Jeckelmann B., Bühlmann H., Houdré R. and Ilegems M., Influence of the device width on the accuracy of quantization in the integer quantum Hall effect, *IEEE Trans. Instrum. Meas.*, **44** (1995) 254.

[135] Komiyama S., Hirai H., Sasa S. and Fujii T., Non-equilibrium population of edge states and a role of contacts in the quantum Hall regime, *Surf. Sci.*, **229** (1990) 224.

[136] Jeanneret B., Jeckelmann B., Bühlmann H. and Ilegems M., Influence of infrared illumination on the accuracy of the quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **46** (1997) 285.

[137] Jeckelmann B. and Jeanneret B., Influence of the voltage contacts on the four-terminal quantized Hall resistance in the nonlinear regime, *IEEE Trans. Instrum. Meas.*, **46** (1997) 276.

[138] Hirai H. and Komiyama S., A contact limited precision of the quantized Hall resistance, *J. Appl. Phys.*, **68** (1990) 655.

[139] Delahaye F. and Jeckelmann B., Revised technical guidelines for reliable dc measurements of the quantized Hall resistance, *Metrologia*, **40** (2003) 217.

[140] Delahaye F., Witt T., Jeckelmann B. and Jeanneret B., Comparison of quantum Hall effect resistance standards of the OFMET and the BIPM, *Metrologia*, **32** (1996) 385.

[141] Ricketts B. and Kemeny P., Quantum Hall effect devices as circuits elements, *J. Phys. D: Appl. Phys.*, **21** (1988) 483.

[142] Delahaye F., Series and parallel connection of multiple quantum Hall-effect devices, *J. Appl. Phys.*, **73** (1993) 7914.

[143] Poirier W., Bounouh A., Piquemal F. and André J., A new generation of QHARS: discussion about the technical criteria for quantization, *Metrologia*, **41** (2004) 285.

[144] Melcher J., Warnecke P. and Hanke R., Comparison of precision ac and dc measurements of the quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **42** (1993) 292.

[145] Delahaye F., Accurate AC measurements of the quantized Hall resistance from 1 Hz to 1.6 kHz, *Metrologia*, **31** (1994/95) 367.

[146] Hartland A., Kibble B. P., Rodgers P. J. and Bohàček J., Ac measurements of the quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **44** (1995) 245.

[147] Piquemal F., Trapon G. and Genevès G., Ac measurements of the minimum longitudinal resistance of a QHE sample from 10 Hz to 10 kHz, *IEEE Trans. Instrum. Meas.*, **45** (1996) 918.

[148] Wood B., Inglis D., Côté M. and Young R., Improved AC quantized Hall measurements, *IEEE Trans. Instrum. Meas.*, **48** (1999) 305.

[149] Cabiati F., Callegaro L., Cassiago C., D'Elia V. and Reedtz G., Measurements of the AC longitudinal resistance of a GaAs-AlGaAs quantum Hall device, *IEEE Trans. Instrum. Meas.*, **48** (1999) 314.

[150] Chua S., Hartland A. and Kibble B., Measurement of the AC quantized Hall resistance, *IEEE Trans. Instrum. Meas.*, **48** (1999) 309.

[151] Delahaye F., Kibble B. and Zarka A., Controlling ac losses in quantum Hall effect devices, *Metrologia*, **37** (2000) 659.

[152] Schurr J., Melcher J., von Campenhausen A. and Pierz K., Adjusting the losses in an ac quantum Hall sample, *Metrologia*, **39** (2002) 13.

[153] Overney F., Jeanneret B. and Jeckelmann B., Effect of metallic gates on AC measurements of the quantum Hall resistance, *IEEE Trans. Instrum. Meas.*, **52** (2003) 574.

[154] Jeanneret B. and Overney F., Phenomenological model for frequency related dissipation in the quantized Hall resistance, to appear in *IEEE Trans. Instrum. Meas.*, **56** (2007).

[155] Overney F., Jeanneret B., Jeckelmann B., Wood B. M. and Schurr J., The quantized Hall resistance: towards a primary standard of impedance, *Metrologia*, **43** (2006) 409.

[156] Cage M., Field B., Dziuba R., Girvin S., Gossard A. and Tsui D., Temperature dependence of the quantum Hall resistance, *Phys. Rev. B*, **30** (1984) 2286.

[157] Schurr J., Ahlers F.-J., Hein G. and Pierz K., The ac quantum Hall effect as primary standard of impedance, *Metrologia*, **44** (2007) 15.

[158] Schwitz W., Jeckelmann B. and Richard P., Towards a new kilogram definition based on a fundamental constant, *C. R. Physique*, **5** (2004) 881.

[159] Mills I. M., Mohr P. J., Quinn T. J., Taylor B. N. and Williams E. R., Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM recommendation 1 CI-2005, *Metrologia*, **43** (2006) 227.

This page intentionally left blank

Single charge transport standards and quantum-metrological triangle experiments

F. Piquemal, L. Devoille, N. Feltin and B. Steck

Laboratoire National de Métrologie et d'Essais LNE - Trappes F 78197, France

1. – Introduction

In the present Système International d'Unités SI, the link between electrical and mechanical units is made through a realisation of ampere [1]. However, the direct determination of ampere cannot be carried out with a sufficient accuracy. In practice, it is more relevant to realise first the derivative electrical units, on the one hand farad and ohm, on the other hand volt (fig. 1). That allows the determination of ampere afterwards with a better uncertainty. The farad occupies a special place in the realisation of electrical units by means of a Thompson-Lampard calculable capacitor [2]. The setting-up of this calculable capacitor makes the SI realisation of the ohm possible through a comparison between impedances of capacitor and resistor [3]. That leads to a determination of the von Klitzing constant R_K originated from the quantum Hall effect (QHE) [4-6]. This effect links a resistance to a fundamental constant as the ac Josephson effect (JE) [6, 7] links electromotive force to another fundamental constant, the Josephson constant K_J. Furthermore the theory predicts that $R_\mathrm{K} = h/e^2$ and $K_\mathrm{J} = 2e/h$.

These two quantum phenomena have a great impact in metrology because firstly they provide fundamental standards with reproducible values independent of space and time, getting unique the representation of the ohm and the volt. Secondly, through SI realisation of electrical units, QHE and JE contribute significantly in the improvement of the knowledge of constants of nature [8]. For instance, the SI realisations of the ohm and the watt balance experiments [9, 10] lead to determine the well-known fine-structure constant $\alpha = \mu_0 c/(2h/e^2)$ and the Planck constant if one assumes that QHE and JE give h/e^2 and $2e/h$ exactly.

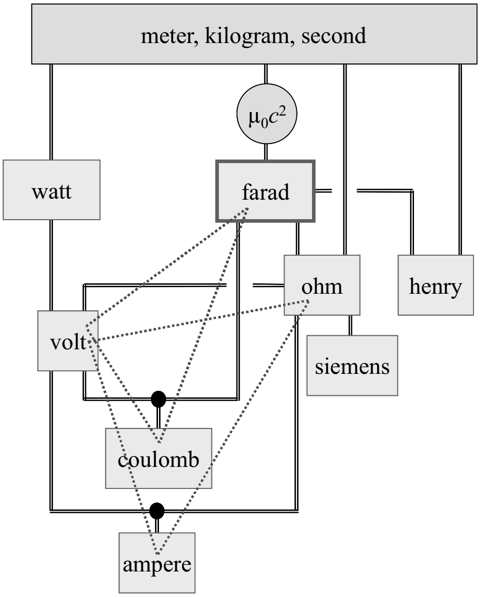


Fig. 1. – Chain of SI realisations of electrical units and metrological triangles. According to the present definition of ampere, the value of the permeability of vacuum μ_0 is fixed: $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2}$. The value of the speed of light in vacuum being fixed for the definition of the meter, leads to conventionally exact values of the permittivity of vacuum $\varepsilon_0 = 1/\mu_0 c^2$ ($\approx 113\,\mathrm{pF/m}$) and the free-space impedance $Z_0 = (\mu_0/\varepsilon_0)^{1/2}$ ($\approx 377\,\Omega$).

The paper deals with a third quantum phenomenon, the Single Electron Tunnelling (SET), and its main applications that could disrupt again the electrical metrology. This phenomenon indeed makes the development of quantum standard of current possible whose amplitude is directly linked to the elementary charge. The Quantum Metrological Triangle (QMT) experiment originally suggested by Likharev and Zorin [11] enables to test directly the coherence of the constants involved in QHE, JE and SET phenomena which are strongly presumed to provide the free-space values of h/e^2, $2e/h$ and e. This experiment consists either in applying Ohm's law $U = RI$ or in following $Q = CU$ [12] from the realisation of an electron counting capacitance standard [13]. Moreover combining QMT with watt balance and Thompson-Lampard calculable capacitor will lead to a determination of the elementary charge. These issues point out the important role that SET experiments should play toward the foundation of new SI system fully based on fundamental constants, for example by fixing h and e for a redefinition of the kilogram and the ampere[1].

[1] Other competitive proposals are to fix Avogadro number N_A instead of h for a new definition of the unit of mass and to keep μ_0 fixed for a reformulation of the ampere and the electrical units putting forward the free-space impedance Z_0.

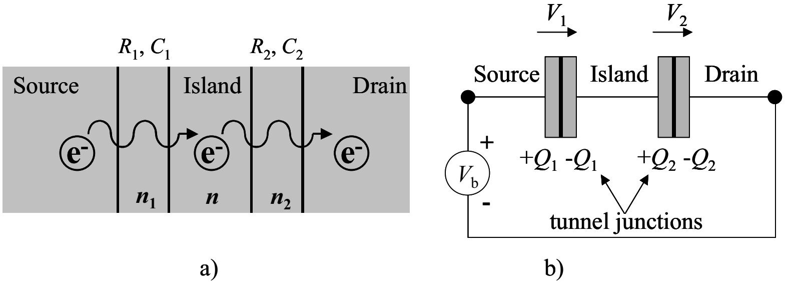


Fig. 2. – a) Schematic representation of a double tunnel junction. R_j is the tunnel resistance and C_j the junction capacitance. b) Symbolic representation of the circuit.

Section **2** describes basic theoretical elements on single electron tunnelling devices from double tunnel junctions to electron pump. For a detailed description on the theory, the reader is referred to review articles [14-16]. Section **3** deals with other single charge transport devices. Section **4** mainly covers the metrological triangle experiments and their impacts for fundamental constants. Conclusions and prospects are given in sect. **5**.

2. – Single electron devices

2˙1. *The elementary device based on Coulomb blockade.* – The Coulomb blockade of electron tunnelling, observed for the first time in disorder granular materials [17], appears when a part of a circuit, named "island", is electrically isolated from the rest of the circuit due to two tunnel junctions. On fig. 2, n_1 (respectively, n_2) is defined as the number of electrons which can be transferred through the first (respectively, the second) junction and n is the number of excess electrons on the island: $n = n_1 - n_2$. The charges Q_1 and Q_2 on the electrodes of capacitances C_1 and C_2 are continuous variables. n denotes the excess charge on the metallic island and changes only with a tunnel event inducing the entrance or the exit of an electron of the island. It leads to a quantization of the island charge and this feature is the cause of the single-electron effect in these systems.

Let us consider the circuit on fig. 2 b). A bias voltage source V_b is added to the double junction described on fig. 2 a). When this circuit is not voltage biased ($V_\mathrm{b} = 0$), there is no excess electron on the island. The island is strictly neutral in charge. On contrary, an applied voltage ($V_\mathrm{b} \neq 0$) will lead to a tunnel transfer through one of the junctions and to the presence of excess charges on the island. This charge variation is necessarily discrete as demonstrated below.

The electron transfer through a double tunnel junction can be treated from thermodynamics. The variation of Helmholtz free energy is defined as the difference between the electrostatic energy stored in the device described on fig. 1 and the work supplied by the voltage source V_b. This energy variation is

$$\Delta F = \Delta E_\mathrm{C} - \Delta W. \tag{1}$$

From a phenomenological point of view, the system will tend to minimize the free energy by means of tunnel transfers. The charge on the island can be written as

$$Q = Q_2 - Q_1 = -ne. \tag{2}$$

By assuming for each junction that, on the one hand, the system has time to energetically relax between two tunnel events, and on the other hand, the charge transfers are fast enough, the variation of free energy of each junction can be calculated. The basic idea is to express the voltages V_1 and V_2 across each junction, in order to calculate the electrostatic energy and the work of the sources. The voltage at each junction terminal is

$$V_1 = (-C_2V_\mathrm{b} + ne)/C_\Sigma, \tag{3}$$
$$V_2 = (-C_1V_\mathrm{b} - ne)/C_\Sigma \tag{4}$$

with $C_\Sigma = C_1 + C_2$. Moreover, the electrostatic energy stored by both junctions is

$$E_\mathrm{C} = {Q_1}^2/2C_1 + {Q_2}^2/2C_2 = \left[{V_\mathrm{b}}^2C_1C_2 + (ne)^2\right]/2C_\Sigma. \tag{5}$$

To get the free energy, the work of the source has to be calculated. If one electron crosses the first junction, the charges on the island and on the right (${Q_1}^-$) and left (${Q_1}^+$) electrodes of the first junction will be changed. It will lead to an electrostatic unbalance and the source will have to oppose to the voltage change due to the tunneling event. During this charge transfer the voltage V_1 varies by a quantity $-e/C_\Sigma$ corresponding to a charge $-eC_1/C_\Sigma$. But, in order to reach the electrostatic balance, the voltage source V_b has to bring the total polarization charge $-eC_2/C_\Sigma$. The work inherent to the transfer of n_1 electrons through the first junction and then of n_2 electrons through the second one becomes

$$W_1 = -n_1eV_\mathrm{b}C_2/C_\Sigma, \tag{6}$$
$$W_2 = -n_2eV_\mathrm{b}C_1/C_\Sigma. \tag{7}$$

Consequently, the tunnel event of one electron in one of the both directions through the first or the second junction leads to the free-energy variation $\Delta F = \Delta E_\mathrm{C} - \Delta W$

$$\Delta {F_1}^\pm = F(n_1 \pm 1, n_2) - F(n_1, n_2) = (e/C_\Sigma)[e/2 \pm (V_\mathrm{b}C_2 + ne)] = e^2/2C_\Sigma \pm eV_1, \tag{8}$$
$$\Delta {F_2}^\pm = F(n_1, n_2 \pm 1) - F(n_1, n_2) = (e/C_\Sigma)[e/2 \pm (V_\mathrm{b}C_1 - ne)] = e^2/2C_\Sigma \pm eV_2. \tag{9}$$

To make a charge transfer possible through the double junction, a negative free-energy variation is needed. From previous expressions, the condition on the bias voltage (by assuming no excess charge on the island, $n = 0$) leads to the appearance of a threshold voltage V_t given by

$$|V_\mathrm{b}| \geq V_\mathrm{t} = e/C_\Sigma. \tag{10}$$

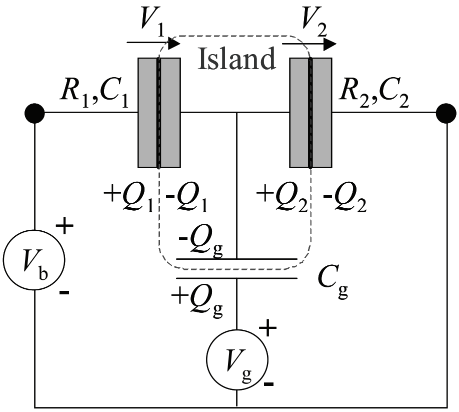


Fig. 3. – Schematic view of a SET transistor.

As long as the condition (10) is not fulfilled, no electron can be transferred, the current is blocked. This phenomenon, based on Coulomb repulsion, is named Coulomb blockade. The addition of one charge generates an electric field E which can stop the tunnel transfer of an excess electron through the first junction. $e^2/2C_\Sigma$ is the energy associated to the Coulomb blockade and this energy is in the first term of eqs. (8), (9). This expression reminds the purely classical model of electron-electron interaction based on the capacitive charge energy defined by Coulomb.

2˙2. *The single-electron transistor.* – In the previous section the principle of Coulomb blockade has been presented for an elementary device: two tunnel junctions in series. In this part let us introduce the SET transistor schematised in fig. 3. A gate electrode of capacitance C_g coupled to the island has been added in order to change the charge state of the island. Thus the number of charges on the island can be controlled by means of the gate voltage and the charge can be written as

$$Q_2 - Q_1 = -ne - C_\mathrm{g}(V_\mathrm{g} + V_2). \tag{11}$$

Note that a voltage offset applied to this additional gate electrode can compensate for the effects due to background charges coming from impurities or vacancies. From the relations (8) and (9) and with taking the energy stored by the gate capacitor into account, the free-energy changes during a tunnel event through the first or the second junction become

$$\Delta F_1{}^{\pm} = e^2/2C_\Sigma \pm eV_1 = (e/C_\Sigma)\big[e/2 \pm \big(-\big(C_2 + C_\mathrm{g}\big)V_\mathrm{b} + C_\mathrm{g}V_\mathrm{g} - ne\big)\big], \tag{12}$$

$$\Delta F_2{}^{\pm} = e^2/2C_\Sigma \pm eV_2 = (e/C_\Sigma)\big[e/2 \pm \big(-C_1V_\mathrm{b} - C_\mathrm{g}V_\mathrm{g} - ne\big)\big] \tag{13}$$

with $C_\Sigma = C_1 + C_2 + C_\mathrm{g}$.

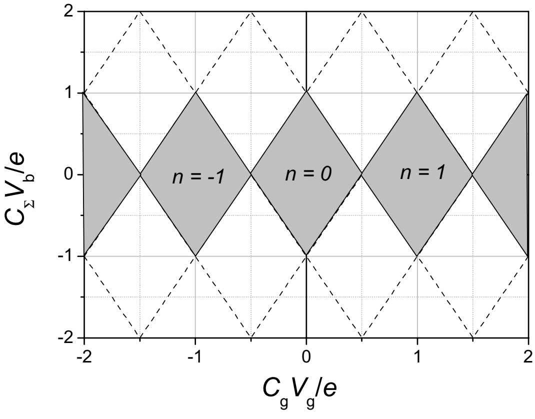


Fig. 4. – Stability diagram of a single-electron transistor showing the blocked (grey) and open (white) state domains.

As previously mentioned a tunnel event occurs only if it involves a decrease of the free energy. As a result, from inequalities $\Delta F_1{}^{\pm} \leq 0$, $\Delta F_2{}^{\pm} \leq 0$ and relations (12), (13), a stability diagram can be constructed. Such a diagram with a diamond shape allows us to display the Coulomb blockade regions in the $V_{\mathrm{b}} \otimes V_{\mathrm{g}}$ plane (fig. 4).

The grey regions correspond to stability domains with an integer number of excess electrons on the island. The probability of transmission through the barrier is very low, the current intensity is zero and the device is in the so-called blockade state. Everywhere else the transistor is in an open state and in this case the tunnelling of electrons through the circuit is possible. Note that at finite bias voltage below the threshold value $|V_{\mathrm{t}}| = e/C_\Sigma$ given in (10) the current oscillates with a period e/C_{g} with increasing gate voltage. Therefore the gate capacitance C_{g} can be estimated from the diamond diagram. Above $|V_{\mathrm{t}}|$ any Coulomb blockade does not arise. Whatever the voltage applied to gate electrode is, the current intensity is non-zero.

In order to well understand the origin of Coulomb blockade, let us refer to the energy band diagram sketched in fig. 5.

The Coulomb electrostatic energy $e^2/2C_\Sigma$ derived from the formulas (8) and (9) involves an energy band gap e^2/C_Σ shown in fig. 5a. The energy level denoted E_{n+1} corresponding to a single excess electron within the island is above the Fermi source energy, which makes an electron tunnelling from the left electrode impossible. The device is in a blockade state and the current is zero: this is the Coulomb blockade. Applying a voltage V_{g} to the gate electrode induces the lowering of the island energy levels (Fermi energy and E_{n+1}). Therefore E_{n+1} ends up being sandwiched between the both electrodes energy level in fig. 5b. As a result the electrons can cross the transistor from the source to the drain. Note that a similar situation can be achieved with increasing the bias voltage V_{b}. Each excess electron on the island tunnelling through the second junction leads to a drop in the energy to E_n which allows a second electron to penetrate into the island.

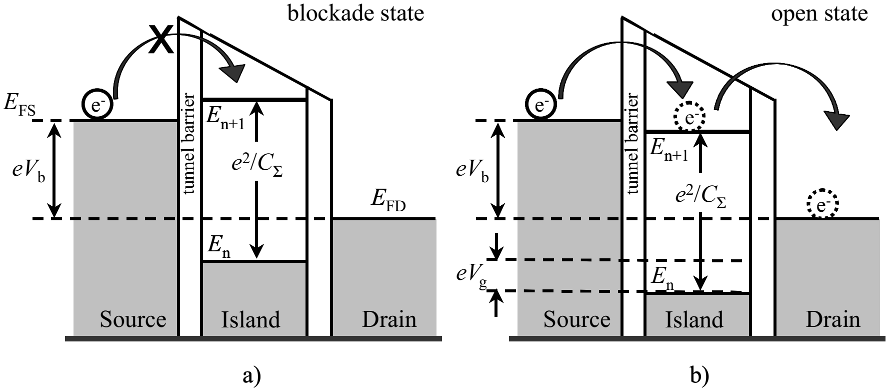


Fig. 5. – Energy band diagram before (a) and after (b) changing gate voltage. $E_{\mathrm{F,S}}$, E_{n} and $E_{\mathrm{F,D}}$ denotes the Fermi energy level of the source, island and drain, respectively. The gap between the source and drain energy levels is due to the bias voltage. n represents the free-electron number contained within island before tunnelling.

Finally, the Coulomb blockade phenomenon is illustrated by the measurements given in fig. 6. On the left picture the charge effects involve periodical changes of the transport properties with the gate voltage V_{g}. The period corresponds to an addition of one electron to the island. V_{g} can be adjusted so that the electron transfer through the device is blocked and so the current is zero. Consequently, the current can be suppressed thanks to two parameters: V_{g} and V_{b}. Below the threshold voltage no electron can tunnel and the current is zero as observed in fig. 6 right.

In this part we have described a device making the transfer of electrons one by one possible. However, a SET transistor is not capable of controlling the electron flow and so the current intensity. Designing a quantum current standard implies a more complex system than a SET transistor as described in the next part.

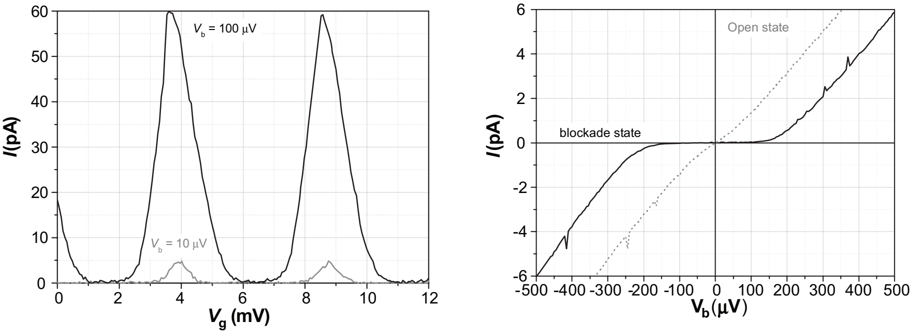


Fig. 6. – I-V characteristics of a SET transistor. Left: Coulomb oscillations obtained with varying gate voltage at various bias voltage. Right: I-V_{b} curve measured in the blockade and open states.

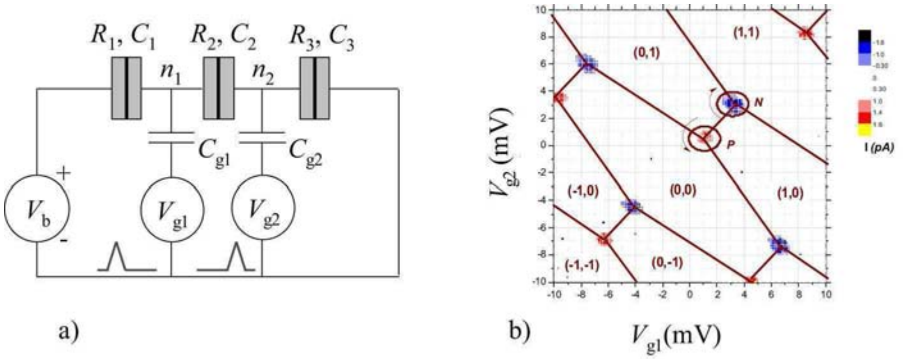


Fig. 7. – a) Schematic view of 3-junctions pump. b) Stability diagram in $V_{g1} \otimes V_{g2}$ plane which displays the stable configurations (n_1, n_2) of numbers of the excess electrons on each island. A RF signal at 10 MHz is applied to the gates. Boundaries between the domains (full lines) form a typical honeycomb pattern. The charges tunneling transfer takes place only in the triple points. These measurements have been carried out at LNE by means of a CCC used as a current amplifier.

2˙3. *The electron pump.* – The SET pump, first investigated by Pothier *et al.* [18] is a device allowing the transfer of electrons one by one at an adjustable clock frequency, f, and of a quasi-adiabatic way. Therefore, the electric current through the electron pump can be expressed by: $I = e \cdot f$. The simplest electron pump consists of two metallic islands separated by three junctions (ideally $C_1 = C_2 = C_3$, typically 150–200 aF). The gate voltages V_{g1} and V_{g2} through the gate capacitance C_{g1} and C_{g2} (typically around few tens of aF) can control the electric potential of each island (fig. 7a). The pump operation can be illustrated by means of the typical diagram given in fig. 7b which displays the stability domains of the different states (n_1, n_2) in the $V_{g1} \otimes V_{g2}$ plane.

The integer couple (n_1, n_2) denotes the number of excess charges located on the first and the second island respectively. The points (fig. 7b), so-called triple points, where conduction can take place, share three neighbouring domains. Everywhere else, the pump is in a blockade state and the electron configuration (n_1, n_2) is stable. Lines represent the boundaries between each stability domains and form a typical honeycomb pattern. The pumping of electrons is based on these topological properties.

The controlled transfer of electrons is obtained in the following way: two periodic signals with the same frequency f but phase shifted by $\Phi \approx 90°$ are superimposed on each applied d.c. gate voltage couple (V_{g10}, V_{g20}) as follows:

$$V_{g1} = V_{g10} + A \cdot \cos(2\pi \cdot f \cdot t),$$
$$V_{g2} = V_{g20} + A \cdot \cos(2\pi \cdot f \cdot t + \Phi).$$

In case the d.c. voltages (V_{g10}, V_{g20}) correspond to coordinates of the point denoted P, the circuit follows a closed trajectory around P as shown in fig. 7b. The

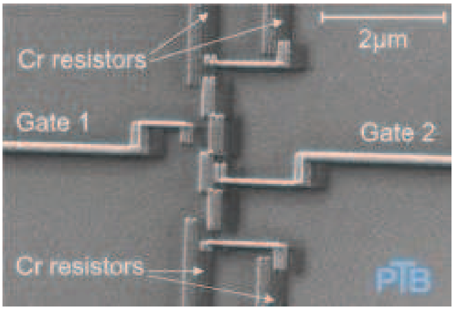


Fig. 8. – SEM-image of a 3-junctions R pump fabricated by PTB [23] (illustration by courtesy of PTB).

configuration changes from (0,0) to (1,0), then from (1,0) to (0,1), and returns to the initial state (0,0). In the real space, the complete sequence involves the transfer of one electron throughout the pump.

The frequency has to be lower than the reciprocal of the tunnel rate ($f \ll R_\mathrm{j}C$, R_j is the junction resistance, typically around 100–150 kΩ). This condition ensures that the system adiabatically returns to its ground state. By adding 180° to the phase shift Φ, the rotation sense is reversed in configuration space, and the electron by electron current takes place in the opposite direction [19]. The honeycomb pattern depends on gates and junctions capacitances and on cross-capacitances between the first gate electrode and the second island and *vice versa*. This effect can be compensated by means of an electronic device connected to both gate wiring inputs which adds a fraction of the voltage applied to one gate to the other gate, with opposite polarity [20].

The accuracy of the charge transfer is limited by three phenomena: thermal errors, frequency errors and co-tunneling effect [21]. The co-tunnelling effect is the most constringent one. This phenomenon involves simultaneous tunnelling of electrons from islands through each junction. In order to avoid errors in the transport rate, a first solution is the increase of the number of junctions. NIST has demonstrated that an error rate at a level of one part in 10^8 or less has been reached with a 7-junctions pump [22]. But, instead of it, PTB has proposed to keep 3-junctions pumps, the easiest to use, and to place on-chip resistive Cr-micro strips of typically 50 kΩ in series with the pump [19, 23], thus named R-pump (fig. 8). As a result, the dissipation of electron tunnelling energy in the resistors suppresses undesirable effects of co-tunnelling([2]) and an increased accuracy can be achieved.

In fig. 9 a set of I-V_b curves, named current steps, are shown and illustrates the quantization and the stability of the current generated by a R-pump with bias conditions at various frequencies. These characteristics are determining for the development of

([2]) A 3-junctions pump with a total Cr resistance of 50 kΩ is roughly equivalent, for co-tunneling, to a 5-junctions pump [23].

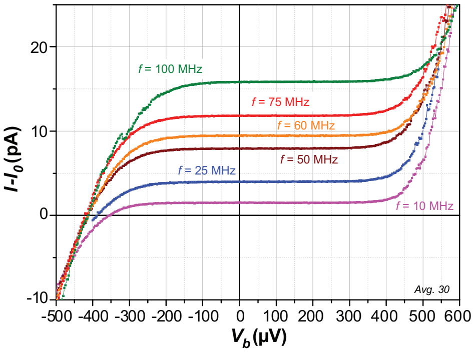


Fig. 9. – Current steps measured with a PTB R-pump operating at various pumping frequencies and at $T_{\mathrm{bath}} = 50\,\mathrm{mK}$.

current standards. Thus, stable current on $300\,\mu$V in a 40 fA range was obtained with a PTB R-pump connected to a CCC [24]. An investigation on long time measurements has shown that these pumps were able to generate a quantified current during more than 12 hours [19].

3. – Other single charge transport devices

3·1. *RF-SET-transistor–based electron counter.* – The bandwidth of a classical SET transistor used as an electrometer is typically around 1 kHz and can achieve 1 MHz with some improvements. But, it is too low to detect a 1pA current with a metrological accuracy, which requires a bandwidth of 10 MHz at least. Therefore, following the principle of the RF SQUID technology, a SET transistor is embedded in a tank circuit. Such a device, called RF-SET, can reach a charge resolution six orders of magnitude better than the commercial conventional detectors, the best result reported so far being $1.10^{-6}\,e/\mathrm{Hz}^{-1/2}$ [25].

The RF-SET is capacitively coupled to a long array of tunnel junctions, makes the electrons counting one-by-one possible (fig. 10) [26]. In a long array of tunnel junctions, charges flow in the form of regularly spaced solitons. Electrons generated by an external current source penetrate into the array of junctions and change the charging state of the island of the transistor when they come close to it. An incident RF signal is partially absorbed by the RF-SET if the transistor is in the open state or totally reflected in the blockade state. Consequently, this system is able to detect the crossing of an individual electron by counting each change of state. In principle, the aim should be to reach a counting speed of at least 60 MHz, corresponding to 10 pA with a 10 parts in 10^6 uncertainty. However, the best measurements reported so far show measured current less than 1 pA [27].

3·2. *SETSAW pump.* – The principle and the design of the electron transfer using a surface acoustic wave (SAW) generating a quantized current is quite different from the

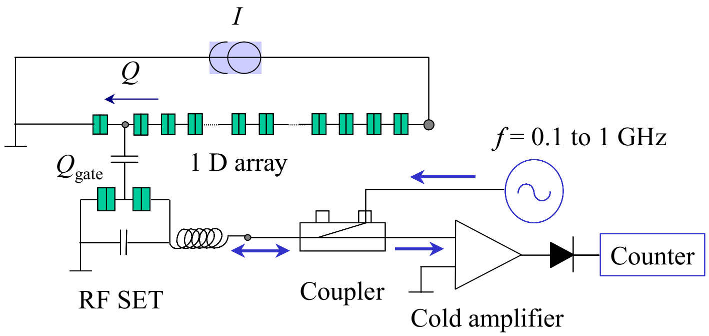


Fig. 10. – Basic circuit of an RF-SET electrometer. I is calibrated in terms of e and f by measuring the average frequency of the signal caused by the time correlated SET oscillations in 1D array.

one of the pumps, but the SETSAW devices remain interesting candidates for developing a current standard source or for quantum computing. A 2DEG in a heterostructure of GaAs/AlGaAs, very similar to those present within QHE devices, is confined to a one-dimensional (1D) channel by using split-gate technique (fig. 11). Thus, this channel is located between two electron reservoirs. By applying an appropriate voltage to the gate, the electron density in the constriction can be reduced to zero and an energy barrier for electrons appears. Due to the piezoelectric effect, a potential modulation is created, propagates through the SETSAW and is superposed to the energy barrier in the constriction area. Based on the Coulomb repulsion, it has been shown that an integer number of electrons, determined by the created well size, can be transferred through a SETSAW device and generates a current $I = N \cdot e \cdot f$ [28]. The maximum speed would be around 10 GHz.

For several years, the collaboration between the University of Cambridge and NPL has extensively investigated and developed a SETSAW current standard [29]. A total

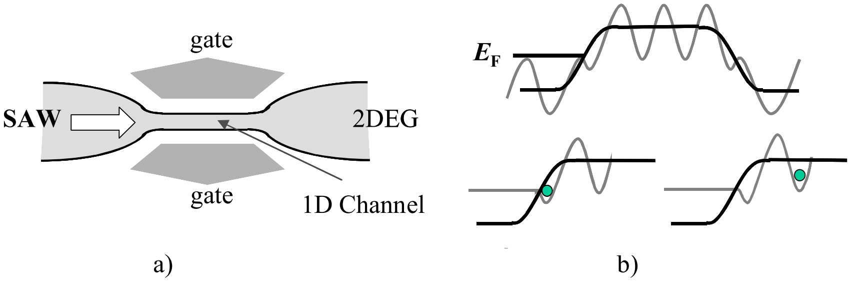


Fig. 11. – a) Schematic of the active part of a SETSAW device. b) Superposition of the surface acoustic wave and the barrier created by the split gates. The Fermi level of the 2DEG is indicated. The hollows of the modulation of the energy act like potential well which can propagate a single electron through the barrier.

current uncertainty of a few parts in 10^4 has been estimated but no real flat plateau has been displayed [30]. This lack of accuracy would no more ascribed to the overheating of electrons due to RF power needed by the transducer and by the speed of switching on and off of the propagating acoustic wave but would be explained by impurity effects [31]. From a model based on Coulomb blockade and a quantum dot within the 1D channel, it has been found that a current quantization $I = ef$ may occur at low RF power when SAW amplitude corresponds to a quantum dot charging energy and at gate voltage slightly exceeding the threshold value above which the channel is depleted. This single electron transport is maintained close to equilibrium, *i.e.* with limited overheating of electrons [31, 32]. We note ongoing developments of SETSAW pump based on carbon nanotubes [33, 34].

3˙3. *Cooper pair pump.* – In principle, the devices consisting of small-capacitance Josephson junctions forming superconducting islands coupled to gate electrodes are able to pump Cooper pairs one-by-one driven by a frequency higher than in the *normal* pump case. However, the tunnelling of cooper pairs is a phenomenon more complex than the one of electrons in the normal state because the Josephson coupling energy, E_{J} ($= hI_{\mathrm{C}}/(4\pi e)$ where I_{C} is the critical current of the Josephson junction) must be taken into account and compared directly to the charging energy, E_{c}. Nevertheless, with $E_{\mathrm{J}} < E_{\mathrm{c}}$, a current $I = \pm 2ef$ generated by a three-junction superconducting pump has been observed by several authors [14, 35, 36]. But, the transfer of the Cooper pairs across the device is disturbed by factors (Cooper pair co-tunnelling, quasi-particles poisoning ...) involving an imperfect plateau of the *I-V* curve. In order to improve the accuracy of the superconducting pumps, Zorin *et al.* have proposed to connect resistors in series to the ends of the array following the example of their R-type normal pumps [36]. The measurements show the through-supercurrent and the unwanted co-tunnelling events are dramatically suppressed.

3˙4. *New devices.* – Another approach to pump single Cooper pairs per cycle has been proposed by Niskanen *et al.* The device, referred to as Cooper pair sluice, consists of two mesoscopic SQUIDs forming between them a superconducting island, which is fitted with a gate [37,38]. The gate provides the possibility of coherent transfer of Cooper pair charges, one at a time, under the influence of an applied RF signal. Quantized currents of 10 up to 100 pA could be obtained with a calculated accuracy of one part in 10^7.

Different particular Josephson devices are currently investigated for the observation of the Bloch oscillations (see, *e.g.*, [39-41]). These are periodic oscillations which manifest on the voltage across a current biased single Josephson junction at frequency $f = I/2e$ [11, 42]. The Bloch voltage oscillations and the Josephson current oscillations are actually dual phenomena. Phase locking Bloch oscillations by an external microwave signal could yield current steps in the *I-V* characteristics. Very recently, it has been shown that the current range of 100 pA–1 nA could be attainable [40]. Following the example of Josephson voltage standards, the Bloch oscillation devices are thus expected to be promising candidates for realising quantum current standards.

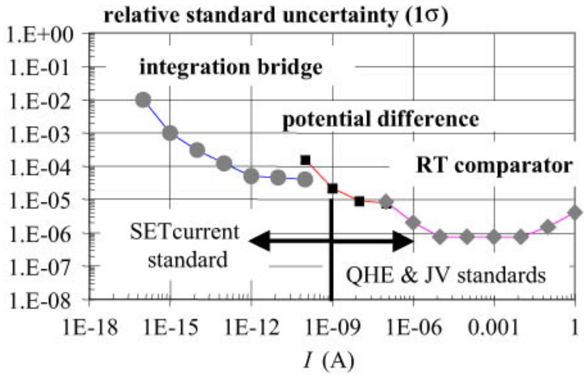


Fig. 12. – Different methods (current comparator at room temperature, potential difference bridge, integration bridge) used for the calibration of current less than 1 A and corresponding uncertainties. The integration method consists in the measurement of the rising time of the supplied current for a given voltage variation applied to a capacitance or inversely in the measurement of a varying voltage across a same capacitance along a known period.

To make up the list of single charge transport devices, we have to come back to non-superconducting devices and to mention the promising development of silicon-based SET pumps. The advantage of the silicon over aluminium lies not only in a possible higher pumping frequency (due to a smaller RC time constant), but also in a high stability of the background charge and a higher operating temperature ($T > 1\,\mathrm{K}$). First measurements on silicon-based SET pump with tuneable barriers have shown current steps at a level of 16 pA ($f = 100\,\mathrm{MHz}$) and at temperature as high as 20 K [43].

4. – New electrical standards and quantum metrological triangle experiments

4˙1. *Electrical standards.* – In practice, the ampere is reproduced by means of the ohm and the volt represented by the quantum Hall resistance standards (QHRS) and the Josephson array voltage standards (JAVS) respectively. On the other hand, the farad is reproduced by implementing a Thompson-Lampard calculable capacitor or by means of the QHRS associated to a measurement chain linking resistance to capacitance. However, the development of the Quantum-Metrological Triangle (QMT) experiments, which will be described below, requires quantum current or capacitance standards. SET devices are particularly well suited for these standards, which will play an important role for the "mise en pratique" of the possible redefinition of ampere and farad within the frame of a new SI to be implemented in mid term.

It is noteworthy that in shorter term the direct use of SET as primary current standard will be relevant for a current range less than 1 nA (fig. 12). This concerns the calibration of sub-nano ammeters (commercial electronic devices or home-made integration bridges) which allow national metrology institutes in electrical and ionising radiation domains to calibrate then their own secondary low-current standards and high-value resistance

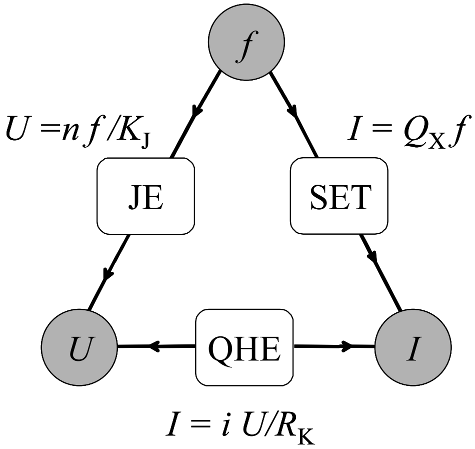


Fig. 13. – Quantum-metrological triangle. Theory predicts that R_K, K_J and Q_X correspond to the fundamental constants h/e^2, $2e/h$ and e.

standards ($R > 1\,\mathrm{T}\Omega$). The improvement of the traceability of small currents should benefit to some instrument manufacturers (detectors or meters of small electrical signals) and to the semiconductor industry (characterization of components, testing of wafers).

4˙2. *Quantum-Metrological Triangle.* – SET, or more widely, Single Charge Transport, provides the missing link of the Quantum-Metrological Triangle (QMT) (fig. 13) by realising a quantum current standard whose amplitude is only given by the product of the elementary charge by a frequency. The closure of the QMT experimentally consists here in applying Ohm's law $U = RI$ directly from the three phenomena SET, Josephson effect (JE) and Quantum Hall Effect (QHE). Another approach to close the triangle proposes to apply $Q = CV$ by means of a quantum capacitance standard.

In practice, the experiment amounts to determine the dimensionless product $R_K K_J Q_X$, expected to be equal to 2, where the constant Q_X is defined as an estimate of the elementary charge [12], $Q_X = e|_{\mathrm{SET}}$, by analogy with the definitions of Josephson and von Klitzing constants, $K_J = 2e/h|_{\mathrm{JE}}$ and $R_K = h/e^2|_{\mathrm{QHE}}$. Checking the equality $R_K K_J Q_X = 2$ with an uncertainty of one part in 10^8 will be a relevant test of the validity of the three theories.

The experiments testing the universal character of JE and QHE by showing that the Josephson-voltage-frequency quotient V_J/f and the product index of the QHE plateau times the resistance of the plateau $i \times R_H(i)$ are independent of materials at a level of parts in 10^{16} [44] and parts in 10^{10} [45,46], respectively, and the high level of agreement shown by numerous comparisons of quantum voltage and resistance standards (part in 10^{10} to a few parts in 10^9, respectively) ([6] and references therein) undoubtedly strengthen our confidence in the universal and fundamental aspects of K_J and R_K and hence in the equalities $K_J = 2e/h$ and $R_K = h/e^2$. However, even if strong theoretical arguments exist, the high reproducibility of the JE and the QHE, from a strictly metrological point of view, does not prove these relations.

The validity of these two relations has been recently tested by the CODATA Task group in the framework of the 2002 fundamental constants adjustment. It is shown that there is no significant deviation between K_J and $2e/h$ and between R_K and h/e^2 but within a fairly large uncertainty in the case of Josephson relation. The uncertainties amount to 8.5 and 2 parts in 10^8, respectively [8]. Very recently, Mohr *et al.* noted opposite significant deviations of K_J from $2e/h$ when the data coming from X-ray Crystal Density (XCRD) measurement on Si sphere $V_\mathrm{m}(\mathrm{Si})$ or from gyromagnetic ratio measurements in low field $\Gamma'(\mathrm{lo})$ are alternatively deleted from the set of input data [47] $K_\mathrm{J}/(2e/h) = 1-(273\pm 95)\times 10^{-9}$ and $K_\mathrm{J}/(2e/h) = 1+(546\pm 161)\times 10^{-9}$, respectively.

These discrepancies (different by 8 parts in 10^7), which perhaps could be correlated to the present discrepancy of one part in 10^6 on the value of h, measured either by means of watt balance or through XCRD measurement, emphasize the usefulness to close the triangle even with an uncertainty level of few parts in 10^7.

4·2.1. QMT by applying Ohm's law $U = RI$. The first way to close the QMT by applying Ohm's law on quantities provided by QHE, JE and SET consists in the direct comparison of the voltage U_J supplied by a Josephson junctions array to the Hall voltage of a QHE sample crossed by a current I delivered by a SET current source. The current is amplified with a high accuracy by means of a cryogenic current comparator (CCC) [12]. This comparison leads to the relation

$$U_\mathrm{J} = R_\mathrm{H} G I, \tag{14}$$

where G is the gain or winding ratio of the CCC. Considering the JE, QHE and SET relationships, eq. (14) becomes

$$n f_\mathrm{J}/K_\mathrm{J} = \big(R_\mathrm{K}/i\big) G Q_\mathrm{X} f_\mathrm{SET}, \tag{15}$$

where n is the index of the voltage step delivered by the JAVS at the microwave frequency f_J, i is the index of the QHE plateau and f_SET is the driving frequency of the SET current source. It leads to the dimensionless product

$$R_\mathrm{K} K_\mathrm{J} Q_\mathrm{X} = n(i/G) f_\mathrm{J}/f_\mathrm{SET}. \tag{16}$$

Another approach which leads to the same relation (16) consists in balancing the current delivred by the SET device against the current applied to a cryogenic resistor of high resistance value ($100\,\mathrm{M}\Omega$) by a Josephson voltage. The current is detected by a CCC operating as an ammeter [48]. Then the same CCC is used for calibrating directly the $100\,\mathrm{M}\Omega$ resistance with the quantum Hall resistance standard [49].

Measuring the deviation of $R_\mathrm{K} K_\mathrm{J} Q_\mathrm{X}$ from 2 will give information on the consistency level of the three quantum phenomena. It is noteworthy that the quantum charge involved in SET will be determined in terms of the elementary charge e if one assume $R_\mathrm{K} = h/e^2$ and $K_\mathrm{J} = 2e/h$.

4·2.2. QMT and electron counting capacitance standard. The development of a capacitance standard from SET devices, the so-called Electron Counting Capacitance Standard (ECCS), is feasible by applying the natural definition of the capacitance: the transfer of a well-known charge Q between the electrodes of a cryogenic capacitor with a capacitance C_cryo and the measurement of the potential difference ΔV between these electrodes: $C_\mathrm{cryo} = Q/\Delta V$ [13, 50].

Considering the controlled number N of pumped electrons during a given period and the measurement of ΔV by comparison to the voltage of a Josephson device biased on n-th Shapiro step and irradiated at frequency f'_J, the capacitance is given by the relation

$$C_\mathrm{cryo} = \big(N/nf'_\mathrm{J}\big)K_\mathrm{J}Q_\mathrm{X}. \tag{17}$$

The capacitance C_cryo is then compared to a known capacitance C_X of a capacitor placed at room temperature. Two kinds of results could be obtained.

1) If C_X has been previously measured in terms of the second and R_K, by means of a complete measurement chain whose the keystone is a quadrature bridge enabling the impedance comparison between capacitance and resistance ($2\pi RCf_\mathrm{q} = 1$ at equilibrium), it can be written in a simplified form as

$$C_\mathrm{X} = A_1/\big(R_\mathrm{K}f_\mathrm{q}\big), \tag{18}$$

where A_1 is a dimensionless factor issued from the measurement and f_q is the balance frequency of the quadrature bridge. Combining the two last relations (17) and (18) leads to a new expression of the dimensionless product $R_\mathrm{K}K_\mathrm{J}Q_\mathrm{X}$:

$$R_\mathrm{K}K_\mathrm{J}Q_\mathrm{X} = A_1(n/N)(C_\mathrm{cryo}/C_\mathrm{X})f'_\mathrm{J}/f_\mathrm{q}. \tag{19}$$

2) If the capacitance C_X has been directly compared to the capacitance variation ΔC of the Thompson-Lampard calculable capacitor ([2, 3] and references therein), and consequently known with a value expressed in SI units:

$$C_\mathrm{X} = \{C_\mathrm{X}\}_\mathrm{SI}\mathrm{F} = A_2\{\Delta C\}_\mathrm{SI}\mathrm{F}, \tag{20}$$

where the quantity inside brackets $\{\}_\mathrm{SI}$ is a dimensionless numerical value and, using the same notation as before, A_2 is a dimensionless factor. Then, from relation (17) a SI value of the product $K_\mathrm{J}Q_\mathrm{X}$ can be deduced

$$K_\mathrm{J}Q_\mathrm{X} = A_2(N/n)(C_\mathrm{cryo}/C_\mathrm{X})\big\{f_\mathrm{J}\Delta C\big\}_\mathrm{SI}\Omega^{-1} = A_3\big\{f'_\mathrm{J}\Delta C\big\}_\mathrm{SI}\Omega^{-1} \tag{21}$$

where $A_3 = A_2(N/n)(C_{\text{cryo}}/C_{\text{X}})$. Combining this SI determination of $K_{\text{J}}Q_{\text{X}}$ with the closure of the triangle *via* $U = RI$ leads to a new SI realisation of R_{K}

$$R_{\text{K}} = A_4\{f_{\text{SET}}\Delta C\}_{\text{SI}}{}^{-1}\,\Omega, \tag{22}$$

where $A_4 = A_3{}^{-1}n(i/G_{\text{CCC}})(f'_{\text{J}}/f_{\text{J}})$. The two relations (21) and (22) give rise to measurements of a quantity whose value might be compared to the values of h/e^2 deduced from measurements in atomic physics (anomalous magnetic moment of electron, ground state hyperfine transition frequency of muonium, quotient of Planck constant and either relative atomic mass of cesium or neutron mass times lattice spacing of a crystal), or in solid state physics (shielded gyromagnetic ratio of proton). These provides new determinations of α if one assumes exact the relations $R_{\text{K}} = h/e^2$, $K_{\text{J}} = 2e/h$ and $Q_{\text{X}} = e$. It is noteworthy that the first one is independent of QHE.

4˙2.3. Observational equations. It is thus shown that the QMT experiments do not consist solely in verifying the consistency of QHE, JE and SET. The closure of QMT *via* the two approaches $U = R\cdot I$ or $Q = C\cdot V$ might give significant information that can be taken into account in the adjustment periodically made on fundamental constants by the CODATA task group [8]. The adjustment is based on the method of least squares described in details by Mohr *et al.* in [51]. Using the same notation of the authors, the adjustment involves a large number of input data q_{i} (more than 80 in 2002), each of them being expressed as a function f_{i} of constants to be adjusted z_{i} (for example $\alpha, h, R_\infty, \ldots$) and giving rise to the set of observational equations.

$$q_{\text{i}} \doteq f_{\text{i}}(z_1, z_2, \ldots, z_{\text{M}}). \tag{23}$$

The previous relations (16), (19), (21) and (22) correspond to new observational equations as follows [52]:

$$Q_{\text{X-90}} \doteq \left[(K_{\text{J}}R_{\text{K}})/(K_{\text{J-90}}R_{\text{K-90}})\right]\cdot[2\alpha h/(\mu_0 c)]^{1/2}, \tag{24}$$

$$K_{\text{J}}Q_{\text{X}} \doteq 4\alpha/(\mu_0 c), \tag{25}$$

$$R_{\text{K}} \doteq \mu_0 c/(2\alpha), \tag{26}$$

where $Q_{\text{X-90}} = [I\times\text{A}/\text{A}_{90}]/f$, I is measured in conventional unit A_{90} and the assumptions $R_{\text{K}} = h/e^2$ and $K_{\text{J}} = 2e/h$ being relaxed.

4˙2.4. Determination of the elementary charge. It is noteworthy the great interest to combine all the three experiments, QMT, calculable capacitor and watt balance in a same laboratory or not. This would lead to a first direct determination of the quantum charge involved in SET devices, the expected electron charge, without assuming that $R_{\text{K}} = h/e^2$ and $K_{\text{J}} = 2e/h$. The best known value of e is at present the 2002 CODATA value ($e = 1.60217653\,\text{C}$, with an uncertainty of 8.5 parts in 10^8) mainly issued from values of the Planck constant h (*via* watt balance) and the fine-structure constant α (g-2, h/m ratio and QHE).

The watt balance provides the SI value of the product ${K_{\mathrm{J}}}^2 R_{\mathrm{K}}$

$$(27) \qquad {K_{\mathrm{J}}}^2 R_{\mathrm{K}} = A_5\{{f_{\mathrm{J}}}^2/[Mgv]\}_{\mathrm{SI}} \quad \Omega\,\mathrm{V}^{-2}\,\mathrm{s}^{-2},$$

where, as before, A_5 is a dimensionless factor, f_{J} is the Josephson frequency. M, g and v correspond to the suspended mass, Earth's gravitational acceleration and the constant speed.

The determination of R_{K} from the complete experiment linking the Thompson-Lampard calculable capacitor to the quantum Hall resistance standard gives

$$(28) \qquad R_{\mathrm{K}} = A_6\{(\Delta C f_{\mathrm{q}})^{-1}\}_{\mathrm{SI}}\,\Omega,$$

where A_6 is a dimensionless factor andf_{q} is the frequency of the balanced quadrature bridge.

These two experiments combined with QMT lead to the new observation equation

$$(29) \qquad Q_{\mathrm{X}} \doteq [2\alpha h/(\mu_0 c)]^{1/2}.$$

This relation can be deduced from (24) by considering that here the current I is measured in SI unit A. From the two approaches of QMT ($U = RI$ or $Q = CV$), the measured SI value of Q_{X} is given by

$$(30) \qquad Q_{\mathrm{X}} = A_7\Big\{[\Delta C f_{\mathrm{q}} Mgv]^{1/2}/f_{\mathrm{SET}}\Big\}_{\mathrm{SI}}\,\mathrm{C},$$

or

$$(31) \qquad Q_{\mathrm{X}} = A_8\Big\{[\Delta C Mgv/f_{\mathrm{q}}]^{1/2}\Big\}_{\mathrm{SI}}\,\mathrm{C},$$

where A_7 and A_8 are dimensionless factors.

4˙3. *QMT experimental set-up using a CCC*

4˙3.1. Cryogenic current comparator (CCC). In general, the CCC is used in NMIs to calibrate resistances against quantum Hall resistance standards. This is the instrument which has allowed to demonstrate the universality of QHE with the highest accuracy. The CCC can also be used as a low-current amplifier with two characteristics never reached by any conventional device. The CCC may exhibit a current resolution around $1\,\mathrm{fA/Hz}^{1/2}$ or less over the white-noise frequency range. This excellent resolution is mainly due to the low-noise properties of the magnetic flux detector used, currently a SQUID (Superconducting Quantum Interference Device) [53]. The second extraordinary feature of this cryogenic amplifier is the exactness of the current gain. The CCC is shortly described below. More details can be found in the literature [54].

The principle of CCC, invented by Harvey in 1972 [55], rests on Ampère's law and the perfect diamagnetism of the superconductor in the Meissner state. Given two wires

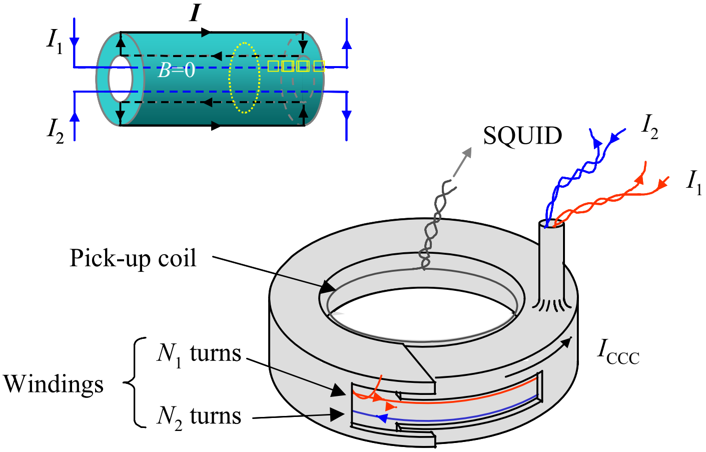


Fig. 14. – Toroidal structure of a CCC and principle (in insert). The supercurrent flowing up the inner surface of the toroidal shield is given by $I = N_1 I_1 + N_2 I_2$.

inserted in a superconducting tube (fig. 14), currents I_1 and I_2 circulating through these wires will induce a supercurrent I flowing up the inner surface of the tube and backing down the outer surface in such a way to maintain a null magnetic flux density B inside the tube.

Application of Ampère's law to a closed contour (a) in the bulk gives

$$\oint_a B \cdot \mathrm{d}l = 0 = \mu_0 \cdot (I_1 + I_2 - I) \tag{32}$$

and leads to the equality of the currents

$$I = I_1 + I_2. \tag{33}$$

If the tube contains N_1 and N_2 wires crossed, respectively, by currents I_1 and I_2, then (33) becomes

$$I = N_1 I_1 + N_2 I_2. \tag{34}$$

These equalities are valid independently of the position of the wires inside the tube. Here is the key reason of the high accuracy of the CCC. In practice, a CCC is made of two windings with N_1 and N_2 turns crossed by currents I_1 and I_2 circulating in opposite directions. These windings are enclosed in a superconducting torus [56], whose extremities overlap without being electrically connected on a length large enough to overcome the end effects, which distort the current equality in the real case of a finite length tube (fig. 14).

The outside magnetic flux Φ, which results only from the supercurrent I_{CCC}, is detected by a SQUID through a flux transformer composed of a pickup coil wound very close to the toroidal shield (on its inner or outer surface) and the input coil of the SQUID. The output voltage of the SQUID is then converted in a current, which feeds back one of the two windings to null the magnetomotive forces

$$N_1 I_1 - N_2 I_2 = 0. \tag{35}$$

From this ampere turn balance the equality of the ratios results

$$I_2/I_1 = N_1/N_2. \tag{36}$$

Except the case of a wrong number of turns in the windings, the error in the current ratio or equivalently in the current gain is in general very small, least measurable values of 10^{-11} having been reported so far. The ratio error comes mainly from a lack of efficiency of the superconducting toroidal shield (in a.c. measurements carried out at frequencies higher than 1 Hz, error sources arise from various capacitances inside the CCC).

The second relevant characteristic of the CCC is its current resolution δI_{CCC} in terms of $\mathrm{A/Hz}^{1/2}$ and is defined as the square root of the power spectral density of current noise referred to the CCC input, or equivalently as the minimum measurable supercurrent circulating in the overlapping tube of the CCC. The relation below gives a complete expression for δI_{CCC},

$$\delta I_{\mathrm{CCC}} = \left[4k_{\mathrm{B}}T/R_{\mathrm{in}} + 8\varepsilon/{N_1}^2 k^2 L'_{\mathrm{CCC}} + (S_{\Phi\mathrm{ext}}/N_1 L'_{\mathrm{CCC}})^2\right]^{1/2}, \tag{37}$$

where N_1 is the number of turns of the primary winding of the CCC, k is a coupling parameter between the pickup coil and the overlapping toroidal shield characterized by an effective inductance L'_{CCC} [54]. The first term corresponds to the Johnson noise of the input resistor R_{in} at temperature T. The second term is the contribution of the SQUID with an energy resolution ε when the optimal sensitivity of the CCC is reached. The third term comes from the external magnetic flux noise with a power spectral density $S_{\Phi\mathrm{ext}}$. This last term becomes negligible with careful shielding as described below. The dominant noise arises from one of the two first terms, depending on the CCC application. When a CCC is used for comparing resistance standards [6], the Johnson noise they deliver cannot be avoided and consequently the number of turns of the primary winding is increased to a limiting value (typically around 2000) above which the noise contribution of the SQUID becomes negligible. For low current measurements implying CCC-based current amplifier, very high input resistances are involved ($R_{\mathrm{in}} \gg 100\,\mathrm{M}\Omega$), and consequently only the SQUID noise contributes

$$\delta I_{\mathrm{CCC}} \approx \left[8\varepsilon/{N_1}^2 k^2 L'_{\mathrm{CCC}}\right]^{1/2}. \tag{38}$$

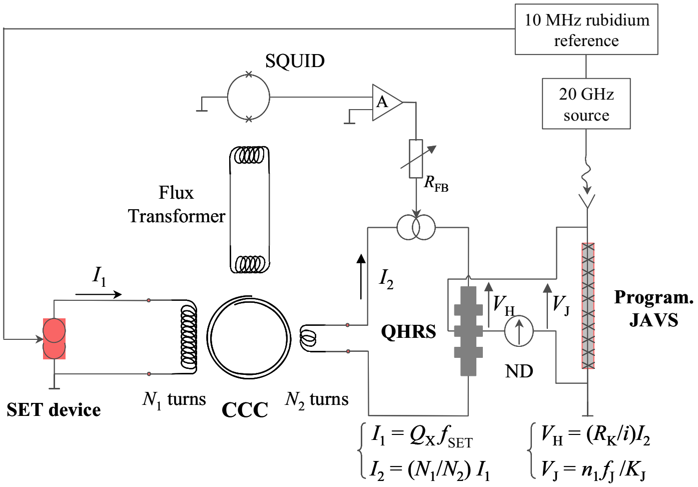


Fig. 15. – Basic circuit for closing the quantum metrological triangle.

4˙3.2. CCC used as a current amplifier. An experimental set-up for testing QMT is sketched in fig. 15. The current amplifier is composed of a CCC of high winding ratio, $G = N_1/N_2$, a d.c. SQUID with low white-noise level and low corner frequency f_c, and a secondary current source, servo controlled by the SQUID in such a way that the latter works at null magnetic flux and the relation (36) is verified [12]. In order to minimize the contribution from $1/f$ flicker noise, the polarity of the current to be amplified is periodically reversed. The Hall voltage is simultaneously compared to the voltage of a programmable Josephson junction array voltage standard (JAVS), well suited here because of the low voltage level and the requirement of periodic reversal of polarity [57]. The null detector will be balanced by adjusting the operating frequency of the SET source f_{SET}. This frequency and the irradiation frequency of the Josephson array are both referred to a 10 MHz rubidium clock.

The great challenge of this experiment is to reduce the type-A uncertainties to the level of few parts in 10^7 and ultimately one part in 10^8, taking into account the low current delivered by the SET source. The largest type-B uncertainties are estimated to be on the order of one part in 10^8 or less, and depend weakly on current level. They arise from the CCC ($u_{\mathrm{CCC}} \approx 10^{-8}$ including capacitive leakage, finite open loop gain and winding ratio error), the quantum Hall resistance standard ($u_{\mathrm{QHRS}} < 10^{-9}$ considering the effects of finite temperature, contact resistance and resistive leakage), the Josephson voltage set-up ($u_{\mathrm{JE}} < 10^{-8}$ mainly due to residual e.m.f., resistive leakage, detector and frequency error) and the SET experimental set-up ($u_{\mathrm{SET}} < 10^{-9}$ coming from current leakage and frequency error) without taking into account intrinsic errors of SET device (missing events due to pumping frequency, cotunneling and thermal effects).

In order to analyze the different noise sources and to estimate the possible type-A uncertainties, let us consider a simplified diagram of the QMT experiment as shown

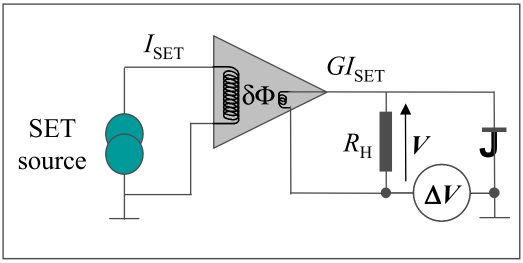


Fig. 16. – Principle of a QMT experiment involving a CCC-based current amplifier.

in fig. 16. Two detectors are used: the SQUID which detects the magnetic flux $\delta\Phi$ induced by the supercurrent in the CCC and to be zeroed, and a voltmeter to measure the deviation ΔV between the Hall and Josephson voltages.

The voltage noise recorded by the voltmeter and referred to the nominal voltage V is given by the expression

$$(39)\qquad \delta V/V = \left[\delta I^2{}_{\mathrm{CCC}} + 4k_{\mathrm{B}}T/G^2R_{\mathrm{H}} + \left(\delta V^2{}_{\mathrm{ND}} + \delta V^2{}_{\mathrm{emf}}\right)/G^2R^2{}_{\mathrm{H}}\right]^{1/2}/I_{\mathrm{SET}},$$

where one finds again in the first term the current resolution of the CCC directly linked to magnetic flux noise $\delta\Phi$ at the SQUID input. The second term reflects the Johnson noise of the quantum Hall resistance, while the third term corresponds to the noise generated by the null detector (ND) itself including the instability of unwanted electromotive forces.

The required CCC for amplifying the very small current generated by the SET source must present high winding ratio and ultra low noise performances. In this framework, some CCCs with winding ratios from 10 000:1 to 109 999:1 have been investigated by NMIs [58-63]. They present input current noise $\delta I_{\mathrm{CCC}}|_{\mathrm{exp}}$ from 0.8 to 4 fA/Hz$^{1/2}$ in the white-noise range (typically $f > 0.1$ Hz) although measured in the favourable case where the CCC was not connected to SET device. The input current noise undoubtedly increases once the input winding of the CCC is connected to a SET current source (due to increased influence of antenna loop effects, microphonics effects). For example, δI_{CCC} has been found in the order of 12 fA/Hz$^{1/2}$ with a CCC connected to an electron pump compared to the initial value 4 fA/Hz$^{1/2}$ [64]. All The experimental values $\delta I_{\mathrm{CCC}}|_{\mathrm{exp}}$ are around ten times higher than the expected values $\delta I^2{}_{\mathrm{CCC}}|_{\mathrm{theo}}$ varying from 80 to 700 aA/Hz$^{1/2}$. Some improvements have thus to be done for reducing this margin and a value of 1 fA/Hz$^{1/2}$ might be considered as a first reasonable target.

The Johnson noise term is made negligible by designing CCC with a gain higher than 10^4. Considering a QHE sample cooled at current temperature of 1.3 K and operating on the $i = 2$ resistance plateau ($R_{\mathrm{H}}(i = 2) = R_{\mathrm{K}}/2 \approx 13\,\mathrm{k}\Omega$), the term $(4k_{\mathrm{B}}T/G^2R_{\mathrm{H}})^{1/2}$ is only in the order of 7 aA/Hz$^{1/2}$.

The null detector generates both noise voltage δU_{ND} and noise current δI_{ND}. The latter depends on the resistance placed at the input of the detector and might become dominant if a threshold resistance value is exceeded. The last term of relation (39) can then be expressed by

$$(40)\quad \big(\delta V^2{}_{\mathrm{ND}} + \delta V^2{}_{\mathrm{e.m.f.}}\big)/G^2R^2{}_{\mathrm{H}} = \big(\delta U^2{}_{\mathrm{ND}} + \delta V^2{}_{\mathrm{e.m.f.}}\big)/G^2R^2{}_{\mathrm{H}} + \delta I^2{}_{\mathrm{ND}}(R_{\mathrm{H}})/G^2.$$

Considering the previous case with $G = 10^4$ and $R_{\mathrm{H}} \approx 13\,\mathrm{k}\Omega$, the noise characteristics $\delta I_{\mathrm{ND}} \approx 100\,\mathrm{fA/Hz}^{1/2}$ and $\delta U_{\mathrm{ND}} \approx 20\,\mathrm{pV/Hz}^{1/2}$ (corresponding to an equivalent noise resistance of $40\,\Omega$) of the EM model N11 nanovoltmeter and $\delta V_{\mathrm{e.m.f.}} \approx 1\,\mathrm{nV/Hz}^{1/2}$ typically, it results

$$\big[\big(\delta V^2{}_{\mathrm{ND}} + \delta V^2{}_{\mathrm{e.m.f.}}\big)/G^2R^2{}_{\mathrm{H}}\big]^{1/2} \approx 14\,\mathrm{aA/Hz}^{1/2}.$$

This value is well below the CCC current resolution.

With the target value of $1\,\mathrm{fA/Hz}^{1/2}$ for δI_{CCC}, the total noise detected by the voltmeter and referred to the input voltage might amount to $\delta V/V \approx 1 \times 10^{-3}/\mathrm{Hz}^{1/2}$ for a current of 1 pA.

Finally, the experimental standard deviation of the mean for a set of data can be estimated, either with a power spectral density calculus either with an Allan deviation one [65, 66]. These calculi can be carried out only if the noise is white. In this case, $s(\bar{I})$ represents the type-A uncertainty associated to the mean value of the current $\bar{I}$ [67] and the Allan deviation is an unbiased estimator of $s(\bar{I})$ [65, 66].

$$(41)\qquad s\big(\bar{I}\big)/\bar{I} = [h_0 f_0/(2N)]^{1/2},$$

where h_0 ($= \delta V^2 \cdot \bar{I}/V^2$) is the white-noise level (in A^2/Hz) as defined in [66], f_0 is the sampling frequency of the measurement and N is the total number of values.

The condition of white noise will be fulfilled if the current to be measured is periodically reversed at a frequency slightly higher than the corner frequency f_{C} of the SQUID. f_{C} defines the frontier between the domains of $1/f$ noise and white noise. In fig. 17 we report a typical set of data over a measurement time T_{N} and composed of N reversals of current. f_{M} denotes the modulation frequency of a single measurement (two current reversals) and n the number of values per trace taken with an integrating frequency $f_{\mathrm{i}} : f_{\mathrm{M}} = f_{\mathrm{i}}/n$, $T_{\mathrm{N}} = N/f_{\mathrm{M}}$.

The sampling frequency to be considered is equal to the frequency f_M and consequently

$$s\big(\bar{I}\big)/\bar{I} = [h_0 f_{\mathrm{M}}/(2N)]^{1/2},$$

or

$$(42)\qquad s\big(\bar{I}\big)/\bar{I} = [h_0/(2T_{\mathrm{N}})]^{1/2}.$$

For $h_0^{1/2} = 1\,\mathrm{fA/Hz}^{1/2}$ and considering the typical values used at LNE [24], $f_{\mathrm{M}} = 0.14\,\mathrm{Hz}$, the experimental standard deviation of the mean $s(\bar{I})$ from a one-hour measurement

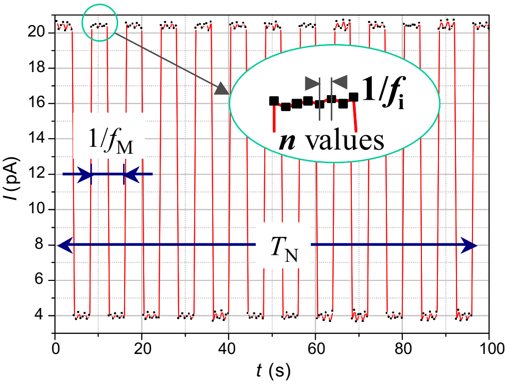


Fig. 17. – Set of values recorded at the output of the SQUID of the CCC during a measurement carried out at LNE. The current delivered by an electron pump is periodically reversed at a frequency so that the SQUID works in the white-noise regime.

would amount to 10 aA, *i.e.* around 7.4 parts in 10^6 for a current of 1.6 pA delivered by an electron pump operating at 10 MHz. This is already lower than the uncertainties obtained for the same current amplitude with the best conventional bridge (involving integration method) so far.

With the system at LNE, using a CCC in non-optimal working mode([3]) (the SQUID being flux-locked by feeding the current to its modulation coil and not to the secondary winding of the CCC), the current of 16 pA generated by an electron pump at 100 MHz has been measured with a type A uncertainty of 55 aA (*i.e.* a relative uncertainty of 3.5 parts in 10^6) after 6.5 hours of measurement [64] in agreement with ${h_0}^{1/2}$ measured at the level of 12 fA/Hz$^{1/2}$. From this result, some improvements have to be made both on R pump and CCC used as current amplifier to reach an uncertainty below one part in 10^6. SET devices capable to supply currents up to 100 pA and generating lower noise must be developed in order to attempt the ultimate uncertainty of 1 part in 10^8. A new amplifier with a better sensitivity is also needed. This can be obtained by increasing the CCC gain with a factor 5 and by using a SQUID well suited to the experiment.

4·3.3. CCC used both in the calibration of a cryogenic resistance and as current detector. The principle of QMT experiment implying a CCC as a current detector instead of a current amplifier is sketched on fig. 18. It presents the advantage of involving a single detector with a very low input current noise and allowing current reversals faster than in the previous case, *i.e.* at frequencies of 1 Hz or higher [48], thus operating the SQUID far from the $1/f$ noise regime.

[3] *i.e.* in an internal feedback mode, where the SQUID is flux-locked by feeding the current to its modulation coil and not to the secondary winding of the CCC.

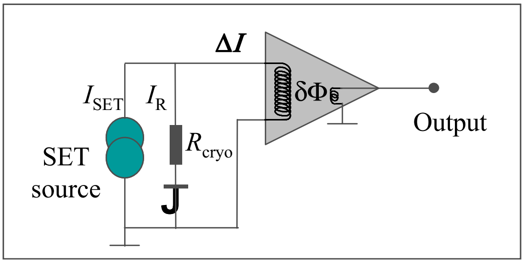


Fig. 18. – Principle of a QMT experiment with a CCC operating as a SQUID ammeter, the primary winding being the input coil coupled to the SQUID via a flux transformer. Induced by a voltage generated by a Josephson array to a cryogenic resistor, a current I_R is opposed to a current I_{SET} delivered by a SET device.

The expected value of the type-A uncertainty on the current deviation $\Delta I = I_{SET} - I_R$ will be given by

$$s(\bar{I})/\bar{I} = \big(\delta I/\bar{I}\big)/T_N{}^{1/2} \tag{43}$$

with

$$\delta I = \big(4k_B T/R_{cryo} + \delta I^2{}_{CCC}\big)^{1/2} \tag{44}$$

by taking into account the intrinsic current noise of the CCC given by the relation (38). A cryogenic resistor of 100 MΩ, as proposed at NIST [48], cooled down to 1.3 K (in a simple pumped helium bath) generates a current noise of about 800 aA/Hz$^{1/2}$, in the same order of the noise of the null detector described in 5.2.2. making the two approaches equivalent in terms of type-A uncertainties which could be reached.

In this experiment the challenge to overcome is the calibration of the cryogenic resistance, whose value of 100 MΩ has to be measured in terms of R_K with an uncertainty as low as one part in 10^8. This requires a specific CCC bridge enabling a direct comparison with QHR [49].

4·4. *QMT experimental set-up using an electron counting capacitance standard.* – The QMT has been successfully closed within an uncertainty of one part in 10^6 at NIST [50]. As shown in fig. 19, the system consists of a seven junctions electron pump, a SET transistor/electrometer with a charge detection threshold of the order of $e/100$, and a cryogenic capacitor of 1.8 pF capacitance. Two mechanical cryogenic switches N_1 and N_2 allow two working phases:

a) *N_1 closed, N_2 open*

In this phase, the cryogenic capacitance C_{cryo} ($\approx$ 1 pF) is charged with N electrons generated one by one through the pump. The process is stopped for a short time (20 s) to measure the voltage $V_c{}^+$. Then, the pump is forced to transfer N electrons in the opposite

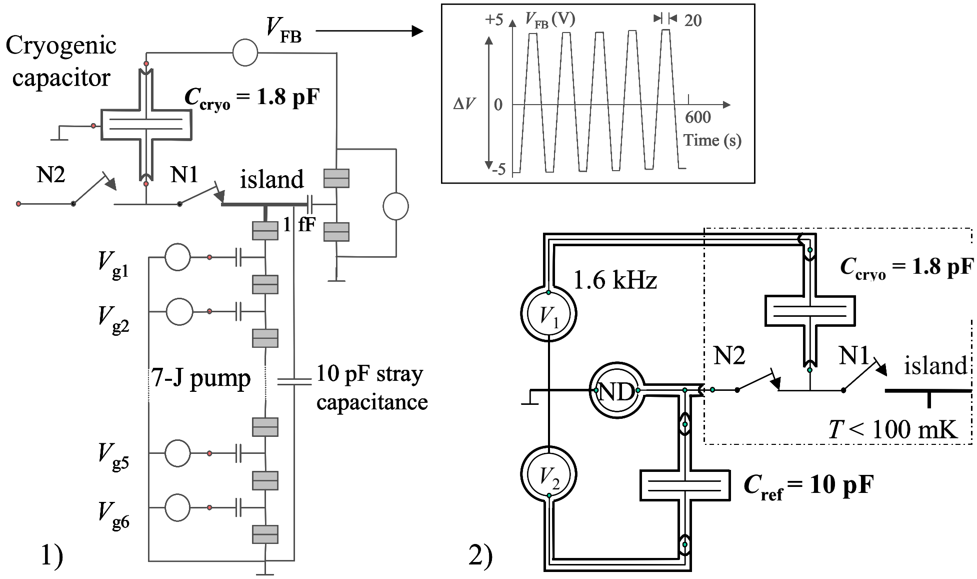


Fig. 19. – Principle of electron counting capacitance standard [50]. Two operating modes: 1) N_1 closed, N_2 open: electrons are periodically pumped forward and backward, 2) N_1 open, N_2 closed: The cryogenic capacitance is compared to a reference capacitance.

direction. Another stop occurs to measure a voltage ${V_c}^-$, and so on. The successive voltages ${V_c}^+$ and ${V_c}^-$ are compared to those of a JAVS and the differences $\Delta V = {V_c}^+ - {V_c}^-$ are calculated. The average of these differences $\langle \Delta V \rangle$ gives the capacitance

$$C_{\mathrm{cryo}} = Ne/\langle \Delta V \rangle = \big(N/nf_{\mathrm{J}}\big)K_{\mathrm{J}}Q_{\mathrm{X}} \tag{45}$$

from the relation $\langle \Delta V \rangle = nf_{\mathrm{J}}/K_{\mathrm{J}}$, where n is the index of the voltage step provided by the binary Josephson array at a frequency f_{J}.

b) N_1 *open,* N_2 *closed*

In this second configuration, C_{cryo} is compared with the capacitance C_{x} of a capacitor at room temperature using a capacitance bridge. This capacitance comparison is carried out at a frequency in the kHz range, much higher than the effective frequency of electron counting (25 mHz) and at 15 V of rms voltage value (compared to 3.5 V in the first phase).

Keller *et al.* [68] has recently established a complete uncertainty budget on the results reported in 1999 [50]. The combined total uncertainty amounts to around 9.5 parts in 10^7 and mainly results from the uncertainty components given below.

For the electron counting phase, the relative standard deviation of C_{cryo} values is in the order of 1.4 to 2.4 parts in 10^7 depending on the run performed (three runs in total). The type-A uncertainties lie within 1.2 and 2.3 parts in 10^7 while the two most important type-B uncertainties amount to 5 parts in 10^8 and 4 parts in 10^8 corresponding to the calibration of the digital voltmeter against JAVS and the capacitive leakage of

C_{cryo}, respectively. About the electron pumping error, earlier measurements have shown uncertainties of one part in 10^8 at frequencies of a few MHz [22].

For the capacitance comparison phase, the commercial bridge used was traceable to the calculable capacitance standard at NIST with a calibration uncertainty of 8.5 parts in 10^7. This corresponds to the most important type-B uncertainty, the other are in the order of 2 parts in 10^7 (for example the correction of cable loading effects), while the type-A uncertainty here has been found ten times lower.

Last but not least, type-B uncertainties corresponding to the frequency and voltage dependences of the cryogenic capacitance have to be taken into account. They amount to 2 parts in 10^7 [69] and 9 parts in 10^8 [68], respectively.

Unlike the $U = RI$ approach, this experiment does not need a SET source supplying currents higher than a few pA to close the triangle at least with an uncertainty of one part in 10^7. This uncertainty level will be reached by this method if efforts are undertaken particularly on the capacitance measurement. A drastic improvement will be in implementing a coaxial capacitance bridge based on two terminal-pair method [70]. This will be absolutely required in order to reduce more the uncertainty and to reach the level of few parts in 10^8. This also needs a better knowledge of the frequency dependence of the cryogenic capacitor. Presently the observed logarithmic increase of the capacitance when the frequency decreases below few 100 Hz might be due to dielectric dispersion and dissipation of insulating films of $\mathrm{Cu_2O}$ formed on the surface of the electrodes [69].

Within this frame, developments of ECCS are also in progress at other NMIs. A cryogenic capacitor with highly symmetrical coaxial electrode arrangement has been developed at the PTB: $C_{\mathrm{cryo}} = 1.435\,\mathrm{pF}$ at 4.2 K, drift $< 1\,10^{-7}$/day when the capacitor is maintained at low temperature, $\Delta C_{\mathrm{cryo}}/C_{\mathrm{cryo}} = 1.3\,10^{-6}$ during several thermal cycling [71]. In addition to the highly reliable cryogenic switches, METAS has designed and fabricated a tuned capacitor [72]. The capacitance value can be adjusted at room temperature so that the value is equal to the nominal value 1 pF within 3 parts in 10^5 allowing to take advantage of high-precision capacitance bridges.

5. – Conclusion and prospects

The development of the Coulomb blockade nanodevices opens extended prospects for applications in fundamental electrical metrology, *i.e.* the development of current and capacitance standards, and, more crucially, the closure of the quantum-metrological triangle and the determination of the elementary charge. These experiments could contribute to establish a new frame of the SI, fully based on fundamental constants by creating a direct link between the fundamental physics and the units. The target uncertainty needs to be around few parts in 10^7 and then ultimately one part in 10^8. If there is no deviation, our confidence on the three phenomena to provide us with $2e/h$, h/e^2 and e will be considerably enhanced. Any significant discrepancy will prompt further experimental and theoretical work. The closure of the quantum-metrology triangle, at these required uncertainties, should be assisted by improvements in new SET devices which could generate accurate currents as high as 100 pA. Ideas currently under investigation for the implementation of a higher-frequency-locked current source include improved R-pumps, silicon-based electron pumps, SETSAW devices, superconducting Cooper pair pumps as a generalization of a single electron pump, Cooper pair sluices, Bloch oscillation devices, *etc.* Efforts have also to be pursued and encouraged in the improvement of CCCs, cryogenic capacitors and associated measurement techniques.

The Coulomb blockade nanodevices also present a high metrological potential in the applied domain in electricity and ionising radiation (calibration of sub-nano ammeters and development of charge detector), in thermometry (absolute cryogenic thermometer with so-called Coulomb blockade thermometer invented by Pekola *et al.* [73] and commercially available), in nanometrology (nanometer scale displacement sensor) and in new fields based on single-photon sources (single- or multiple-photon discrimination metrology, quantum cryptography and computing). Moreover some encouraging preliminary results and the advances in nanofabrication techniques (miniaturization of the tunnel junctions) will improve the performances of SET devices and allow them to operate at higher temperatures in future.

All these emerging applications in addition to those reaching maturity in fundamental electrical metrology (quantum standards based on QHE samples arrays and Josephson junctions arrays) and in time and frequency domain (microwave frequency standards by cooling atoms, femtosecond optical frequency combs, ion optical clocks ...) give signs on the metrology of the future, a science more and more focused on the measurement of discrete quantities rather than continuous quantities by detecting, manipulating, counting elementary entities (electron or charge quantum, flux quantum, photon ...). This evolution explains the present discussion around the new formulation of the SI which should express more our present knowledge on quantum physics and the date of its implementing will mark this crossover.

REFERENCES

[1] Kovalevsky J. and Quinn T. J., *C. R. Phys.* **5** (2004).

[2] Thompson A. M. and Lampard D. G., *Nature*, **177** (1956) 888.

[3] Trapon G., Thévenot O., Lacueille J. C. and Poirier W., *Metrologia*, **40** (2003) 159.

[4] Von Klitzing K., Dorda G. and Pepper M., *Phys. Rev. Let.*, **45** (1980) 494.

[5] Prange R. E. and Girvin S. M. (Editors), *The Quantum Hall Effect* (Springer-Verlag, New York) 1990.

[6] Jeckelmann B and Jeanneret B., this volume, p. 135.

[7] Josephson B. D., *Phys. Lett.*, **1** (1962) 251.

[8] Mohr P. J. and Taylor B. N., *Rev. Mod. Phys.*, **77** (2005) 1.

[9] Kibble B. P., in *Atomic Masses and Fundamental constants 5*, edited by Sanders J. H. and Wapstra A. H. (Plenum, New York) 1976, pp. 545-551.

[10] Genevès G. *et al.*, *IEEE Trans. Instrum. Meas.*, **54** (2005) 850.

[11] Likharev K. and Zorin A., *J. Low Temp. Phys.*, **59** (1985) 347.

[12] Piquemal F. and Genevès G., *Metrologia*, **37** (2000) 207.

[13] Williams E. R., Ghosh R. N. and Martinis J. M., *J. Res. Natl. Stand. Technol.*, **97** (1992) 299.

[14] Averin D. V. and Likharev K. K., in *Mesoscopic Phenomena in Solids*, edited by Al'tshuler B. I., Lee P. A. and Webb R. A. (Elsevier, Amsterdam) 1990, p. 173.

[15] Grabert H. and Devoret M. H. (Editors), *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures* (Plenum Press, New York) 1992.

[16] Likharev K. K., *Proc. IEEE*, **87** (1999) 606.

[17] Zeller H. R. and Giaver I., *Phys. Rev.*, **181** (1969) 789.

[18] Pothier H. *et al.*, *Europhys. Lett.*, **17** (1992) 249.

[19] Feltin N., Devoille L., Piquemal F., Lotkhov S. and Zorin A., *IEEE Trans. Instrum. Meas.*, **52** (2003) 599.

[20] Keller M. W., Martinis J. M., Steinbach A. H and Zimmerman N. M., *IEEE Trans. Instrum. Meas.*, **46** (1997) 307.

[21] Jensen H. D. and Martinis J. M., *Phys. Rev. B*, **46** (1992) 13407.

[22] Keller M. W., Martinis J. M., Zimmerman N. M and Steinbach A. H, *Appl. Phys. Lett.*, **69** (1996) 1804.

[23] Lotkhov S. V., Bogoslovsky S. A., Zorin A. B. and Niemeyer J., *Appl. Phys. Lett.*, **78** (2001) 946.

[24] Steck B. *et al.*, *CPEM Digest* (2006) 158.

[25] Brenning H., Kafanov S., Duty T., Kubatkin S. and Delsing P., *J. Appl. Phys.*, **100** (2006) 114321.

[26] Schoelkopf R. J., Wahlgren P., Kozhevnikov A. A., Delsing P. and Prober D. E., *Science*, **280** (1998) 1238.

[27] Bylander J. *et al.*, *Nature*, **434** (2005) 361.

[28] Shilton J. M. *et al.*, *J. Phys.: Condens. Matter*, **8** (1996) L531.

[29] Ebbecke J., Fletcher N. E., Ahlers F. J., Hartland A. and Janssen T. J. B. M., *IEEE Trans. Instrum. Meas.*, **52** (2003) 594.

[30] Fletcher N. E., Janssen T. J. B. M. and Hartland A., *BEMC Digest* (2001).

[31] Fletcher N. E. *et al.*, *Phys. Rev. B*, **68** (2003) 245310.

[32] Alhers F. J., Kieler O. F. O., Sagol B. E., Pierz K. and Siegner U., *CPEM Digest* (2006) 154.

[33] Leek P. *et al.*, *Phys. Rev. Lett.*, **95** (2005) 256802-1.

[34] Shin Yun-Sok *et al.*, *CPEM Digest* (2006) 228.

[35] Geerligs L. J. *et al.*, *Z. Phys. B*, **85** (1991) 349.

[36] Zorin A. B., Bogoslovsky S. A., Lotkhov S. V. and Niemeyer J., Cond-mat/0012177 (2000).

[37] Niskanen A. O., Pekola J. P. and Seppä H., *Phys. Rev. Lett.*, **91** (2003) 177003/1.

[38] Kemppinen A. *et al.*, *CPEM Digest* (2006) 164.

[39] Haviland D. B., Kuzmin L. S., Delsing P., Likharev K. K. and Claeson T., *Z. Phys. B: Condens. Matter*, **85** (1991) 339.

[40] Boulant N. *et al.*, Cond-mat/0605061 (2006).

[41] Lotkhov S. V., Krupenin V. A. and Zorin A. B., *CPEM Digest* (2006) 166.

[42] Gallop J. C., *Philos. Trans. R. Soc. London, Ser. A*, **363** (2005) 2221.

[43] Fujiwara A., Zimmerman N. M., Ono Y. and Takahashi Y., *Appl. Phys. Lett.*, **84** (2004) 1323.

[44] Tsai J. S., Jain A. K. and Lukens J. E., *Phys. Rev. Lett.*, **51** (1983) 316.

[45] Hartland A., Jones K., Williams J. M., Gallagher B. L. and Galloway T., *Phys. Rev. Lett.*, **66** (1991) 969.

[46] Jeckelmann B., Inglis A. D. and Jeanneret B., *IEEE Trans. Instrum. Meas.*, **44** (1995) 269.

[47] Mohr P. J., Newell D. B. and Taylor B. N., private communication (2006).

[48] Elmquist R., Zimmerman N. M. and Huber W. H., *IEEE Trans. Instrum. Meas.*, **52** (2003) 590.

[49] Elmquist R. E., Hourdakis E., Jarret D. G. and Zimmerman N. M., *IEEE Trans. Instrum Meas.*, **54** (2005) 525.

[50] Keller M. W., Eichenberger A. L., Martinis J. M. and Zimmerman N. M, *Science*, **285** (1999) 1706.

[51] Mohr P. J. and Taylor B. N, *J. Phys. Chem. Ref. Data*, **28** (1999) 1713.

[52] Piquemal F. *et al.*, *C. R. Physique*, **5** (2004) 857.

[53] Clarke J. and Braginski A. I. (Editors), *The SQUID Handbook*, Vol. **I** and Vol. **II** (WILEY-VCH Verlag GmbH&Co.FgaA, Weinheim), 2004-2006.

[54] Gallop J. C. and Piquemal F., chapter 9 in *The SQUID Handbook Vol.* ***II***, edited by Clarke J. and Braginski A. I. (WILEY-VCH Verlag GmbH&Co.FgaA, Weinheim) 2006, pp. 95-137.

[55] Harvey I. K., *Rev. Sci. Instrum.*, **43** (1972) 1626.

[56] Sullivan D. B. and Dziuba R. F., *Rev. Sci. Instrum.*, **45** (1974) 517.

[57] Hamilton C. A., Burroughs C. J. and Kautz R. L., *IEEE Trans. Instrum. Meas.*, **44** (1995) 223.

[58] Hartland A., *BEMC Digest* (1993) 18/1.

[59] Janssen T. B. J. M. and Hartland A., *Phys. B.*, **284-288** (2000) 1790.

[60] Gay F., Piquemal F. and Genevès G., *Rev. Sci. Instrum.*, **71** (2000) 4592.

[61] Gay F., PhD Thesis, Conservatoire National des Arts et Métiers, Paris, France (2000).

[62] Bartolomé E., PhD Thesis, Twente University, The Netherlands (2002).

[63] Rietveld G. *et al.*, *IEEE Trans. Instrum. Meas.*, **52** (2003) 621.

[64] Steck B. *et al.*, to be submitted to *Metrologia* (2007).

[65] Allan D. W., *IEEE Trans. Instrum. Meas.*, **36** (1987) 646.

[66] Witt T., *IEEE Trans. Instrum. Meas.*, vol. **50** (2001) 445.

[67] *Guide to the Expression of Uncertainty in Measurement*, International Standardization Organisation (ISO), ISBN 92-67-10188-9 (1993).

[68] Keller M. W. and Zimmerman N. M., to be submitted to *Metrologia* (2007).

[69] Zimmerman N. M., Simonds B. J. and Wang Y., *Metrologia*, **43** (2006) 383.

[70] Kibble B. P and Rayner G. H., *Coaxial AC bridges*, edited by Bailey A. E. (Adam Hilger Ltd., Bristol) 1984.

[71] Willemberg G. D. and Warnecke P., *IEEE Trans. Instrum. Meas.*, **50** (2001) 235.

[72] Overney F., Jeanneret B. and Furlan M., *IEEE Trans. Instrum. Meas.*, **49** (2000) 1326.

[73] Pekola J. K., Hirvi K. P., Kauppinen J. P. and Paalanen M. A., *Phys. Rev. Lett.*, **73** (1994) 2903.

Cooling, trapping and manipulation of atoms and Bose-Einstein condensates: Applications to metrology(*)

K. Helmerson and W. D. Phillips

Atomic Physics Division, Physics Laboratory, National Institute of Standards and Technology - Gaithersburg, MD 20899, USA

1. – Introduction

The thermal motion of atoms often imposes limitations on the precision and accuracy of measurements involving those atoms. Atomic beam cesium clocks are one outstanding example, where the typical thermal velocity of the cesium atoms, on the order of 200 m/s, limits the observation time to a few milliseconds in a typical apparatus. This finite observation time results in a linewidth for the microwave resonance whose frequency defines the unit of time. Other motional effects include first- and second-order Doppler effects whose presence limits the performance of such clocks. Because conventional refrigeration would result in the condensation of cesium or other atoms into a solid, researchers sought other means of reducing the thermal motion, a quest that led to laser cooling and electromagnetic manipulation of neutral atoms. Similar techniques have been developed for trapped ions and used in metrological applications, but these are not treated here.

These notes are based on a series of lectures given at the Enrico Fermi Summer School on Recent Advances in Metrology and Fundamental Constants held in Varenna in July and August of 2006. The notes include a general discussion of the mechanical effects of light; laser cooling and trapping techniques for neutral atoms; experiments with Bose-

(*) Official contribution of the National Institute of Standards and Technology, not subject to copyright in the United States.

Einstein condensed sodium and rubidium atoms at the National Institute of Standards and Technology (NIST) in Gaithersburg, Maryland; and applications of these techniques to metrology.

The topics to be covered in the general discussion of the mechanical effects of light and laser cooling and trapping techniques include: radiative forces, both spontaneous and dipole; various topics related to trapping neutral atoms, including magnetic traps, both static and time-averaged orbiting potentials (TOP); laser dipole traps and far-off resonant traps (FORTs); radiation pressure traps and the optical Earnshaw theorem; magneto-optical traps (MOTs); mechanisms limiting the loading of MOTs and techniques to increase the densities, including dark spot MOTs; Doppler cooling and the Doppler cooling limit; deceleration and cooling of atomic beams; optical molasses and the discovery of sub-Doppler laser cooling; and new laser cooling mechanisms, including polarization gradient and Sisyphus heating.

These topics have been selected as representing some of the important developments in the interaction of electromagnetic fields and atoms which have enabled the development of new metrological tools as well as the creation and manipulation of Bose-Einstein condensates. More details on these and related topics can be found in a number of notes, references, and review articles [1-13].

The later sections of these notes discuss metrology applications of cold atoms, particularly clocks and atom interferometers, as well as Bose-Einstein condensation (BEC) and its potential applications to metrology.

Before beginning the discussion of the mechanical effect of atom-light interactions, we define, for convenience, some frequently used notation and symbols, with frequencies expressed in radians/second. The reader is cautioned that other authors may have used other conventions:

$\delta = \omega_{\text{laser}} - \omega_{\text{atom}}$ is the detuning of the laser frequency from the natural resonant frequency of the atom.

$\Gamma = \tau^{-1}$ is the decay rate of the population in the excited state, the inverse of the natural lifetime, the linewidth of the transition.

$k = 2\pi/\lambda$ is the laser photon wave vector.

k_{B} is Boltzmann's constant.

Ω is the on-resonance Rabi frequency for a laser field, the precession frequency of the Bloch vector representing the 2-level atom when $\delta = 0$.

$I/I_0 = 2\Omega^2/\Gamma^2$ is the normalized intensity of the laser.

M is the mass of the atom.

$v_{\text{rec}} = \hbar k/M$ is the recoil velocity of an atom upon emission or absorption of a single photon.

$E_{\text{rec}} = Mv_{\text{rec}}^2/2 = \hbar^2 k^2/2M$ is the kinetic energy of an atom having velocity v_{rec}.

2. – Radiative forces

Light comes in quantized packets of energy and momentum called photons. The transfer of energy and momentum between the photon and an atom through either coherent or incoherent scattering results in a force exerted on the atom. This is the basis for the mechanical effects of atom-light interactions. Radiative forces or the mechanical effects of light are generally divided into two categories—the scattering force and the dipole force.

2˙1. *The scattering force.* – The scattering force, also called the spontaneous or radiation pressure force, is the force exerted on an atom by the incoherent scattering of photons. More precisely it is the force on an atom corresponding to absorption of a photon followed by spontaneous emission. The photon absorbed transfers $\hbar k$ of momentum to the atom. The spontaneous emission of the photon is symmetrically distributed and so the momentum transfer due to spontaneous emission, averaged over many absorption-emission events, is zero. The average force on a two-level atom moving with velocity $\mathbf{v}$ in a plane wave of wave vector $\mathbf{k}$ and detuning δ is

$$\mathbf{F}(\mathbf{v}) = \hbar \mathbf{k} \frac{\Gamma}{2} \frac{I/I_0}{1 + I/I_0 + [2(\delta - \mathbf{k}\cdot\mathbf{v})/\Gamma]^2}. \tag{1}$$

This force is the maximum scattering rate $\Gamma/2$, times the resonance Lorentzian times the photon momentum, *i.e.* the rate of absorbing photon momentum. The detuning $\delta - \mathbf{k}\cdot\mathbf{v}$ accounts for the Doppler shift, and the force is large when $|\delta - \mathbf{k}\cdot\mathbf{v}| \leq \Gamma$. The spontaneous force is limited by the rate at which spontaneous emissions can occur. These occur at a rate Γ for excited atoms, whose maximum fractional population is 1/2.

For real, multilevel atoms, the situation can be more complicated. A common occurrence, typical of alkali atoms, is that the ground state is split by the hyperfine interaction into two states separated in frequency by many times the optical linewidth Γ. An atom excited by a laser from one of these hyperfine levels to an optically excited state may decay by spontaneous emission to the other hyperfine level. Transitions from this level are then so far out of resonance that effectively no further absorption occurs and no force is applied to the atom. While various schemes involving selection rules and polarization of the light may be used to avoid this problem of optical pumping, the most straightforward method is to apply a second laser frequency, tuned to resonance between the "wrong" hyperfine state and the optically excited state. This "repumper" keeps the atom out of the wrong ground state and allows the atom to effectively feel the force of the laser acting on the main transition.

Equation (1) is only valid if the force can be meaningfully averaged over many absorption-emission events. If a single event changes the atomic velocity so much that the resonance condition is not satisfied, such an average is not possible. This imposes a validity condition $\hbar k^2/M \ll \Gamma$ on eq. (1). This condition is well satisfied for most atoms of interest in laser cooling. For example, $M\Gamma/\hbar k^2$ is 1200 for cesium laser cooled on its resonance transition at 852 nm and is 200 for sodium cooled at 589 nm.

2˙2. *The dipole force.* – The dipole force, also called the gradient force or stimulated force, is the force exerted on an atom due to coherent redistribution of photons. The dipole force can be considered as arising from stimulated Raman events—the absorption

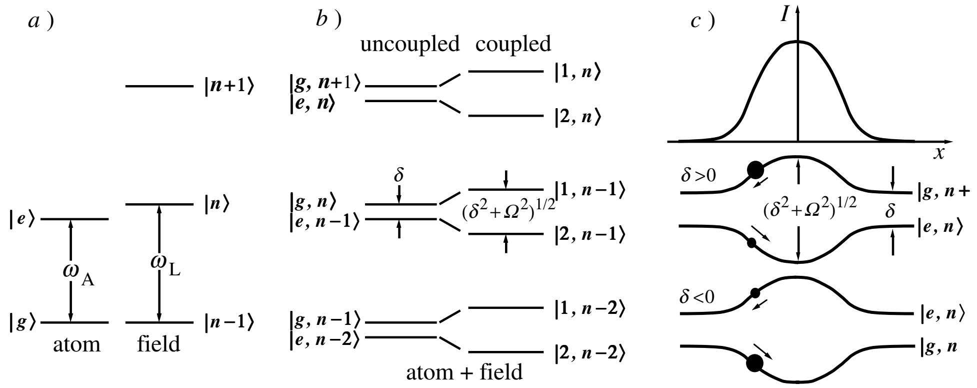


Fig. 1. – *a*) Energy levels of a 2-level atom and a laser field in the bare basis, *b*) dressed basis with the atom uncoupled and coupled to the field. *c*) Top: the intensity profile of the laser field. Bottom: the dressed energy levels as a function of position for this intensity profile, for laser detuning above and below resonance. The sizes of the black dots indicate the relative population of the dressed states.

and stimulated emission of photons. (Note that the absorption and stimulated emission cannot be thought of as successive and independent events; their correlation is central to the proper understanding of the force.)

We can also understand the dipole force in analogy to a driven, classical oscillator. A harmonically bound charge driven by an oscillating electric field **E** has an oscillating dipole moment μ which is in phase with the driving field when driven below resonance, and out of phase when driven above resonance. The energy of interaction between the dipole and field is $W = -\mu \cdot \mathbf{E}$. Below resonance the energy is negative and the oscillator will be drawn toward a more intense field, while above resonance it will be drawn to the weaker part of the driving field.

A particularly powerful and intuitive way of describing the dipole force is in the dressed-atom picture. This has been treated in detail by Dalibard and Cohen-Tannoudji [14] and we present the basic idea here. Consider a 2-level atom with ground state $|g\rangle$ and excited state $|e\rangle$. Separately consider a single mode of the radiation field, close to resonance with the atomic transition, having energy levels labeled $\ldots, |n-1\rangle, |n\rangle, |n+1\rangle, \ldots$ according to the number of photons in the mode. The energy levels of these two distinct systems are shown in fig. 1*a* for the case where the photon energy ω_L is greater than the difference in energy between the atomic states, ω_A. This is the "bare" basis in which we consider the atom's energy levels and those of the photon field separately. If we now consider the atom and laser field together as a single system we have the dressed basis (atom dressed by laser photons) of fig. 1*b*. If there is no interaction between the atom and the laser field (as, for example, when the laser field does not spatially overlap the atom's position) the dressed level energies are simply the sum of the atom and field energies. This gives a ladder where each rung is a pair of nearly degenerate energy levels, since an atom in the ground state with n photons in the laser field has nearly the same energy as an atom in the excited state with $n-1$ photons in the field (with the energy difference being equal to the laser detuning). When the atom interacts with the field, the dressed levels (the eigenstates of the full Hamiltonian) become superpositions of ground and excited states and of the numbers of photons in the field. By convention, the higher of the two levels in a rung is called $|1\rangle$, and the rungs are labeled by the photon number n associated with the excited state in the uncoupled basis [15]. According to the general rule that interacting energy levels repel each other by an amount depending on the coupling, the spacing between the dressed levels increases from the detuning δ to the "effective Rabi frequency" $\Omega_{\text{eff}} = \sqrt{\delta^2 + \Omega^2}$ as the interaction turns on. Because each of the dressed levels has both ground and excited state character, an atom can make spontaneous emission transitions between the rungs of the ladder. These transitions establish an equilibrium population between the two types of levels [14].

Now consider a light field whose intensity varies in space, such as a focused laser beam with a Gaussian intensity profile, or a plane standing wave. The Rabi frequency seen by an atom now varies in space, so the energy of the dressed levels also varies. The dipole force arises from this variation in energy and the relative populations of the dressed levels. Figure 1*c* illustrates the idea. For laser detuning above resonance ($\delta > 0$) the upper of the two dressed levels is the one that connects to the ground state in the limit of small interaction. This upper level always has the largest population, and its potential tends to repel the atom from the region of most intense light. For $\delta < 0$ it is the lower level that connects to the ground state and that has the higher population; its potential attracts the atom to the higher intensity. The actual dipole force is the average force, weighted by population, for the two potentials. The dipole force is derivable from a potential [14, 16-19], which can be written as

$$U = \frac{\hbar\delta}{2} \log\left[1 + \frac{I/I_0}{1 + (2\delta/\Gamma)^2}\right], \tag{2}$$

where the spatial dependence of the potential comes in through the dependence of the intensity I on position.

2˙3. *Doppler cooling.* – Doppler cooling results from near resonant radiation pressure acting on an atom and thus it is a particularly instructive example of an application of the scattering force. To see how cooling works we shall consider the one-dimensional motion of an atom in counterpropagating plane waves. If the laser beams are tuned below the atomic resonance, an atom moving in the direction opposite to one of the beams will, because of the Doppler shift, see that beam shifted closer to its resonance frequency. At the same time it sees the other laser beam, propagating in the same direction as its velocity, shifted further away from resonance. The atom absorbs more photons from the beam propagating opposite to its velocity and thus slows down. From eq. (1), the forces due to the positive- and negative-going waves, along the direction of propagation, are

$$F_{\pm}(v) = \pm \hbar k \frac{\Gamma}{2} \frac{I/I_0}{1 + I/I_0 + [2(\delta \mp kv)/\Gamma]^2}. \tag{3}$$

If we assume that $I/I_0 \ll 1$, so that the intensity is low enough for us to add the forces from the two waves independently, and that $kv \ll \Gamma$, the total force is given by

$$F_{\text{tot}} = F_+ + F_- = \frac{4\hbar k^2 I/I_0}{[1 + (2\delta/\Gamma)^2]^2} \frac{2\delta}{\Gamma} v = -\alpha v. \tag{4}$$

For $\delta < 0$, α is positive and the force damps the velocity at a rate $\dot{v}/v = -\alpha/M$. The largest damping force is obtained for $\delta \cong -\Gamma/2$.

Atomic motion in a strong standing wave [14, 16, 17, 20] is beyond the scope of this treatment, however a simple approximation for small velocity, ignoring interference of the two beams and stimulated redistribution of photons between the beams [21], shows that the friction coefficient α maximizes at about $\alpha = \hbar k^2/4$ for $I/I_0 = 1$ in each beam and $2\delta/\Gamma = -1$. Under these conditions the velocity damping time $v/\dot{v}$ for sodium, cooled on the 589 nm resonance line, would be 13 μs.

So far we have only considered the average force on an atom in a light field. We should remember that the force arises from discrete photon scatterings, and so must fluctuate about its average. The fluctuations can be thought of as arising from two sources: fluctuations in the number of photons absorbed in a given time and fluctuations in the direction of the spontaneously emitted photons. Both of these effects arise because of the randomness of spontaneous emission.

The fluctuations represent a random walk of the atomic momentum, with each random walk step being of magnitude $\hbar k$. For simplicity we will assume a fictitious one-dimensional situation where photons are emitted as well as absorbed along a single axis. Each scattering event represents two random walk steps, one from the absorption, which could be from either of the counterpropagating beams, and one from the spontaneous emission, which can be in either direction along the axis. The mean square momentum of the atom increases linearly with the number of scattering events (random walk steps), with a rate

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle p^2 \rangle = 2R\hbar^2 k^2, \tag{5}$$

where R is the scattering rate; the factor of 2 comes from the two steps per scattering. We define a momentum diffusion coefficient

$$2D_p = \frac{\mathrm{d}}{\mathrm{d}t}\langle p^2 \rangle, \tag{6}$$

so that D_p/M is the rate of increase in kinetic energy, the heating rate. The damping force decreases the kinetic energy as $\mathbf{F} \cdot \mathbf{v} = -\alpha v^2$, the cooling rate. At equilibrium we set the sum of the heating and cooling rates to zero, finding

$$\frac{D_p}{\alpha} = M\langle v^2\rangle = k_{\mathrm{B}}T. \tag{7}$$

Here we have replaced v^2 with its mean value and used the equipartition theorem to identify $k_{\mathrm{B}}T/2$ as the mean kinetic energy $M\langle v^2\rangle/2$ in the single degree of freedom. Using eq. (4) to get α, eqs. (5) and (6) to get D_p, and remembering that the total scattering rate from the two laser beams, each of intensity I, is

$$R = \Gamma\frac{I/I_0}{1+(2\delta/\Gamma)^2}, \tag{8}$$

we find, for low intensity and small velocity:

$$k_{\mathrm{B}}T = \frac{\hbar\Gamma}{4}\frac{1+(2\delta/\Gamma)^2}{2\delta/\Gamma}. \tag{9}$$

This is the Doppler temperature, and we emphasize that it applies to the fictional one-dimensional case we have constructed. A true 1D experiment, such as cooling an atomic beam along one axis, would produce a lower temperature depending on the distribution of scattered photons in 3D. However, it can be shown [21] that in a symmetrical three-dimensional case, the temperature is also given by eq. (9). The temperature minimizes at a detuning of $\delta = -\Gamma/2$, where

$$k_{\mathrm{B}}T_{\mathrm{Dopp}} = \frac{\hbar\Gamma}{2} \tag{10}$$

defines the Doppler cooling limit. To derive the Doppler temperature we assumed that the velocity was small enough that $kv \ll \Gamma$. At the Doppler limit, sodium atoms would have $v_{\mathrm{r.m.s.}} = 30\,\mathrm{cm/s}$, corresponding to a temperature of $240\,\mu\mathrm{K}$ and $kv_{\mathrm{r.m.s.}} = \Gamma/20$, so the assumption is justified in this case (and most others). We may also ask whether the velocity distribution corresponds to a temperature. That is, is it Maxwell-Boltzmann? It can be shown that for sodium, and similar atoms where the recoil energy $E_{\mathrm{rec}} = \hbar^2k^2/2M \ll \hbar\Gamma$, the velocity distribution is indeed very close to being thermal [21]. What happens if the recoil energy is not small? Then we violate the validity condition on eq. (1), and the detuning due to the Doppler shift changes significantly with each emission or absorption. In the limit where the linewidth Γ is small compared to the recoil energy, we feel intuitively that the cooling limit, rather than being related to Γ, as in eq. (10), is related to the recoil energy. Indeed, it can be shown [21, 22] that in this case the lowest temperature attainable (the recoil temperature) is given by

$$\frac{1}{2}k_{\mathrm{B}}T = \frac{\hbar^2k^2}{2M} = E_{\mathrm{rec}} \tag{11}$$

that is, one recoil energy per degree of freedom. While it might seem that this is the ultimate limit of laser cooling, in fact it is possible to break the recoil limit under certain circumstances where the interaction with the light is turned off as the atomic velocity becomes small. One way this may be achieved is by velocity selective coherent population trapping in multilevel atoms [23, 24]. Atoms are optically pumped into a coherent superposition of both internal states and center-of-mass velocities, a superposition that cannot absorb the laser light. Another way is velocity-space optical pumping [25, 26]. Atoms are cooled on a transition with a narrow linewidth, but with a distribution of laser frequencies such that the excitation rate for zero velocity atoms is small or vanishing. This has been accomplished by use the use of pulsed, two-photon Raman transitions [27].

The damping force of eq. (4) is similar to the viscous force on an object moving in a fluid. Because of this, the configuration of pairs of counterpropagating laser beams is often called "optical molasses". The viscosity is so high that a velocity corresponding to the Doppler cooling limit is damped out and randomized while the atom travels only a few tens of micrometers, much smaller than the size of the molasses, which is typically a centimeter. Thus the atoms executes a Brownian-like motion with a short mean free path, moving diffusively rather than ballistically [28,29]. The evidence for this is the long residence time of atoms in optical molasses. Atoms require several seconds to diffuse out of a typical optical molasses [29, 30], whereas, moving ballistically, atoms cooled to the Doppler cooling limit would traverse a region the size of a typical molasses in a few tens of milliseconds.

3. – Deceleration and cooling of an atomic beam

Experimentally, atoms must first be at reasonably slow velocities ($kv \leq \Gamma$) before Doppler cooling can be effective. The two general ways this has been accomplished are by laser deceleration of an atomic beam [28, 31] and by collection of slow atoms from a thermal gas [32, 33]. Loading from a thermal gas has the advantage of allowing a more compact and simpler apparatus, while the beam deceleration technique usually allows lower background pressure and faster production of slow atoms.

Deceleration of an atomic beam is usually accomplished by directing a near resonant laser beam so as to oppose the atomic beam. The atoms absorb photons at a rate determined by the intensity of the laser beam, the detuning from resonance and the atoms velocity. For each photon absorbed, the atomic velocity changes by v_{rec} in the direction of the laser propagation. The spontaneously emitted photons are emitted randomly in a pattern that is symmetric on reflection through the atom, so there is no net average change in the atomic velocity due to these emissions. If the absorption is followed by stimulated emission into the same direction as the incident laser beam (we assume the laser beam is a plane wave), there is no net momentum transfer from the absorption-emission process. Only absorption followed by spontaneous emission contributes to the average force, which is given by the rate of scattering photons times the momentum of a photon. For a two-level atom this force is given by eq. (1). At high intensity this force saturates to the value $\hbar k\Gamma/2$.

The acceleration of an atom due to the saturated radiation pressure force is $a_{\rm max} = \hbar k\Gamma/2M = v_{\rm rec}\Gamma/2$, which can be quite large. For sodium with $\lambda = 2\pi/k = 589\,{\rm nm}$, $1/\Gamma = 16\,{\rm ns}$ and $M = 23\,{\rm a.m.u.}$, $v_{\rm rec} \approx 3\,{\rm cm/s}$ and $a_{\rm max} \approx 10^6\,{\rm m/s^2}$. For cesium, with $\lambda = 2\pi/k = 852\,{\rm nm}$, $1/\Gamma = 30\,{\rm ns}$ and $M = 133\,{\rm a.m.u.}$, $v_{\rm rec} \approx 3.5\,{\rm mm/s}$ and $a_{\rm max} \approx 6 \times 10^4\,{\rm m/s^2}$. This acceleration would stop in 50 cm, a thermal, 1000 m/s Na atom scattering ≈ 33000 photons in 1 ms and in 80 cm, a thermal, 300 m/s Cs atom scattering ≈ 84000 photons in 5 ms.

Implicit in eq. (1) is one of the major impediments to effective deceleration of an atomic beam using a counterpropagating laser beam. The force acting on the atom is large if $|2(\delta - \mathbf{k}\cdot\mathbf{v})| \leq \Gamma\sqrt{1 + I/I_0}$. Atoms much outside this resonant-velocity range will experience little deceleration, and atoms initially within this range will be decelerated out of it. This process results in a cooling or velocity compression of a portion of the atomic beam's velocity distribution, and was first observed by Andreev *et al.* [34]. Atoms initially at the resonant velocity decelerate out of resonance. Other atoms with nearby velocities will also decelerate, those with larger velocities first decelerating into resonance, then to slower velocities out of resonance, while initially slower atoms decelerate to still lower velocities. The atoms will "pile up" at a velocity somewhat lower than the resonant velocity. Both deceleration and cooling occur because a range of velocities around the resonant velocity are compressed into a narrower range at lower velocity. However, only a small portion of the total velocity distribution has been decelerated by only a small amount.

There are a number of possible solutions to this problem, some of which have been discussed in ref. [28]. These include Zeeman tuning [28] where a spatially varying magnetic field compensates the changing Doppler shift as the atoms decelerate, so as to keep the atoms near resonance; white-light deceleration [35] where a range of laser frequencies ensures that some light is resonant with the atoms, regardless of their velocity (within the range to be decelerated); diffuse-light deceleration [36] where light impinges on the atoms from all angles so that, with the Doppler shift, some of the light is resonant with each velocity; Stark cooling [37] where a spatially varying electric field is used to Stark shift atoms and keep them near resonance as they Doppler shift due to deceleration; intense standing wave deceleration [38] where the linewidth is sufficiently power broadened to capture a large velocity distribution for laser deceleration; and "chirp cooling" [39] in which the frequency of the laser is swept up, or chirped, in time such that the laser stays in resonance with atoms that have been decelerated, and they continue to absorb photons and decelerate. Zeeman tuning and chirp cooling were among the earliest atomic beam slowing techniques demonstrated and they continue to be the most widely used techniques to date.

3·1. *Chirp cooling.* – In chirp cooling, the frequency of the laser is swept up, or chirped, in time [39]. Because of the chirp, atoms that have been decelerated by the laser stay in resonance, continue to absorb photons, and continue to decelerate. Furthermore, the chirp brings the laser into resonance with additional atoms having lower velocities than the original group around the velocity initially resonant with the laser.

In order to analyze this process, let us consider atoms having positive velocities near some velocity V opposed by a laser beam propagating in the negative direction. We express any atomic velocity as $v = V + v'$. The acceleration of atoms having velocity V $(v' = 0)$ is $a = F(V)/M$, where $F(V)$ is given by eq. (1). Therefore, we write $V(t) = V(0) + at$. Also we let the detuning vary as $\delta(t) = \delta' - kV(t)$. That is, we chirp the laser frequency so as to stay a constant detuning δ' from resonance with atoms having the decelerating velocity $V(t)$. Now we transform to a frame decelerating with $V(t)$. In this frame the atomic velocity is v' and the laser detuning is Doppler shifted to δ'. The force on an atom in this frame, for small v', is

$$F(v') = 2\hbar k^2 \frac{I}{I_0} \frac{(2\delta'/\Gamma)v'}{\left[1 + I/I_0 + (2\delta'/\Gamma)^2\right]^2}. \tag{12}$$

The term multiplying v' is minus the friction coefficient α. When $\delta' < 0$, the force opposes the velocity v' and tends to damp all velocities to zero in the decelerating frame, which is $V(t)$ in the laboratory frame. Maximum damping occurs for $I/I_0 = 2$ and $2\delta'/\Gamma = -1$. The final velocity to which the atoms are decelerated is determined in practice by the final frequency to which the laser is chirped. The result of chirp cooling an atomic beam is that all of the atoms in the initial distribution below the velocity resonant with the laser at the beginning of its chirp are decelerated.

The first definitive experiment showing such chirp cooling is in ref. [40], with deceleration to zero velocity first achieved in ref. [41]. The analysis given above is similar to that given in ref. [42]. The robust character of this sort of cooling is evident. Atoms within a range of velocities around $V(t)$ are damped (in velocity) toward $V(t)$. Lower velocities, not initially close to $V(t)$, come within range as the laser chirp brings $V(t)$ into coincidence with them. If the laser intensity changes during the time an atom is being decelerated (because, for example, the laser beam is not collimated), the atoms will continue to decelerate according to the chosen chirp rate, but with a different effective detuning δ'. The chosen chirp rate, however, must be consistent with the achievable deceleration for the given I/I_0. That is, the chirp rate must satisfy

$$\dot{\delta} = ka = \frac{\hbar k^2 \Gamma}{2M} \frac{I/I_0}{1 + I/I_0 + [2\delta'/\Gamma]^2}. \tag{13}$$

This means that $\dot{\delta}$ has an allowable upper limit of $ka_{\max}$. We have noted that for the velocities to be damped in the decelerating frame we must have $\delta < 0$ and it is easy to show that the conditions for best damping lead to a deceleration half as large as the maximum.

3˙2. *Zeeman tuning.* – In the Zeeman tuning technique, a spatially varying magnetic field is used to keep the frequency of an atom resonant with a counterpropagating laser beam, as the atom slows down by scattering photons from this laser. Because the atoms are slowed down and the spread of the velocity distribution is narrowed, this process is

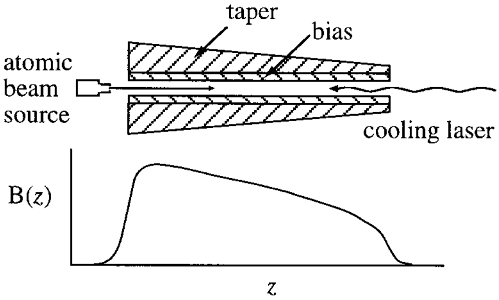


Fig. 2. – Upper: Schematic representation of a Zeeman slower. Lower: Variation of the axial field with position.

often referred to as "Zeeman cooling." Figure 2 illustrates the general idea of this scheme. The atomic beam source directs atoms, which have a range of velocities, along the axis (z-direction) of a tapered solenoid. This magnet has more windings at its entrance end, near the source, so the field is higher at that end. The laser is tuned so that, given the field-induced Zeeman shift and the velocity-induced Doppler shift of the atomic transition frequency, atoms with velocity v_0 are resonant with the laser when they reach the point where the field is maximum. Those atoms then absorb light and begin to slow down. As their velocity changes, their Doppler shift changes, but is compensated by the change in the Zeeman shift as the atoms move to a point where the field is weaker. At this point, atoms with initial velocities slightly lower than v_0 come into resonance and begin to slow down. The process continues with the initially fast atoms decelerating and staying in resonance while initially slower atoms come into resonance and begin to be slowed as they move further down the solenoid. Eventually all the atoms with velocities lower than v_0 are brought to a final velocity that depends on the details of the magnetic field and laser detuning. The magnetic field profile of the tapered solenoid is

$$B(z) = B_0\sqrt{1 - \frac{2az}{v_0^2}} \tag{14}$$

with $0 \le 2az \le v_0^2$. B_0 is the magnetic field producing a Zeeman shift equal to the Doppler shift for atoms with velocity v_0 and $a \le a_{\mathrm{max}}$ is the deceleration rate.

The first experiment on the deceleration of atoms using the Zeeman technique is in ref. [43]. Subsequently, neutral sodium atoms were stopped in ref. [44]. Figure 3 shows the velocity distribution resulting from Zeeman cooling: a large fraction of the initial distribution had been swept down into a narrow final velocity group. The velocity distribution after deceleration was measured in a detection region some distance from the exit end of the solenoid using a separate detection laser. We were able to determine the velocity distribution in the atomic beam by scanning the frequency of the detection laser and observing the fluorescence from atoms having the correct velocity to be resonant.

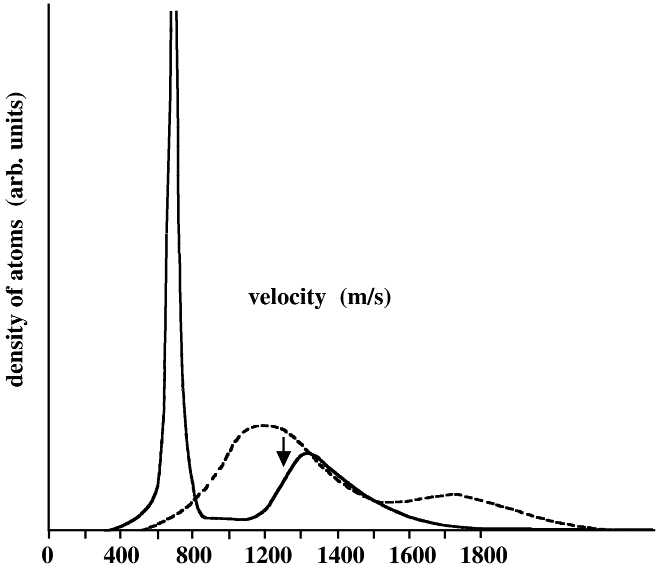


Fig. 3. – Velocity distribution before (dashed) and after (solid) Zeeman cooling. The arrow indicates the highest velocity resonant with the slowing laser. (The extra bump at 1700 m/s is from $F = 1$ atoms, which are optically pumped into $F = 2$ during the cooling process.)

3˙3. *Optical pumping.* – For alkali atoms, whose electronic ground states are split by the hyperfine interaction, deceleration of an atomic beam may be impeded by the optical pumping problem described above in subsect. **2**˙1. For chirp cooling, the usual way of addressing this problem is, as with optical molasses, to use a repumper that excites atoms out of the "wrong" hyperfine state so that they may again be excited by the cooling laser. In Zeeman tuning the large Zeeman shifts from the high magnetic field, the use of a circularly polarized decelerating laser, and the nature of the matrix elements for radiative transitions in the magnetic field, all work together to make the probability of deleterious optical pumping very small, so that a repumper is not usually needed.

4. – Traps for neutral atoms

Trapping of atoms usually refers to their confinement by the application of external fields rather than by the use of a material container. In contrast with the trapping of ions by electric and magnetic fields, the trapping forces that can be applied to neutral atoms are relatively weak. Ions have a charge on which an electromagnetic field can exert a large Coulomb or Lorentz force. Neutral atoms, however, may be acted upon through their permanent magnetic dipole moments or induced electric dipole moments, allowing generally smaller forces to be applied. The strongest traps for neutral atoms have energy depths of only a few kelvin, while ion traps can typically hold room temperature ions, and have trapped particles with energies of a few thousand electron volts.

Another possible difficulty with traps for neutral atoms is that the trapping potentials represent changes in the internal energy of the atoms. That is, the positions of and, in general, the spacings between energy levels are changed by the trapping fields. This means that, in particular, high accuracy spectroscopy of trapped atoms is problematic.

Neutral atoms have the advantage that the lack of space charge effects means that one can generally trap larger numbers and densities of neutral atoms than of ions. Furthermore, some applications demand that one work with neutral atoms (as for example in Bose condensation of an atomic gas). Fortunately, even rather weak forces are capable of trapping atoms that have been laser cooled, and many different kinds of neutral atom traps have been demonstrated. Among these are:

i) Magneto-static traps, first demonstrated in 1985 [45], rely on the force exerted by a gradient magnetic field on the permanent magnetic dipole moment of an atom such as a ground-state alkali.

ii) Laser dipole traps, first demonstrated in 1986 [46], use the dipole force that results from the gradient of the energy of the oscillating dipole moment induced on an atom in an inhomogeneous laser field.

iii) Radiation pressure traps use the scattering force of eq. (1), but are not stable in 3D for two-level atoms and for laser intensities constant in time. The magneto-optical trap (MOT), using multilevel atoms and an inhomogeneous magnetic field, was the first radiation pressure trap to be demonstrated [47].

iv) Magneto-dynamic traps use the micromotion driven by oscillating magnetic field gradients to allow trapping of high-field–seeking states not stably trapped in static magnetic fields. Such a trap, first demonstrated in 1991 [48], is analogous to the radio-frequency Paul trap for ions.

v) Microwave traps are low-frequency, spontaneous-emission-free analogs of laser dipole traps. Such a trap was first demonstrated near a magnetic resonance transition [49]. Similarly, traps based on rf dressed states have recently been realized [50].

vi) Electrostatic traps are similar to laser dipole traps but rely on the interaction between a gradient, DC electric field and the polarization induced in the atom by this field to generate a trapping force. Electrostatic trapping and deceleration of molecules has recently been demonstrated [51].

vii) Gravito-optical traps, which combine optical dipole forces with gravity to produce stable trapping [52], are only one example of hybrid traps that combine different types of forces to achieve trapping of atoms.

viii) TOP traps (for Time-averaged Orbiting Potential) [53] are an important modification of one type of magnetostatic trap. While time dependent, they are not dynamic in the sense of atomic micromotion being essential.

These notes will not treat each of these kinds of traps in detail, nor attempt to give more complete references about them. We shall, however, discuss the radiation pressure, laser dipole and magnetic traps in some more detail below.

4'1. *Dipole force traps.* – The dipole force discussed in subsect. **2**'2 can be used to trap atoms. A single, focused laser beam, tuned below resonance is the simplest dipole trap and was first proposed by Ashkin in 1978 [54]. As an example, consider sodium atoms interacting on their strongest transition ($I_0 = 6\,\mathrm{mW/cm^2}$) with a modest power, 10 mW Gaussian laser beam focused to a $1/e^2$ radius of $10\,\mu$m. This gives $I/I_0 \cong 10^6$ at the focus. For the detuning maximizing U, we find $U_{\max}/k_{\mathrm{B}} \cong 100\,\mathrm{mK}$. For a 1 W beam we find $U_{\max}/k_{\mathrm{B}} \cong 1\,\mathrm{K}$. Such traps can easily confine laser cooled atoms. Gas atoms at temperature above those achieved by laser cooling will not be easily trapped.

The dipole potential is a conservative potential, so it does not have any dissipative mechanism associated with it. It was not until the demonstration of laser cooling in optical molasses [29] that such a trap could be loaded and finally realized [46]. One of the difficulties involved in such trapping is that although the trap is tuned below resonance (the proper sign of the detuning to achieve cooling) the cooling provided by the trapping light does not reduce the thermal energy below the trap depth. Auxiliary cooling [55] is required, but even this is difficult because the inhomogeneous light shifts induced by the trapping laser interfere with the cooling process. Dalibard *et al.* [56, 57] proposed a solution in which the trapping and cooling are alternated in time, and this procedure was used for the first dipole trap [46].

Another difficulty with a single focus dipole trap is that the radiation pressure force pushes the atoms away from the focus, while the dipole force is attracting them to it. While more complicated, counterpropagating beam geometries can avoid this difficulty [54, 58, 59], one can, at the expense of reduced trap depth, solve the problem by detuning the laser [46]. According to eq. (1), the destabilizing radiation pressure force varies as $1/\delta^2$ (for sufficiently large δ), while the dipole trapping force obtained from the dipole potential given by eq. (2) varies as $1/\delta$. Thus, for large enough detuning, the radiation pressure will be negligible compared to the dipole trapping force.

Large detuning has other advantages, particularly when coupled with multi-level laser cooling. When the detuning is large enough that the trap depth is comparable to the natural linewidth, the thermal energy of laser cooled multilevel atoms can still be considerably less than this depth, which is comparable to the Doppler cooling limit of eq. (10). Furthermore, the optimum detuning for good multilevel laser cooling is at least several times the linewidth, the inhomogeneous light shifts should not much affect the cooling. Such a far-off-resonance-trap (FORT) has been demonstrated to work without the need to alternate cooling and trapping phases [60, 61].

Another advantage [62] of such a FORT is that for sufficiently large detuning the population of the trapped atoms is almost entirely in the ground state. The trap is then nearly free of heating due to spontaneous emission [63] and of many of the collisional perturbations involving excited atoms. FORTs have been used to hold atoms for collision experiments [64], to trap atoms for evaporative cooling [65], and to confine Bose-Einstein condensates [66].

Variations of the dipole force trap include using the evanescent wave created by total internal reflection of light (detuned blue of resonance) to act as a mirror to reflect atoms [52, 67], crossing two red-detuned laser beams at their foci to achieve a strong

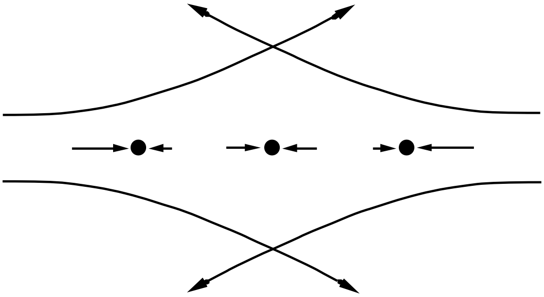


Fig. 4. – A one-dimensional radiation pressure force trap formed from two counterpropagating, focussed laser beams with separated foci. The length of each of the pairs of arrows indicates the magnitude of the force from the respective laser beams on an atom at the indicated positions.

gradient in all directions [65], and crossing sheets of blue-detuned light to form a box bounded by light (with confinement in the vertical direction sometimes provided by gravity) [68, 69].

4˙2. *Radiation pressure force traps.* – Radiation pressure force traps use the spontaneous or scattering force to confine atoms. Unlike the dipole force, the force generated by absorption of a photon followed by spontaneous emission does not depend on the gradient of the intensity. Hence larger volume traps can be created using the scattering force compared to the dipole force for a given flux of incident photons. A simple radiation force trap, shown in fig. 4, can be made in 1D using two counterpropagating focused laser beams with separated foci [54]. Midway between the foci the radiation pressure from the two beams is balanced and the net force on an atom is zero. As the atom moves away from this equilibrium point toward one of the foci it is pushed back by the higher intensity near that focus. In contrast to the dipole force trap, radiation pressure trap depths on the order of a kelvin can be achieved with a modest, near saturation intensity, *i.e.*, a few milliwatts per square centimeter [70]. While the trap of fig. 4 produces radiation pressure force trapping along only one axis, the dipole force can provide trapping along the two orthogonal axes [54], a configuration first demonstrated in ref. [59].

It is tempting to extend the trap idea of fig. 4 to three dimensions, but it can be shown that such an extension is not possible when the radiation pressure force is proportional to the photon flux from the laser beams. The impossibility of such trapping is related to the fact that the divergence of the Poynting vector is zero, and has been called the optical Earnshaw theorem [71] by analogy with the theorem from electrostatics forbidding stable trapping of a test charge in a charge-free region.

In spite of the optical Earnshaw theorem, which applies to 2-level atoms in a weak, static laser field, where the scattering force is proportional to the Poynting vector, it is possible to create a 3D radiation force trap by making use of such features as saturation, multiple levels, optical pumping and Zeeman shifts, which make the scattering force not proportional to the Poynting vector. The most successful radiation pressure trap

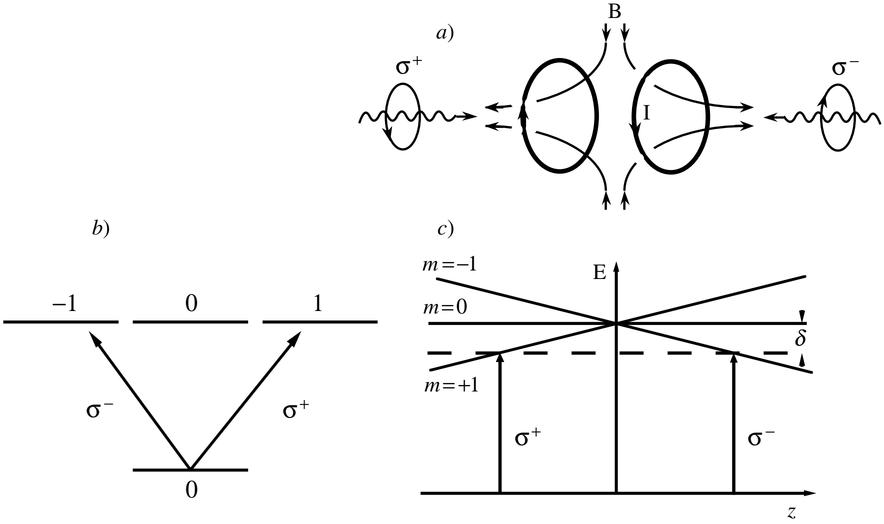


Fig. 5. – *a*) Magnetic field and laser configuration for a 1D MOT. *b*) Transition scheme. *c*) Energy levels and transitions in the spatially varying magnetic field. The designation of the m-state is with respect to a space-fixed axis, as is the laser polarization.

that circumvents the optical Earnshaw theorem is the magneto-optical trap or MOT. Conceived by Dalibard [72] and demonstrated in an MIT-Bell Labs collaboration [47], its principle is illustrated in fig. 5 for 1D operation and a simple $J = 0 \rightarrow J = 1$ transition.

A pair of current-carrying coils with opposing currents creates a quadrupole magnetic field that is zero at the origin and whose vector value is proportional to the displacement from the origin. The simple $J = 0 \rightarrow J = 1$ transition gives us a Zeeman shifted transition frequency with a non-degenerate ground state. (This non-degeneracy leads to Doppler rather than sub-Doppler cooling.) Two circularly polarized laser beams with opposite helicity counterpropagate along the coils' axis. The $\sigma^{\pm}$ beam excites atoms to the $m = \pm 1$ excited states, respectively. Thus, for a red-detuned laser frequency ($\delta < 0$) atoms displaced in the positive direction will experience a Zeeman shift that brings the $m = -1$ state into resonance with the laser frequency, and the σ^- laser beam that excites this state is the one that pushes it back toward the origin. Similarly, an atom displaced in the negative direction is pushed back by the σ^+ beam. In addition to the restoring force, there is also the usual Doppler-cooling damping force.

The force on an atom with velocity v and position z can be obtained from eqs. (3) and (4) by replacing the effective detuning $\delta \mp kv$ with $\delta \mp (kv + \beta z)$, where βz is the magnitude of the Zeeman frequency shift at position z:

$$F(v, z) \cong \frac{4\hbar k I/I_0}{\left[1 + (2\delta/\Gamma)^2\right]^2} \left(\frac{2\delta}{\Gamma}\right)(kv + \beta z), \tag{15}$$

where in the second expression we are in the limit of low velocity, magnetic field and laser intensity. For negative detuning the above force represents a damped harmonic oscillator. Typical operating conditions for a MOT might be $I/I_0 = 1$, $2\delta/\Gamma = -1$ and $\beta = 2\,\pi \times 10\,\mathrm{MHz/cm}$. This would lead to an oscillation frequency of about 1 kHz for Na, but with strong overdamping.

While we have considered only the 1D case for a simple transition, the MOT was first demonstrated in 3D, on an atom with a degenerate ground state (Na). The theory in 3D has been worked out in detail for the $J = 0 \to J = 1$ transition [73]. For transitions allowing sub-Doppler cooling (see sect. **5** below), some insights have been gained by studying the forces in 1D on a moving atom in a magnetic field [74-77]. Experiments have shown that, with a degenerate ground state, sub-Doppler temperatures are achieved in a MOT [76] along with larger trapping and damping than predicted by the $J = 0 \to J = 1$ theory.

The MOT has become an important tool in the study of cold atoms. It is routinely used to capture atoms that have been slowed by chirp cooling or Zeeman tuning. A particularly useful feature of the MOT is that it can capture atoms from an uncooled, thermal atomic vapor [32, 33]. This often allows considerable simplification of the apparatus compared to one where an atomic beam is first decelerated.

A MOT can concentrate atoms to the point that collisions [78, 79] and radiation pressure exerted by the atoms' fluorescence [80] are factors limiting the density. A major advance in reducing such problems is the "dark spot" MOT [81] in which atoms are, for the most part, optically pumped into a state from which the excitation rate is considerably smaller than in a conventional MOT. In practice this is accomplished, for atoms with ground-state hyperfine structure, by pumping the atoms into one of the hyperfine states. Normally laser cooling and trapping of such atoms is performed by applying a separate laser frequency to excite each of the ground hyperfine states to the electronically excited state, ensuring atoms are not pumped into a state from which they cannot be excited. In a dark spot MOT, one of the laser frequencies (the re-pumper) is eliminated from the central region of the trap. Atoms then accumulate in the hyperfine state that would have been excited by the missing frequency. These atoms are rarely excited (only by off-resonant light or indirectly scattered resonant light) so problems due to excited atoms are greatly reduced, even though the atoms are still cooled and trapped. Such techniques have achieved atomic densities near $10^{12}\,\mathrm{cm}^{-3}$.

4˙3. *Magneto-static traps.* – Both laser dipole and radiation pressure traps involve optical excitation of the atoms being trapped. Although in principle a sufficiently intense laser dipole trap can be tuned far enough off resonance for the excitation rate to be small, it is difficult in practice to make it truly negligible. As a result, optical trapping generally results in some heating of the atoms due to the random nature of spontaneous emission, and cooling is required to keep the atoms in equilibrium. Furthermore, whenever the atoms are excited they are much more likely to undergo collisions that will heat them and eject them from the trap. Magnetic traps do not suffer this problem and so can be used to store atoms for long periods of time without the need for additional cooling, and without suffering much collisional loss. On the other hand, such traps only work for atoms having a magnetic dipole moment and only for those states of such atoms whose Zeeman energy increases in increasing magnetic field.

The idea of magnetic trapping was an extension of the magnetic focusing of atomic beams [82] and was discussed by Paul during the 1950s [83]. The first published proposals for magnetic trapping of neutral atoms were in the 1960s [84-86]. It was not until after the demonstration of laser cooling of neutral atoms that the first magnetic trapping of atoms was achieved [45].

The principle of the magnetic trap can be understood by considering an atom with a Zeeman sub-structure such as that shown in the excited state of fig. 5*c*. (Note that magnetic trapping is generally done on ground-state atoms with such Zeeman structure.) Since the energy varies with magnetic field, an atom feels a force in a magnetic field gradient. For states whose energy increases with magnetic field (low-field–seeking states), a trap is formed by the field of the coils in fig. 5*a*. This quadrupole field is zero on the axis midway between the coils, and its magnitude increases linearly along any line away from this central point. Low-field–seeking atoms experience a restoring force towards and are trapped around this point. (It can be shown [87,88] that in a current-free region no magnetic field can have a relative maximum in its magnitude, so that high-field seekers cannot be trapped.) The depth of a magnetic trap is given by the maximum Zeeman shift along the easiest escape path. A magnetic field of 1 mT (10 G) gives a typical shift of 14 MHz, equivalent to 670 μK, so laser-cooled atoms are easily trapped by modest fields.

The quadrupole magnetic field was used in the first magnetic trap [45] and is the simplest of magnetic traps. It works for low-field seekers as long as they do not change their spin orientation with respect to the local field. This means than they must adiabatically follow the changes in the direction of the field as they orbit in the trap. The condition for adiabatic following is that

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} \ll \omega_{\mathrm{Zeeman}}, \tag{16}$$

where θ is the angle of the magnetic field at the atom's position and ω_{Zeeman} is the Zeeman frequency separating states of different spin orientation. Since laser-cooled atoms move so slowly, they generally experience small field rotation rates, and their motion is usually adiabatic. The quadrupole trap, however, has a point where the magnetic field is zero and the adiabatic condition is impossible to satisfy. Atoms passing sufficiently close to the trap center will fail to follow adiabatically, change their spin orientation (undergo Majorana transitions) and be ejected from the trap [45]. This difficulty can be avoided by a variety of traps that do not have a point of zero magnetic field [89], but such traps are generally not as "stiff" as a quadrupole trap. That is, the restoring force near the center of the trap is not as large as for a quadrupole trap. In practice, non-adiabaticity is usually a problem only for very cold atoms confined very near the center of a quadrupole trap (which is exactly what occurs on the way to BEC.)

In order to achieve Bose-Einstein condensation through runaway evaporative cooling, it is desirable to have a "stiff" trap to maintain a high collision or thermalization rate. For atoms cold enough to Bose condense, the problem of non-adiabatic spin flip transitions at the zero-field point of a quadrupole trap is quite severe [90, 91]. The TOP trap provided one solution to this problem by superposing onto the quadrupole field a magnetic field rotating in the plane of symmetry. The rotating field guarantees that the field is not zero at the trap center, and produces a rounding of the sharp-cusp, quadrupole, trapping potential. This time-averaged, orbiting potential (TOP) trap [53], although time-dependent, is not a magneto-dynamic trap in the sense that the TOP trap does not rely on the micromotion of the atoms. The atoms respond to the time-averaged potential at a given point in the trap.

5. – Sub-Doppler laser cooling

The early experiments [29, 92] on optical molasses produced a satisfying agreement of the observed temperature and spatial diffusion with the predictions of the theory of Doppler cooling as outlined in subsect. **2**˙3. Some experiments, however, were in disagreement with expectations [93]. Finally, in 1988, careful temperature measurements [94] showed conclusively that atoms in optical molasses were much colder than the Doppler cooling limit. The time-of flight (TOF) method used to measure the temperature in those experiments has become a standard technique. Atoms are collected and cooled in the optical molasses at the intersection of the molasses laser beams. The molasses beams are suddenly extinguished and the released atoms fall toward a probe. As the atoms pass through the probe laser beam they absorb and fluoresce light. This fluorescence is measured, with time resolution, by a photodetector. The distribution of detected fluorescence in time gives the distribution of times of flight from the molasses to the probe, and the temperature can be determined from that distribution. With the ultralow temperatures and high densities of or near Bose-Einstein condensation, researchers often use the direct measurement of the spatial expansion of a cloud of atoms (typically obtained by imaging the spatial profile of the absorption of a resonant or near-resonant laser beam) to determine the temperature.

5˙1. *Observation of sub-Doppler temperatures.* – Figure 6*a* shows an example of an experimental TOF signal for sodium atoms cooled in optical molasses [21]. The experimental points correspond reasonably well with the predicted signal for a temperature of $25\,\mu$K, while $250\,\mu$K, about the Doppler cooling limit for Na, is completely inconsistent with the experimental points. Furthermore, the dependence of the temperature on detuning was found to be inconsistent with the theory of Doppler cooling. Figure 6*b* (points) shows the measured temperature as a function of laser detuning for Na, where the linewidth is 10 MHz. The temperature decreases for detunings larger than $\Gamma/2$ (until the laser frequency approaches resonance with another hyperfine state, about 60 MHz to the red of the chosen resonance). This is in sharp contrast to the prediction of Doppler cooling theory (fig. 6*b*, solid line), which has the temperature increasing for detunings greater than $\Gamma/2$. In addition, the temperature was found to be linearly dependent on laser intensity [95], again in contrast to the predictions of Doppler cooling theory, and

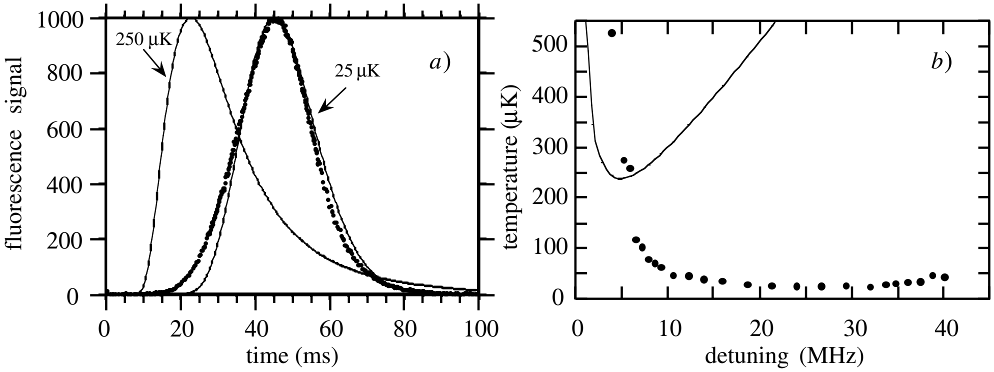


Fig. 6. – *a*) Experimental time-of-flight distribution (points) for Na atoms released from an optical molasses. Expected TOF distributions for temperatures of 25 μK and for 250 μK (approximately the Doppler cooling limit) are shown. *b*) Measured temperature as a function of laser detuning for sodium atoms cooled by a 3D optical molasses. The temperature predicted by Doppler cooling theory is shown by the solid line.

the temperature was found to depend on magnetic field and laser polarization [94, 95]. These latter facts, in particular, suggest that the magnetic sublevels of the atom play an important role.

The observation of sub-Doppler-limit temperatures was quite surprising. Doppler cooling theory at low intensity was simple and compelling. Furthermore, at least for 1D, there was a complete theory, taking into account the effects of high intensity and interference between laser beams that were ignored in the treatment presented in the previous section. The theory had been restricted to 2-level atoms, but it was widely believed that this restriction was not particularly important. At low intensity the Doppler temperature depended on the transition linewidth and the detuning, and these were the same for any of the degenerate Zeeman sublevels in a given alkali hyperfine level. Nevertheless, the magnetic field and polarization dependence of the sub-Doppler-limit temperatures pointed to the importance of the Zeeman sublevels.

5˙2. *New cooling mechanisms.* – The explanation for the sub-Doppler temperatures soon came in the form of a new theory of multilevel laser cooling. The key elements of the new theory [74, 96, 97] were optical pumping among the magnetic sublevels of the electronic ground level and differential light shifts of the sublevels. While the original theories treated cases where a spatial gradient of the polarization of the optical field was important, it was later demonstrated that such multilevel laser cooling was possible even without polarization gradients [97-103].

Since the theory of multilevel laser cooling is treated extensively in the literature a detailed description will not be given here. The most important of the multilevel cooling mechanisms is Sisyphus cooling [19, 74, 104, 105]. Semiclassically (when the atom can be considered to be well-localized on the scale of an optical wavelength) the simplified

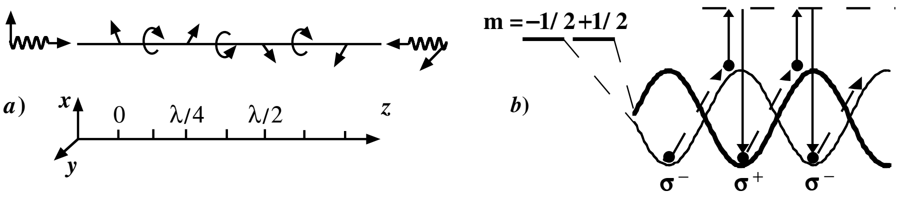


Fig. 7. – *a*) Interfering, counterpropagating beams having orthogonal, linear polarizations create a polarization gradient. *b*) The different Zeeman sublevels are shifted differently in light fields with different polarizations; optical pumping tends to put atomic population on the lowest energy level, but non-adiabatic motion results in "Sisyphus" cooling.

physical picture for Sisyphus cooling can be understood by considering the atom and laser field situation illustrated in fig. 7.

Figure 7*a* shows a 1D set of counterpropagating beams with equal intensity and orthogonal, linear polarizations. The interference of these beams produces a standing wave whose polarization varies on a sub-wavelength distance scale. At points in space where the linear polarizations of the two beams are in phase with each other, the resultant polarization is linear, with an axis that bisects the polarization axes of the two individual beams. Where the phases are in quadrature, the resultant polarization is circular and at other places the polarization is elliptical. An atom in such a standing wave experiences a fortunate combination of light shifts and optical pumping processes.

Because of the differing Clebsch-Gordan coefficients governing the strength of coupling between the various ground and excited sublevels of the atom, the light shifts of the different sublevels are different, and they change with polarization (and therefore with position). Figure 7*b* shows the sinusoidal variation of the ground-state energy levels (reflecting the varying light shifts or dipole forces) of a hypothetical $J_{\rm g} = 1/2 \rightarrow J_{\rm e} = 3/2$ atomic system. Now imagine an atom to be at rest at a place where the polarization is σ^- at $z = \lambda/8$ in fig. 7*a*. As the atom absorbs light with negative angular momentum and radiates back to the ground states, it will eventually be optically pumped into the $m_{\rm g} = -1/2$ ground state, and simply cycle between this state and the excited $m_{\rm e} = -3/2$ state. For low enough intensity and large enough detuning we can ignore the time the atom spends in the excited state and consider only the motion of the atom on the ground state potential. In the $m_{\rm g} = -1/2$ state, the atom is in the lower energy level at $z = \lambda/8$, as shown in fig. 7*b*. As the atom moves, it climbs the potential hill of the $m_{\rm g} = -1/2$ state, but as it nears the top of the hill at $z = 3\lambda/8$, the polarization of the light becomes σ^+ and the optical pumping process tends to excite the atom in such a way that it decays to the $m_{\rm g} = +1/2$ state. In the $m_{\rm g} = +1/2$ state, the atom is now again at the bottom of a hill, and it again must climb, losing kinetic energy, as it moves. The continual climbing of hills recalls the Greek myth of Sisyphus, so this process, by which the atom rapidly slows down while passing through the polarization gradient, is called Sisyphus cooling. In Sisyphus cooling, the radiated photons, in comparison with the absorbed photons, have an excess energy equal to the light shift, while in Doppler cooling, the energy excess comes from the Doppler shift.

In contrast to the case for Doppler cooling (see eq. (4)) the friction force is independent of laser intensity, and proportional to detuning at low (but not too low) intensity and low velocity, while the momentum diffusion coefficient is proportional to intensity and independent of detuning. This leads, according to eq. (7), to a temperature that depends linearly on intensity and inversely on detuning. That is, the temperature is proportional to the light shift

$$k_{\mathrm{B}}T \propto \hbar\Delta_{\mathrm{light\,shift}} = \frac{\hbar\Omega^2}{4\delta}, \tag{17}$$

where the expression for the light shift is valid in the limit of low intensity and large detuning.

A less restricted treatment [106] shows that the friction force is not linear in the atomic velocity, nor is the momentum diffusion constant independent of velocity. Nevertheless, the temperature remains approximately linear in the light shift as long as the intensity is sufficiently above a critical intensity. The lower limit to the temperature obtainable by Sisyphus cooling is set by this lowest intensity for which the cooling process works [106, 107]. This is the intensity at which the light shift is comparable to the recoil energy, and it leads to a lower limit for the thermal velocity on the order of a few times the recoil velocity.

These qualitative features of multi-level laser cooling have been confirmed by experiments on atoms cooled in 3D optical molasses [21, 99, 108]. The experimental results showed the linear dependence of the temperature on intensity and light shift for all but the largest intensity at the smallest detuning, outside the limits of validity of the simple results listed above. The constancy of the lowest temperature, once the detuning is large enough, is consistent with its depending only on the recoil energy. The temperature of 2.5 μK obtained for Cs represents an r.m.s. velocity of about three recoil velocities, similar to the case for the lowest temperatures observed for Na [21, 99] and Rb [109].

6. – Metrology with cold atoms

Since laser cooled atom move much more slowly than ordinary, thermal atoms, metrology with such atoms is less subject to errors and uncertainties associated with the motion. These include the first-order Doppler shift of atomic transition frequencies and the second-order Doppler shift or relativistic time dilation shift. Furthermore, the slow velocity means that atoms remain longer in the apparatus, allowing observation of narrower linewidths for observation-time–limited features.

The wave nature of atoms become more evident as the atoms become slower because the deBroglie wavelength is inversely proportional to the velocity. This can make atom interferometry easier with cold atoms, and atom interferometers can be used for measurements, as optical interferometers are often used in metrology.

This section briefly describes the use of laser-cooled atoms in atomic clocks and in metrology applications of atom interferometers. General background on cold atom clocks can be found in ref. [110], and on atom interferometers in ref. [11].

6˙1. *Atomic fountain clocks.* – Primary frequency standards, the devices by which the SI definition of the second is realized, have traditionally been cesium atomic beams using the Ramsey method of separated oscillatory fields [82]. In this method the atomic beam is first state-selected so that the atoms are in one (let us say, the lower) of the two hyperfine states (clock states) whose frequency difference (the clock frequency) defines the unit of time. The atoms then enter a microwave cavity where a microwave field tuned close to the clock frequency induces coherent transitions in the atoms. The strength of the field and the duration of the passage of the atoms through the cavity are adjusted so that the atoms receive approximately a $\pi/2$-pulse, putting the atoms into a coherent superposition of the two hyperfine states.

After the atoms exit the cavity, they proceed in a nearly field-free region for a time T to a second, similar cavity, where they receive another $\pi/2$-pulse, while in the field-free region, the atoms, being in a superposition of the two hyperfine states, evolve at the clock frequency. When they encounter the second cavity, with its second oscillatory microwave field, the action of that field depends on the difference in the phase evolution of the atoms and the phase evolution of the microwaves. When the microwave and clock frequencies are identical, the second pulse completes the transition begun by the first pulse, and the atoms are all in the upper hyperfine state. If the frequencies are different enough so that the atom and field are out of phase by π, then the second pulse puts the atoms back into the lower hyperfine state. The population of the lower (upper) state varies sinusoidally (cosinusoidally) with the phase difference. The upper state population varies from maximum to minimum for a difference of a half-cycle between the atom and microwave frequencies over the time required for the atoms to travel between the two cavities. This is also the frequency difference corresponding to the full width at half-maximum (FWHM) of the feature corresponding to the hyperfine resonance. Thus, the FWHM in hertz is given by $\Delta\nu = 1/2T$. For cesium atoms traveling at about 200 m/s over a distance of 1 m between the cavities, this means a FWHM of about 100 Hz, a rather narrow line, but one whose center must be determined to a part in 10^{-6} in order to achieve the 10^{-14} accuracy that is typical of the best laboratories' primary frequency standards.

Slower (colder) atoms would allow a longer Ramsey time T, a narrower FWHM, and with it potentially better accuracy. The way in which such colder atoms could be used to make a better atomic clock was suggested by Zacharias in the early 1950s [82]. In this scheme, called an atomic fountain clock, state selected atoms are launched vertically through a microwave cavity, where they experience a $\pi/2$-pulse as in the first passage through a microwave cavity in the traditional beam clock. The atoms continue along a vertical trajectory, falling back through the cavity, where they experience the second $\pi/2$-pulse. For a trajectory height of 1.2 m above the cavity, the time between cavity passages is about 1 s, an improvement of more than two orders of magnitude over a typical laboratory primary reference standard. Furthermore, the fact that the second passage is with the same cavity as the first passage eliminates some systematic errors associated with having two different cavities in the traditional design.

The first realization of an atomic fountain clock of the Zacharias design [111] had only a 5 cm fountain height, which still gave a 200 ms Ramsey time, significantly longer than any beam clock. (An earlier atomic fountain design [112] did not have the atoms passing through the cavity and falling back, but instead exposed the atoms to two pulses of microwaves while they remained in the cavity near the top of their trajectory.) Today atomic fountains at the LPTF in Paris [113], PTB in Germany [114] and at NIST in the USA [115] with significantly larger heights achieve accuracy on the order of 10^{-15}.

Achieving temperatures below the Doppler cooling limit was an important feature in the success of cesium fountain clocks. If the temperature of laser-cooled cesium atoms had been as high as the $125\,\mu$K predicted by Doppler cooling theory, a cloud of cesium atoms in a meter-high fountain would spread too much during the fountain time that less than 1% of the launched atoms would return through the microwave cavity for the second passage. (The opening in the 9.2 GHz microwave cavity, which allows the passage of the atoms, cannot be any larger than about 1 cm without causing significant distortion of the microwave field.) But even at temperatures as low as $1\,\mu$K most of the launched atoms do not return for detection. One might suppose that a modest loss of atoms due to the thermal spreading at $1\,\mu$K could be compensated by simply launching a larger number of atoms. Unfortunately, collisional shifts of the clock frequency are large enough [116] that such a procedure would introduce significant difficulties. A cesium density of 10^6 cm^{-3} at $1\,\mu$K causes a fractional shift on the order of 10^{-15} [117].

The collisional shift implies that is would be advantageous to further lower at least the transverse velocity spread of the launched atoms, so as to insure that more of the launched atoms are detected. One proposed method for achieving temperatures lower than those usually achieved by laser cooling is Raman cooling [118, 119]. If we consider the possibility of further improving the clock performance by increasing the Ramsey time, such cooling becomes even more attractive. We cannot, however, increase the Ramsey time significantly in an earth-bound atomic fountain clock. A Ramsey time of even 10 s would require a fountain 120 m high, a daunting engineering prospect that would also make shielding of magnetic fields quite difficult. For this reason, a number of groups are considering the use of the "microgravity" environment of earth orbiting satellites [120, 121].

6˙2. *Atom interferometers.* – Optical interferometers have long been used for metrology. Indeed, practical length metrology at the highest levels of precision is almost exclusively accomplished by laser interferometry. Over the past two decades atom interferometry, where the wave nature of atoms is used in a way analogous to that of light, has developed rapidly to the point where it too is an important metrological tool. In many respects, this development has been driven or encouraged by the developments in laser cooling and trapping of atoms. Two important reasons are that the longer deBroglie wavelengths associated with cold atoms make their interference effects more accessible and that the smaller velocity spreads available with laser cooling increase the deBroglie wave coherence of the atoms.

In many respects the traditional Ramsey-resonance atomic clock, and the atomic fountain clock that is based on it, may be thought of as atom interferometers: the first interaction with the microwaves splits the atoms, originally state-selected in one quantum state, into two components, the two "clock-state" hyperfine energy levels. The second interaction splits or projects each of these states into the same two states. The components of each state interfere so as to produce the Ramsey pattern, often called the Ramsey interference pattern. This picture provides an analogy with optical Michelson or Mach-Zehnder interferometers where a first interaction (with a beamsplitter) divides the light and a second interaction (with the same or another beamsplitter) divides each of these in such a way that they interfere with each other at the output.

One clear difference between the Ramsey interferometer and those optical interferometers is that in the Ramsey case the atoms are not physically separated while in the optical case they are. The Ramsey interference is (at least in the traditional description) between internal energy states of the atoms and does not involve the center-of-mass motion, deBroglie-wave character of the atoms. This apparent difference disappears when we introduce photon momentum transfer into the interactions.

Consider the interferometer with the following configuration: An atom, initially at rest and in internal state $|1\rangle$, is split by a $\pi/2$ pulse into a superposition of $|1\rangle$ and $|2\rangle$. The pulse delivers the photon momentum to that component of the wave function transferred to state $|2\rangle$, which then separates from state $|1\rangle$. A π pulse, applied after a time T, transforms $|1\rangle$ into $|2\rangle$ and vice versa. The photon momentum associated with the transition stops the motion of the portion of the wave function that had been in $|2\rangle$ as it transforms to $|1\rangle$, and gives momentum to the state that had been in $|1\rangle$ as it transforms to $|2\rangle$. Finally, after an additional time T, when the two paths of the atoms' wave function physically overlap, a $\pi/2$ pulse projects part of the amplitude of each component of the wave function into final states $|1\rangle$ and $|2\rangle$ with their appropriate momenta. This example is a time-domain, atomic analog of a Mach-Zehnder interferometer where the $\pi/2$ pulses act as beamsplitters and the π pulses act as mirrors. An example of a Mach-Zehnder type, atom interferometer based on stimulated Raman transitions is shown in fig. 13 of subsect. **8**˙4.

Steve Chu and his colleagues at Stanford [122] have used this kind of interferometer for precision measurement of the gravitational acceleration of atoms in an atomic fountain geometry. The π and $\pi/2$ pulses in those experiments induce Raman transitions between hyperfine states, using two laser beams whose frequencies differ by the hyperfine separation. The laser beams are counterpragating, so that the momentum of two optical photons is transferred to the atom when the hyperfine transition is induced. Because of the different trajectories followed by the falling atoms in the two arms of the interferometer, they experience different phase shifts, with the difference depending on the gravitational acceleration and the square of the time (in the same way as does the distance a body falls due to gravity). In the most recent reports, these experiments have a resolution better than 10^{-10} and an accuracy on the order of parts in 10^9. These results, whose accuracy is comparable to the best measurements with falling corner cubes, are expected to improve with further experiments. Experiments using somewhat different interferometer configurations have been used at Stanford to measure h/M [122], and at MIT [123] and Yale [124] to measure absolute rotations. Bloch oscillations of laser-cooled atoms in an optical standing wave have been used for a new determination of α, the fine-structure constant [125]. While not yet competitive, α has also been measured using atom interferometry with a BEC [126].

7. – Coherent manipulation of Bose-Einstein condensates with light

The creation of Bose-Einstein condensates (BECs) in dilute atomic vapors of Rb [127], Na [128] and Li [129] is one of the major triumphs of laser cooling and trapping of neutral atoms. Furthermore, the creation of BECs has renewed interest in the applications of laser cooling and trapping techniques for atom optics, the manipulation of atoms with mirrors, beamsplitters and lenses analogous to the manipulation of light.

Until recently, atom optic experiments have used thermal sources of atoms much as early experiments in optics used lamps. What was lacking was a coherent source of matter-waves similar to the laser for light. The creation of a Bose-Einstein condensate (BEC) of a dilute atomic gas has opened up the possibility of realizing a matter-wave source analogous to the optical laser. The macroscopic occupation of the ground state of a trap by a BEC is similar to the occupation of a single mode of an optical cavity by photons. The atoms forming the condensate all occupy the same wave function—both in terms of their internal and external degrees of freedom.

Atoms from a BEC are nearly the ideal, monochromatic source for atom optics. Many atom optical elements involve the interaction of the atoms with an optical field and the associated transfer of the photon momentum to the atoms. Because of the repulsive atom-atom interaction, which can be described by a mean field, the BEC swells to a size significantly larger than the ground-state wave function of the harmonic trap confining the atoms [130]. The spatial extent of the resulting wave function can be several orders of magnitude larger than the optical wavelength. Hence the momentum width, given by the Heisenberg uncertainty principle, can be much less than the photons momentum. Not all experiments will realize this reduced, intrinsic momentum width. The interaction energy may be converted to kinetic energy when the atoms are released from the trap. Nonetheless, the resulting additional momentum spread, due to the atom-atom interaction, can still be significantly less than the momentum of a single photon.

A principal focus of research on Bose-Einstein condensates in the Laser Cooling and Trapping Group of NIST is in atom or matter-wave optics. This includes optically induced diffraction as beamsplitters and mirrors for the condensate atoms, as well as applications of diffraction in atom interferometry and the realization of an "atom laser", the atomic analogue of the optical laser. These experiments are basically demonstrations of "single atom" phenomena applied to a collection of atoms with a high degree of first-order coherence. When the interactions between the atoms is considered, collective matter-wave phenomena can be observed which are analogous to effects in nonlinear optics with light. We have observed such effects with matter-waves and they are described in the subsequent section titled "Nonlinear atom optics with Bose-Einstein condensates".

Our atom optics experiments are primarily performed on a BEC of sodium atoms. We begin with a brief description of our experimental apparatus and our approach for creating a Bose-Einstein condensate. We do this essentially for two reasons: our strategy for achieving BEC is somewhat different than other approaches, and it illustrates an application of many of the techniques of laser cooling and trapping developed over the last 15 years.

As of this writing, only a couple of experiments [126] have been attempted using Bose-Einstein condensates for high-precision metrology; however, these experiments are not yet competitive with laser-cooled atoms, primarily because of systematic errors due to atom-atom interactions. Nevertheless, the properties of condensates as extremely low-temperature samples of atoms and as coherent sources of deBroglie wave for atom interferometers suggests that metrological applications will be realized. For example, a BEC, adiabatically expanded to a 1 cm diameter in a microgravity environment, could reach an "effective" temperature on the order of a picokelvin, with atomic velocities on the order of micrometers per second. Such a sample could be used in an atomic clock with a Ramsey time on the order of 1000 seconds, three orders of magnitude longer than the best earthbound atomic clocks [120, 121]. Similarly, a BEC could be used in a space-borne atom interferometer where the extremely small spreading of the condensate could enable long observation times and high precision.

7˙1. *The triaxial* TOP *trap for sodium.* – Our Bose-Einstein condensates of sodium atoms are produced in a time-averaged orbiting potential or TOP trap [53]. Our TOP trap differs from those in other BEC experiments in two respects. First, all other experiments with TOP traps that have resulted in BEC use rubidium. We are currently the only group making condensates of sodium in a TOP trap. The lighter mass of sodium poses a greater technical challenge in the design of the TOP trap compared to rubidium. Since the oscillation frequency of an atom in a trap is higher for the lighter atom, the frequency of the rotating bias field of the TOP trap must be correspondingly higher so that the atom experiences the time-averaged potential. Typically, sub-Doppler laser cooling can cool a sample of atoms to an energy which is a few times the recoil energy. Since the recoil energy is inversely proportional to the mass of the atom, the TOP trap must also be deeper or stronger to contain the lower-mass, laser-cooled atoms. Second, the geometry of our TOP trap is different from other TOP traps resulting in a totally anisotropic or triaxial, time-averaged potential.

A time orbiting potential or TOP trap is a magnetic trap consisting of a quadrupole magnetic field and a constant magnitude rotating bias field. The potential resulting from a superposition of these two magnetic fields, averaged over a rotation period, is harmonic for small displacements. The time-averaged value of the minimum magnetic field in the TOP trap is just the magnitude of the rotating bias field. That is, the rotating bias field has effectively "plugged" the zero-field region of the quadrupole field. The zero-field point of the quadrupole has been displaced by the bias field and, in fact, rotates with the rotating bias field, producing the so-called "circle of death".

In the NIST-Gaithersburg TOP trap, the bias field rotates in a plane containing the symmetry axis of the quadrupole field. For small displacements, the time-averaged magnetic field is

$$\langle \mathbf{B} \rangle_t = B_0 + \frac{b_\mathrm{q}^2}{4B_0}\left(x^2 + 2y^2 + 4z^2\right), \tag{18}$$

where B_0 is the magnitude of the rotating bias field and $2b_\mathrm{q}$ is the gradient of the quadrupole field along the symmetry (z) axis. The spring constants for this TOP magnetic trap are in the ratio of $1:2:4$ in the x, y and z direction, respectively. In the standard configuration for the TOP trap fields, such as in the original trap of JILA [53], used to create the first BEC of Rb, the bias field rotates in the symmetry plane of the quadrupole field, and the spring constant in the radial (x,y) direction is a factor of eight less than the axial direction (z) producing a disk-shaped time-averaged potential. The magnetic trap given by eq. (18) is triaxial; that is, unlike the JILA TOP trap, it has no rotational symmetry. In addition, it is closer to spherical than the JILA TOP trap and better matched for loading from the nearly spherical clouds of laser-cooled atoms from the MOT.

Although the TOP trap has a number of desirable properties, including some independent adjustments of the spring constants in the three principle directions, there are certain limitations associated with trapping a mixed state such as the $F=1$, $m_F=-1$ state of sodium. The most important is the quadratic Zeeman effect which reduces the "effective" magnetic moment of that state.

Equation (18) shows that in order to achieve the stiffest TOP trap for a given radius of the "circle of death", the distance where the magnitude of the quadrupole field equals the bias field (a distance typically chosen to be larger than the radius of the sample of trapped atoms), the largest possible quadrupole field should be used. If the strength of the quadrupole field is increased, then the magnitude of the rotating field must be increased to keep the radius of the "circle of death" constant. The energy of the $F=1$, $m_F=-1$ state of sodium as a function of magnetic field initially increases linearly (the linear Zeeman effect), but the slope decreases with magnetic field due to the quadratic Zeeman effect. Eventually the energy of this state reaches a maximum around 31.5 mT (315 Gauss) above which it becomes an anti-trapped state. Thus for a fixed radius of the "circle of death", the stiffness of the TOP trap no longer increases linearly with the strength of the quadrupole field, becoming extremely weak for sufficiently large values of the bias field.

7ˑ2. BEC *of sodium in a* TOP *trap*. – Similar to other BEC experiments, our Bose-Einstein condensate of alkali atoms is produced by evaporative cooling in a magnetic trap loaded with laser-cooled and trapped atoms, however, the details of our technique differ from other groups. More specifically, we load a dark spot MOT [81] from an effusive source of sodium at 625 K using Zeeman slowing. The slowing laser beam passes through the trap so that the capture area of the trap subtends a large solid angle of the flux of slow atoms. In order to minimize the effect of the slowing laser on the trapped atoms, we use a hybrid slower geometry [131]. A conventional Zeeman slower is used to slow atoms from about 800 m/s down to about 160 m/s, followed by a short section of reverse slower [132] to slow atoms from about 160 m/s down to a few m/s, which is within the capture velocity of the dark spot MOT. To avoid the loss of atoms in the slowing process due to optical pumping as the atoms pass through the zero-field region between the conventional and reverse Zeeman slowing magnets, a second laser frequency is in the slowing beam to pump atoms from the $F = 1$ hyperfine level to the $F = 2$ level where they can continue to be slowed. In addition, a dark spot is placed in the slowing laser beam, creating a shadow in the region of the cloud of atoms in the dark spot MOT [133]. This further reduces the effect of the slowing laser beam on the trapped atoms. Typically, we load more than 10^{10} atoms into the MOT in less than 0.5 seconds.

After loading into the dark spot MOT, the magnetic fields are rapidly switched off and the atoms are further cooled to $\approx 200\,\mu$K by a brief period of dark molasses (there is a dark spot in the repumper light in the molasses beams), followed by optical pumping of the entire population into the $F = 1$ hyperfine level. The magnetic trap is then rapidly switched on, trapping atoms in the $m_F = -1$ sublevel of the $F = 1$ ground state. Since all three sublevels are present in equal populations at the time when the magnetic trap is turned on, two-thirds of the sample of laser-cooled atoms are necessarily lost in the transfer. Experimentally, we find that we are able to trap 4 to 5 $\times 10^9$ sodium atoms in the magnetic trap.

The atoms are confined in the benign environment of a magnetic trap in order to be evaporatively cooled [134, 135]. We have developed two strategies for evaporatively cooling atoms to Bose-Einstein condensation. The first strategy evolved in response to the quadratic Zeeman effect problem in the $m_F = -1$ state of sodium, discussed in subsect. **7**·1. That is, for the large cloud of atoms initially confined in the TOP trap, the stiffness of the trap could not be increased, while keeping the "circle of death" well outside the cloud, to sufficiently compress the sample for runaway evaporation. Instead, we initially compress and evaporatively cool the sample of atoms trapped in a quadrupole magnetic field. When the sample is sufficiently cold and dense enough, but before there is significant loss at the zero-field region due to Majorana transitions, we transfer them to the TOP trap by rapidly switching on the rotating bias field while the quadrupole field is on, after which further evaporative cooling can proceed in the TOP trap. Evaporative cooling in the TOP trap can be achieved by removing the higher-energy atoms to untrapped states with either rf-induced transitions or the circle of death. For most experiments we use rf-induced transitions because it allows us greater control and flexibility. By appropriate choice of frequency and power, we can control the final energy of the atoms we are removing as well as the width of cut made into the sample of atoms. In addition, the parameters of the rf can be changed rapidly compared to changing values of magnetic fields for the circle of death. This technique works for a large range of initial numbers of trapped atoms.

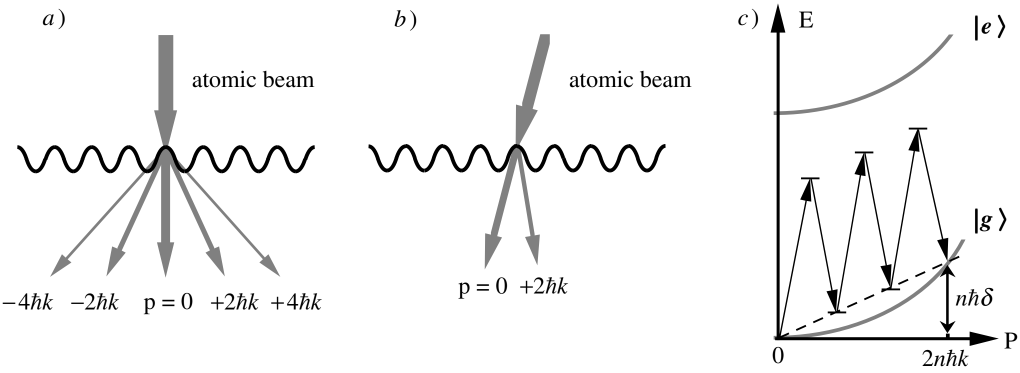


Fig. 8. – *a*) Normal incidence diffraction. *b*) Bragg diffraction. *c*) Bragg diffraction as a $2n$-photon Raman transition.

When the initial number of trapped atoms in the MOT is $\geq 10^9$, we can achieve BEC with atoms in the TOP trap, directly. We initially lose a large number of atoms after transfer into the TOP trap because the circle of death cannot be placed sufficiently outside the cloud of atoms. After this initial loss and the cloud of atoms has rethermalized to a lower temperature and higher density, runaway evaporation can be achieved by compressing the sample and removing high-energy atoms with the circle of death. Both strategies produce approximately the same number of final condensate atoms, about 3×10^6, at a BEC transition temperature of $1.2\,\mu$K.

We probe our samples of atoms using the technique of absorption imaging [127]. The TOP trap is rapidly shut off and after a variable delay, a short laser pulse optically pumps the atoms from $F = 1$ to $F = 2$, after which another short laser pulse resonant with the $3S_{1/2}$, $F = 2 \rightarrow 3P_{3/2}$, $F' = 3$ transition is applied to the atoms along the direction of gravity (the x-direction). The light absorbed from this laser beam is imaged onto a CCD camera. From this image we extract the transverse spatial dependence of the optical depth along the direction of the probe beam.

7·3. *Diffraction of atoms by a standing wave.* – When an atomic beam passes through a periodic optical potential formed by a standing light wave, it diffracts similar to the diffraction of light by a periodic grating. The diffraction can be divided into two regimes, normal and Bragg diffraction. Both diffraction processes can be thought of as arising from the simultaneous absorption of a photon from one laser beam of the optical standing wave, and stimulated emission of a photon due to the counterpropagating laser beam. (This is a similar picture for the origin of the optical dipole force and the momentum transfer resulting from diffraction can be thought of as arising from this force.) This necessarily means that the momentum transfer to the atomic beam by the optical standing wave is quantized in units of $2\hbar k$, twice the momentum associated with a single photon.

In normal diffraction illustrated in fig. 8*a*, the incident atomic beam non-adiabatically enters the light field at normal incidence. As there is no difference in frequency between the two laser beams comprising the standing wave, the exiting atomic beam is symmetri-

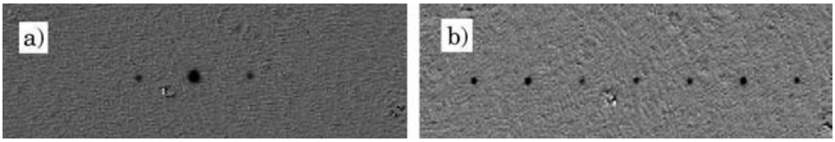


Fig. 9. – Diffraction of a BEC by a short pulse, optical standing wave. *a*) For low intensities only first-order diffraction into $\pm 2\hbar k$ momentum states is visible. *b*) At higher intensities, higher-order diffraction ($\pm 4\hbar k$ and $\pm 6\hbar k$) is observed.

cally diffracted with respect to the incident atomic beam. Energy conservation is satisfied by the spread in energies associated with the non-adiabatic "turn-on" and "turn-off" of the standing wave. For short interaction times such that the atoms do not move appreciably along the direction of the standing wave, the standing wave potential can be considered a thin phase grating that modifies the atomic deBroglie wave with a phase modulation, which for a square profile laser beam is given by $\phi(x) = (U_0\tau/\hbar)\cos^2(kx)$, where U_0 is the maximum depth of the optical potential given by eq. (2) and τ is the interaction time of the atomic beam with the standing wave. An atom with zero momentum is therefore split by the standing wave into multiple components with transverse momenta $p_n = 2n\hbar k$, $(n = 0, \pm 1, \pm 2, \ldots)$, with populations $P_n = J_n^2(U_0\tau/2\hbar)$, where $J_n(x)$ are Bessel functions of the first kind. Normal incidence diffraction of atoms by a near resonant optical standing wave was first demonstrated in Pritchard's group at MIT [136] in 1983.

In the early experiments demonstrating diffraction of atoms by optical standing waves [136, 137], a beam of atoms passed through an optical standing wave and the diffracted beam was detected downstream. In our experiments [138], we start with BEC at a temperature sufficiently below the transition temperature such that no discernible thermal fraction is present. We then adiabatically reduce the strength of the confining potential, which lowers the energy of the condensate by both reducing the mean-field interactions and increasing the size of the condensate wave function. We then expose the atoms to a short pulse of the optical standing wave, while they are either still in the TOP trap, or shortly after releasing them from the trap by rapidly shutting off the magnetic fields. Hence the condensates are essentially at rest and we expose them to the optical standing wave temporally. We detect the momentum transferred to the atoms from the diffraction process by taking an absorption image after a sufficient time delay, such that the various atomic wave-packets with different momenta have spatially separated.

Figure 9 is an example of normal diffraction of BEC by a short pulse of a weak and strong optical standing wave. The optical standing wave is applied along the z-axis, 2 ms after the condensate atoms have been released from the adiabatically expanded trap. For a weak pulse there is only a small phase modulation imposed on the cloud of atoms by the optical standing wave and only the first diffraction orders with momentum $\pm 2\hbar k$ are observed (fig. 9*a*). When the pulse intensity is increased by a factor of 5 there is a substantial phase modulation imposed on the released condensate atoms and higher diffraction orders are observed. In fig. 9*b*, both second- and third-order diffraction, corresponding to momentum transfer of $\pm 4\hbar k$ and $\pm 6\hbar k$, is clearly evident. The images in fig. 9 were taken 10 ms after the application of the diffraction pulse. The momentum spread of the undiffracted atoms is approximately $0.06\,\hbar k$ and so the diffracted components are clearly resolved.

7·4. *Bragg diffraction of atoms.* – When the atoms enter and exit the monochromatic standing wave adiabatically, energy conservation must be explicitly satisfied in the interaction between the atoms and the light field. In this regime, also known as the Bragg regime, the energy difference of the atom before and after the change of momentum of $2\hbar k$ must come from the photon field. This is typically accomplished by having the atomic beam incident on the standing wave at an angle such that the atoms see a differential Doppler shift between the two counterpropagating laser beams comprising the standing wave. This geometry is shown schematically in fig. 8*b*. Under these conditions, Bragg diffraction can be understood as a stimulated Raman transition between two momentum states. Figure 8*c* shows n-th-order Bragg diffraction as a $2n$-photon, stimulated Raman process in which photons are absorbed from one beam and stimulated into the other. Conservation of energy and momentum requires $(n2\hbar k)^2/2M = 2n\hbar kv\sin\theta$, where M is the mass of the atom, v is the longitudinal velocity of the atomic beam and θ is the angle of incidence of the atomic beam on the standing wave. Bragg diffraction of atoms by a near resonant optical standing wave was also first demonstrated in Pritchard's group at MIT [137] in 1988.

In the case where we start with a BEC essentially at rest, two laser frequencies are needed to satisfy energy conservation in the Bragg diffraction process. The difference in frequencies results in a moving standing wave, that would correspond to the differential Doppler shift observed by the atom in the stationary frame of the standing wave. In our experiments, we create our moving standing wave by having a frequency difference δ between the two counterpropagating waves that make up the standing wave. In the presence of this moving standing wave, an atom initially at rest will simultaneously absorb photons from the higher-frequency laser beam and be stimulated to emit photons into lower-frequency beam acquiring momentum $\pm 2n\hbar k$ in the process. In order to satisfy energy conservation, the detuning δ must be chosen such that $n\delta = n^2 4E_{\text{rec}}/\hbar$, where E_{rec} is the recoil energy.

Figure 10 shows first-, second- and third-order Bragg diffraction of Bose condensed atoms released from the magnetic trap. When the frequency difference between the two lasers is 100 kHz, atoms initially at rest can resonantly absorb a photon from the higher-frequency laser beam and be stimulated to emit a photon into the lower-frequency beam. The result of this process is a transfer of two units of photon momentum to the atoms, which then travels ballistically with a velocity of 6 cm/s. Similarly, by setting the frequency difference of the lasers to -100 kHz, the momentum transfer to the atoms from the Raman process will be in the opposite direction. Since Bragg diffraction of the atoms can be thought of as a two-level system (the initial and final momentum states) coupled by the Raman process, it is possible to transfer all of the atoms to the

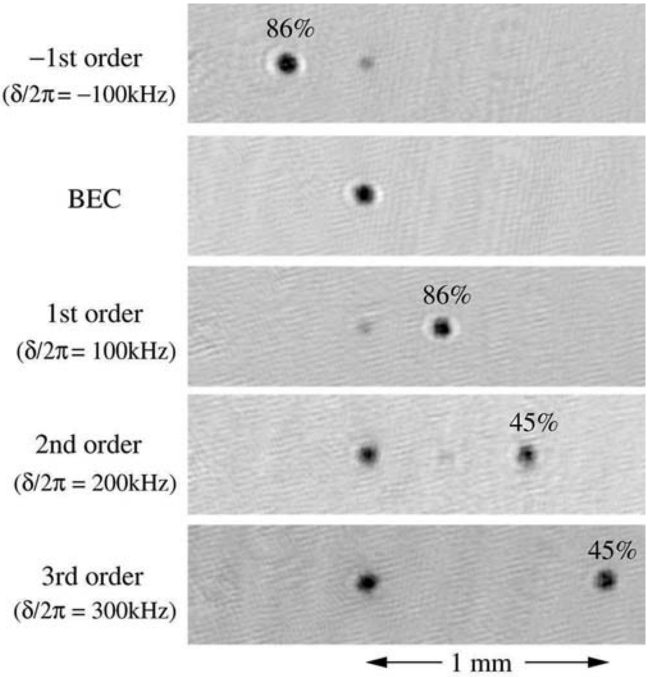


Fig. 10. – Bragg diffraction of a BEC: By applying a moving standing wave (whose velocity is determined by the frequency difference of the two waves comprising the standing wave) we can Bragg diffract a portion of the condensate into a well-defined momentum state.

final momentum state. We have observed first-order Bragg diffraction of 100% of the condensate atoms. The amount of transfer was reduced in the images of fig. 10 so that the location of the condensate atoms, initially at rest, could be easily identified. Second- and third-order Bragg diffraction was observed when the laser detuning was increased to 200 kHz and 300 kHz, respectively. We have observed up to sixth-order Bragg diffraction with a momentum transfer of $\approx 12\hbar k$ (corresponding to a velocity of 0.35 m/s).

7˙5. *Raman output coupling: Demonstration of a* CW *atom laser*. – In order to realize an atom laser from BEC, it is necessary to coherently extract the condensed atoms; that is, an atom output coupler is needed. The first demonstration of an output coupler for BEC was reported in 1997 [139], where coherent, rf-induced transitions were used to change the internal state of the atoms from a trapped state to an untrapped one. This method, however, did not allow the direction of the output coupled atoms to be chosen. The extracted atoms fell under the influence of gravity and expanded because of the intrinsic repulsion of the atoms. We have developed a highly directional method to optically couple out a variable fraction of a condensate. We use stimulated Raman transitions between magnetic sublevels to coherently transfer trapped condensate atoms to an untrapped state while giving them a momentum kick [140].

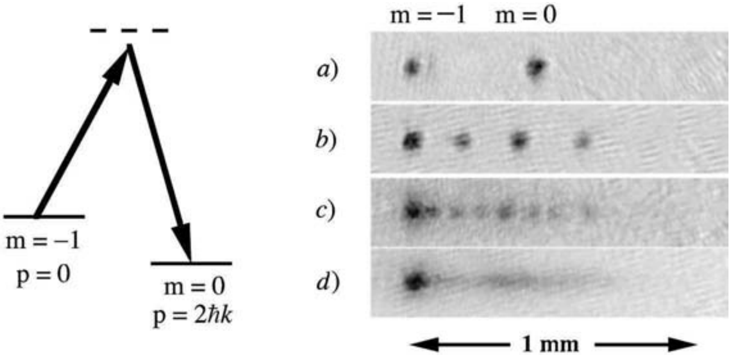


Fig. 11. – Raman output coupler. Left: A stimulated Raman transition is used to transfer $2\,\hbar k$ of momenta, and change the magnetic sublevel from the trapped $m = -1$ to the untrapped $m = 0$ state. Right: Series of images demonstrating multiple Raman output coupling of atoms from BEC en route to demonstrating a continuous stream of coherent atoms. In *a*)-*c*), one, three and seven $6\,\mu$s Raman pulses were applied to the condensate, respectively; *d*) is the result of the application of $1\,\mu$s Raman pulses at the full repetition rate of ≈ 20 kHz imposed by the frequency of the rotating TOP bias field (140 pulses in 7 ms). The pictures are absorption images taken after a time-of-flight period.

The Bragg diffraction of atoms [138] discussed earlier involves a stimulated Raman transition between different momentum states while keeping the atoms in the same magnetic sublevel. If the frequency difference between the lasers includes the additional Zeeman energy between two magnetic sublevels, a simultaneous change in the momentum and internal state of the condensate atoms can be achieved. This is illustrated in fig. 11, where BEC trapped in the $F = 1$, $m_F = -1$ state is transferred to the $F = 1$, $m_F = 0$ state. Two units of photon momentum are transferred in the Raman process, so the cloud of atoms in the $m_F = 0$ state has a velocity of 6 cm/s with respect to those atoms in the $m_F = -1$ state (fig. 11*a*).

We can repeatedly apply the Raman pulses to achieve multiple output coupling of atoms from a BEC. This is shown in fig. 11*a*)-*d*). In order to avoid changes in the Raman resonance frequency between different magnetic sublevels we synchronized the application of the Raman pulses to our rotating TOP field. (Our condensate atoms were displaced by gravity away from the zero of the quadrupole field, such that the local magnetic field was modulated by the rotating TOP bias field.) Figures 11*a*)-*c*) are optical absorption images of the condensate after one, three and seven Raman pulses, respectively. For these images, the TOP trap was held on for a 9 ms window during which time $6\,\mu$s Raman pulses were applied at a subharmonic of the rotating TOP bias frequency. The magnetic fields were then extinguished and the atoms were imaged 1.6 ms later. In fig. 11*d*, the TOP trap was held on for a 7 ms window during which time 140 Raman pulses were applied at the 20 kHz frequency of the rotating bias field and the distribution of atoms was imaged 1.6 ms later. The Raman pulse duration was reduced to $1\,\mu$s in order to couple less atoms out of the condensate during each Raman pulse. In the time between two Raman pulses each output coupled wavepacket moves only $2.9\,\mu$m. These pulses strongly overlap because this spatial separation of $2.9\,\mu$m is much smaller than the $\approx 50\,\mu$m size of the condensate, therefore the output coupled atoms form a continuous coherent matter-wave.

Our Raman output coupling scheme dramatically reduces the transverse momentum width of the extracted atoms compared to other methods such as rf output coupling [139]. This dramatic reduction occurs because the output coupled atoms have received a substantial momentum kick from the Raman process. If the atoms were simply released from the trap with no momentum transfer, they would undergo a burst of expansion due to the repulsive interactions with the other condensate atoms. In our output coupling scheme, however, this additional expansion energy is primarily channeled into the forward direction. The increase in the transverse momentum width due to the interaction between the atoms is reduced by roughly the ratio of the timescale over which the mean field repulsion acts on the freely expanding condensate, divided by the characteristic time it takes the output coupled atoms to leave the still trapped condensate. In our case, the reduction ratio is about a factor of ≈ 20 which results in a well-collimated beam of atoms.

One of the important properties of an optical laser is that the coherence length of the emitted beam of photons is much longer than the size of the cavity. A similar property for an atom laser would be highly desirable. The output coupled beam of our atom laser beam is much longer than the characteristic size of the condensate. Since stimulated Raman transitions are coherent processes, we expect the coherence length of this beam to be much longer than the size of the condensate. While we have not measured the intrinsic coherence length of our atoms laser beam, we have shown, in an interference experiment [141], that successive pulses are fully coherent. The coherence length of an atom laser beam, produced by using rf to continuously extract atoms from a magnetically trapped Bose condensate, was measured first by Hänsch's group in Germany [142].

8. – Nonlinear atom optics with Bose-Einstein condensates

The advent of the laser as an intense, coherent, light source enabled the field of nonlinear optics to flourish. The interaction of light in materials, whose index of refraction depends on the intensity, has led to effects such as multi-wave mixing of optical fields to produce coherent light of a new frequency, and optical solitons, pulses of light that propagate without dispersion. Nonlinear optics now plays an important role in many areas of science and technology. With the experimental realization of Bose-Einstein condensation (many atoms in a single quantum state) and the matter-wave or atom "laser" (atoms coherently extracted from a condensate), we now have an intense source of matter-waves analogous to the source of light from an optical laser. This has led us to the threshold of a new field of physics: nonlinear atom optics [143].

The analogy between nonlinear optics with lasers and nonlinear atom optics with Bose-Einstein condensates can be seen in the similarities between the equations that govern each system. For a condensate of interacting bosons, in a trapping potential V,

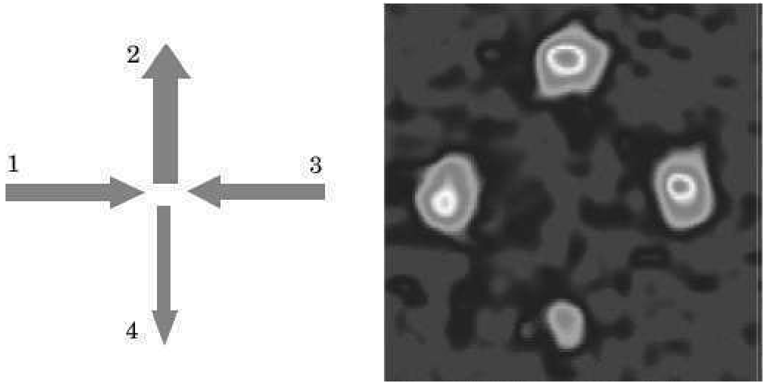


Fig. 12. – The process of four-wave mixing of matter-waves can always be transformed to a reference frame where the mixing process is degenerate (all of the waves have the same energy; left). The nonlinear term describing the mean-field, *s*-wave interaction of the atoms is responsible for the four-wave mixing. Atoms from waves 1 and 3 scatter off each other and go off back-to-back. The scattering process can be stimulated by wave 2, so that it is more likely that one of the scattering pairs goes into this wave. By momentum conservation, wave 4 is created. The small cloud of atoms in the image on the right is the fourth wave generated by four-wave mixing of matter-waves.

the macroscopic wave function Ψ satisfies a nonlinear Schrödinger equation [130],

$$i\hbar\frac{\partial\Psi}{\partial t} = \left(-\frac{\hbar^2}{2M}\nabla^2 + V + g|\Psi|^2 \right)\Psi, \tag{19}$$

where M is the atomic mass, g describes the strength of the atom-atom interaction ($g > 0$ for sodium atoms), and $|\Psi|^2$ is proportional to atomic number density.

8˙1. *Four-wave mixing with matter-waves.* – The nonlinear term in eq. (19) is similar to the third-order susceptibility term, $\chi^{(3)}$, in the wave equation for the electric field describing optical four-wave mixing. We therefore expect that if three coherent matter-waves are spatially overlapped with the appropriate momenta, a fourth matter-wave will be produced due to the nonlinear interaction, analogous to optical four-wave mixing. In contrast to optical four-wave mixing, the nonlinearity in matter-wave four-wave mixing comes from atom-atom interactions, described by a mean field; there is no need for an external nonlinear medium.

Using the atoms from a BEC, we have observed such four-wave mixing of matter-waves. This work is described in detail in ref. [144]. In our four-wave mixing experiment, we used optically induced Bragg diffraction [138] to create three overlapping wavepackets with appropriately chosen momenta. As the three wavepackets spatially separate, a fourth wavepacket, due to the wave-mixing process, is observed (see fig. 12).

The process of four-wave mixing of matter-waves (and also optical waves), can be thought of as Bragg diffraction off of a matter grating. In this picture, two of the matter-waves interfere to form a standing matter-wave grating. The third wave can Bragg diffract off of this grating, giving rise to the fourth wave. An alternative picture of four-wave mixing is in terms of stimulated emission. In this picture it is helpful to view the four-wave mixing process in a reference frame where the process looks like degenerate four-wave mixing; that is, all of the waves have the same energy.

In four-wave mixing, both energy and momentum (corresponding to phase matching) must be conserved. Since atoms, unlike photons, cannot be created out of the vacuum, we have the additional requirement for matter-waves of particle number conservation. (If we included the rest mass of the atom, particle number conservation is contained in energy conservation.) Given these three conditions, one can show that the only four-wave mixing configurations possible with matter-waves are those that can be viewed in some frame of reference as degenerate four-wave mixing. This is also the geometry of phase conjugation. Figure 12 shows the four-wave mixing geometry for matter-waves viewed in the degenerate or phase conjugation frame.

In the picture of four-wave mixing as arising from stimulated emission, atoms in waves 1 and 3 can be considered as undergoing an elastic collision. The scattering process results in atoms going off back-to-back in order to conserve momentum, but at some arbitrary angle with respect to the incident direction. (The scattering process is typically s-wave and the outgoing waves can be considered spherical.) In the presence of wave 2, however, this scattering process can be stimulated. There is an enhanced probability that one of the atoms from the collision of waves 1 and 3 will scatter into wave 2. (This probability is enhanced by the atoms in wave 2.) Because of momentum conservation, the enhanced scattering of atoms into wave 2 results in an enhanced number of atoms in wave 4. In this picture, it is obvious that the four-wave mixing process removes atoms from waves 1 and 3 and puts them into waves 2 and 4. This may have some interesting consequences in terms of quantum correlations between the waves.

8˙2. *Quantum phase engineering.* – A three-dimensional image of an arbitrarily complex object can be constructed by sending light, with sufficient spatial coherence, through the appropriate phase and/or amplitude mask. This is the basic principle behind physical optics, which includes wave phenomena like diffraction and holography. Diffraction can be achieved with a periodic phase and/or amplitude mask, while a more complicated mask is needed to construct a complex holographic image. In each case, the mask modifies the incoming wave and subsequent propagation produces the desired pattern of light. This idea can be readily adapted to atom optics, especially when the "incoming" matter-wave is from a highly coherent source such as a Bose-Einstein condensate.

We have developed a technique to optically imprint complex phase patterns onto a Bose-Einstein condensate in order to create interesting topological states. This technique is analogous to sending a wave through a thin phase mask. The basic idea is to expose the condensate atoms to a short pulse of laser light with a spatially varying intensity pattern. The laser detuning is chosen such that spontaneous emission is negligible. (The phase mask can also serve as an amplitude mask by tuning closer to resonance, so that spontaneous emission is significant.) The pulse duration is sufficiently short such that the atoms do not move an appreciable distance (*i.e.* the wavelength of light) during the pulse. This is sometimes referred to as the Raman-Nath regime. During the laser pulse, the AC Stark effect or optical dipole potential (see eq. (2)) shifts the energy of the atoms by $U(r,t)$. Hence the effect on the atomic wave function is to "instantaneously" change its phase. This effect can be represented by multiplying the wave function by the phase factor $\exp[i\phi(r)]$, where $\phi(r) = -\int U(r,t)\,\mathrm{d}t/\hbar$. Since the AC Stark or light shift is proportional to the intensity of the light, any spatial intensity variation in the light field will be written onto the BEC wave function as a spatial variation in its phase.

Optically induced phase imprinting is a tool for "quantum phase engineering" of the wave function to create a wide variety of states. For example, as discussed in the section on diffraction of the condensate, the application of a short pulse of standing wave light will imprint a sinusoidal phase onto the condensate. The imprinted wave function subsequently evolves in momentum states differing by $2\hbar k$. It should be possible to use quantum phase engineering to produce collective states of excitation of the interacting BEC, such as solitons and vortices. The application of a uniform intensity light field to half of the BEC imprints a relative phase difference between the two halves. This phase step is expected to give rise to dark solitons (see following section). Such solitons will propagate with a speed related to the phase difference [145], which can be adjusted by the intensity of the laser pulse.

It should also be possible to produce one or more vortices by applying a laser pulse which has a linearly varying, azimuthal, intensity dependence [146]. This will produce a topological winding of the BEC phase, which if large enough (*i.e.* 2π) should produce a vortex. Numerical solutions to a 3D Gross-Pitaevskii equation [147] show that this is the case; and also show that such a vortex, although unstable because it is created in a non-rotating trap, will live for a sufficient time to be observable. Increasing the phase winding will generate multiple vortices (vortices with more than $\hbar$ of angular momentum are not stable and will immediately split into multiple vortices each with angular momentum $\hbar$). Quantum phase engineering can generate arbitrary phase patterns, and perhaps other interesting quantum states. In this sense, it is a form of atom holography [148]. The technological challenge is mostly one of imaging. Any complicated pattern must be imaged to the size of the BEC, typically of order $50\,\mu$m.

8˙3. *Solitons in a* BEC. – Solitons are stable, localized waves that propagate in a nonlinear medium without spreading. They may be either bright or dark, depending on the details of the governing nonlinear wave equation. A bright soliton is a peak in the amplitude while a dark soliton is a notch with a characteristic phase step across it. Equation (19), which describes the weakly interacting, zero-temperature BEC, also supports solitons. The solitons propagate without spreading (dispersing) because the nonlinearity balances the dispersion; for eq. (19) the corresponding terms are the nonlinear interaction $g|\Psi|^2$, and the kinetic energy $-(\hbar^2/2M)\nabla^2$, respectively. Our sodium condensate only supports dark solitons because the atom-atom interactions are repulsive [145,149] ($g > 0$).

A distinguishing characteristic of a dark soliton is that its velocity is less than the Bogoliubov speed of sound [145, 149] $v_0 = \sqrt{gn/M}$ (n is the unperturbed condensate density) and they travel opposite to the direction of the phase gradient. The soliton speed v_s can be expressed either in terms of the phase step δ ($0 < \delta \leq \pi$), or the soliton "depth" n_d, which is the difference between n and the density at the bottom of the notch [145,149]:

$$\frac{v_\mathrm{s}}{v_0} = \cos\left(\frac{\delta}{2}\right) = \sqrt{1 - \frac{n_\mathrm{d}}{n}}. \tag{20}$$

For $\delta = \pi$ the soliton has zero velocity, zero density at its center, a width on the order of the healing length [149], and a discontinuous phase step. As δ decreases the velocity increases, approaching the speed of sound. The solitons are shallower and wider, with a more gradual phase step. Because a soliton has a characteristic phase step, optically imprinting a phase step on the BEC wave function should be a way to create a soliton.

8˙4. *Observation of solitons in a* BEC. – We modified the phase distribution of the BEC by employing the technique of quantum phase engineering discussed in an earlier section. The condensate atoms were exposed to a pulsed, off-resonant laser beam, coaxial with the absorption probe beam, with a spatial intensity profile such that only half of the BEC was illuminated. This was accomplished by blocking half of the laser beam with a razor blade and imaging this razor blade onto the condensate. The intensity pattern at the condensate, as observed by our absorption imaging system, had a light to dark (90% to 0%) transition region of 7 μm. The intensity required to imprint a phase of π was checked with a Mach-Zehnder atom interferometer based on optically induced Bragg diffraction [150,151]. Our Bragg interferometer [152] differs from previous ones in that we can independently manipulate atoms in the two arms (because of their large separation) and can resolve the output ports to reveal the spatial distribution of the condensate phase. When a phase of π was imprinted on one half of the condensate relative to the other half, the two output ports of the interferometer displayed the complementary halves of the condensate (see fig. 13).

To observe the creation and propagation of solitons, we measure BEC density distributions with absorption imaging after imprinting a phase step. Figure 14 shows the evolution of the condensate after the top half was phase imprinted with $\phi = 1.5\,\pi$, a phase for which we observe a single deep soliton (the reason for imprinting a phase step larger than π is discussed below). Immediately after the phase imprint, there is a steep phase gradient across the middle of the condensate such that this portion has a large velocity in the $+x$-direction. This velocity can be understood as arising from the impulse imparted by the optical dipole force, and results in a positive density disturbance that travels at or above the speed of sound. A dark notch is left behind, which is a soliton moving slowly in the $-x$-direction (opposite to the direction of the applied force).

A striking feature of the images is the curvature of the soliton. This curvature is due to the 3D geometry of the trapped condensate, and occurs for two reasons. First,

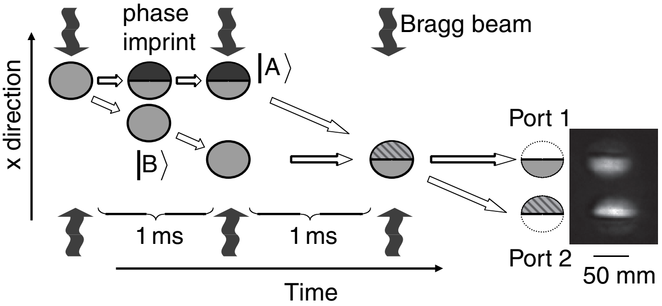


Fig. 13. – Space-time diagram of the matter wave interferometer used to measure the spatial phase distribution imposed on the BEC. Three optically induced Bragg diffraction pulses formed the interferometer. Each pulse consisted of two counter-propagating laser beams detuned by about $-2\,\text{GHz}$ from atomic resonance (so that spontaneous emission is negligible) with their frequencies differing by $100\,\text{kHz}$. The first pulse had a duration of $8\,\mu\text{s}$ and coherently split the condensate into two components $|A\rangle$ and $|B\rangle$ with equal number of atoms. $|A\rangle$ remained at rest and $|B\rangle$ received two photon recoils of momentum. When they were completely separated, we exposed the top half of $|A\rangle$ to a phase imprinting pulse of π, which changed the phase distribution of $|A\rangle$ while $|B\rangle$ served as a phase reference. 1 ms after the first Bragg pulse, a second Bragg pulse of $16\,\mu\text{s}$ duration brought $|B\rangle$ to rest and imparted two photon momenta to $|A\rangle$. When they overlapped again, 1 ms later, a third pulse of $8\,\mu\text{s}$ duration converted their phase differences into density distributions at ports 1 and 2, which appears in the images.

the speed of sound v_0 is largest at the trap center where the density is greatest, and decreases towards the condensate edge. Second, as the soliton moves into regions of lower condensate density, we find numerically that the density at its center, $n - n_\text{d}$, approaches zero, δ approaches π, and v_s decreases to zero before reaching the edge. This is because the soliton depth n_d rather than its phase offset δ appears to be a conserved quantity in a non-uniform medium.

A clear indication that the notches seen in fig. 14 are solitons, rather than simply sound waves, is their subsonic propagation velocity. To determine this velocity, we measure the distance after propagation between the notch and the position of the imprinted phase step along the x-direction. Because the position of our condensate varies randomly from shot-to-shot (presumably due to stray, time-varying fields) we cannot always apply the phase step at the center. A marker for the location of the initial phase step is the intersection of the soliton with the condensate edge, because at this point the soliton has zero velocity. Using images taken 5 ms after the imprint, at which time the soliton has not traveled far from the BEC center, we obtain a mean soliton velocity of 1.8(4) mm/s. This speed is significantly less than the mean Bogoliubov speed of sound $v_0 = 2.8(1)\,\text{mm/s}$. From the propagation of the notch in the numerical solutions (fig. 14, lower images) we obtain a mean soliton velocity, $v_\text{s} = 1.6\,\text{mm/s}$, in agreement with the experimental value. The experimental uncertainty is mainly due to the difficulty in determining the position of the initial phase step.

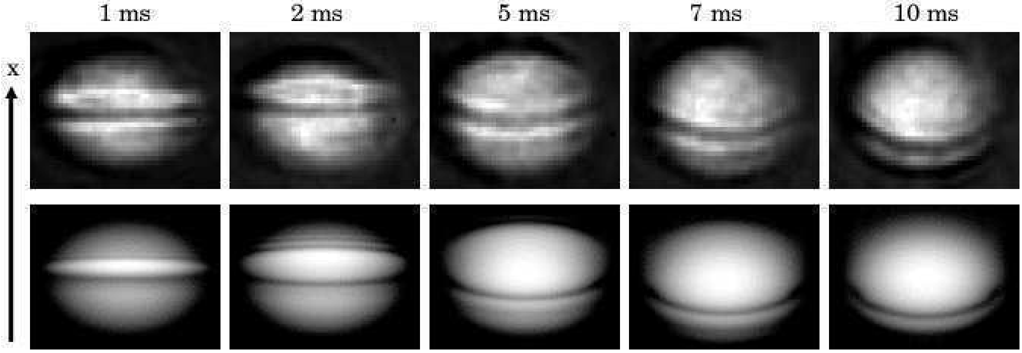


Fig. 14. – Experimental (upper) and theoretical (lower) images of the integrated BEC density for various times after we imprint a phase of about $1.5\,\pi$ on the top half of the condensate with a $1\,\mu$s pulse. The measured number of atoms in the condensate was $1.7(3) \times 10^6$, and this value was used in the calculations. A positive density disturbance moved rapidly in the $+x$-direction and a dark soliton moved oppositely at significantly less than the speed of sound. Because the imaging is destructive, each image shows a different BEC. The width of the images is $70\,\mu$m.

From the lower image of fig. 14 at 5 ms, we can extract the theoretical density and phase profile along the x-axis through the center of the condensate. The dark soliton notch and its phase step are centered at $x = -8\,\mu$m. This phase step, $\delta = 0.58\,\pi$, is less than the imprinted phase of $1.5\,\pi$. The difference is caused by the mismatch between the phase imprint and the phase and depth of the soliton solution of the nonlinear Schrödinger equation (eq. (19)): Our imprinting resolution of $7\,\mu$m is larger than the soliton width, which is on the order of the healing length ($0.7\,\mu$m), and we do not control the amplitude of the wave function.

In order to improve our measurement of the soliton velocity, we avoid the uncertainty in the position of the initial phase step by replacing the razor blade mask with a thin slit. This produces a stripe of light with a Gaussian profile ($1/e^2$ full width $\approx 15\,\mu$m). With this stripe in the center of the condensate, numerical simulations predict the generation of solitons that propagate symmetrically outwards. We select experimental images with solitons symmetrically located about the middle of the condensate, and measure the distance between them. For a small phase imprint of $\phi \approx 0.5\,\pi$ (at Gaussian maximum), we observe solitons moving at the Bogoliubov speed of sound, within experimental uncertainty. For a larger phase imprint of $\phi \approx 1.5\,\pi$, we observe much slower soliton propagation, in agreement with numerical simulations. An even larger phase imprint generates many solitons. The results of these experiments on the creation and propagation of solitons can be found in ref. [152]. Solitons in a BEC have also been observed by a group in Hannover [153].

8˙5. *Quantum atom optics.* – Imagine taking a BEC of N atoms and splitting it in half, putting one half in one trap and the other half in another trap. Furthermore, assume (quite reasonably) that each atom has a 50% probability of ending up in one trap or the other. What would we expect for the distribution of atoms in the two traps? We would find that the mean number of atoms in each trap is $N/2$, with an uncertainty of $\sqrt{N}/2$ (it is extremely unlikely to get exactly 500 000 heads if a coin is flipped 10^6 times.) This uncertainty in the number is intrinsic in physics, and is the source of noise in many experimental situations, such as the fundamental noise (called shot-noise) on a laser beam. Classically, this noise is unavoidable. Interactions, however, offer the ability to change the noise characteristics and create non-classical states. For example, nonlinear optics can be used to create squeezed states of light, where the noise characteristics in one degree of freedom can be less than the standard quantum limit given by the Heisenberg uncertainty principle. Squeezing does not violate the uncertainty principle. Instead, the noise in a degree of freedom that is unimportant (such as the intensity in the measurement of a laser frequency) is increased, while the noise is reduced in a relevant one (such as the phase of the laser). The field concerned with the non-classical statistical properties of light is called quantum optics. By analogy, interactions in a BEC have enabled the generation of non-classical statistical states of atoms, or quantum atom-optics.

The interactions between the atoms in a BEC can be exploited to generate number squeezed states; that is, atoms in traps with reduced number fluctuations. One of the earliest demonstrations of this was an experiment [154] in the group of Mark Kasevich, then at Yale. In this experiment, they loaded a BEC of 10^5 rubidium atoms into a shallow, 1-D optical lattice and then slowly increased in time the depth of the optical lattice (it took $\approx$ 200 ms to reach the final value). Initially, the atoms could tunnel from well to well, but as the depth of the optical lattice was increased, the tunneling was suppressed. The atom-atom interactions responsible for the nonlinear term in the Gross-Pitaevskii equation also means that there is an energy cost when two atoms get close to one another. Hence having many atoms in one well is energetically unfavorable, so the tendency is for the number of atoms in each well to be equal (see fig. 15, left). By raising the lattice intensity slowly enough, the system, starting in the lowest possible energy state (the BEC), could remain in the lowest energy state (by adiabatic following), and eventually evolve into an equal partitioning of the atoms into the wells of the lattice.

The reduction in number fluctuations could be inferred from the disappearance of phase coherence from well to well. That is, as the uncertainty in number decreases, the uncertainty in phase increases due to the number-phase relationship, which is well known in quantum optics. Suddenly releasing the atoms from the optical lattice results in diffraction peaks in the atomic momentum (see subsect. **7**˙3) if there is phase coherence across several wells. The disappearance of the diffraction implies a loss of phase coherence from well to well, which in this experiment was due to the reduction in number fluctuations. A factor of $\approx$ 10 suppression in number fluctuations (from $\approx$ 30 atoms per well to about 2 or 3 per well) was observed in the experiment. More recently, an experiment [155] was performed with a BEC in a 3-D optical lattice in which complete suppression of number fluctuations was observed. This complete disappearance of number fluctuations is a result of a quantum phase transition, which takes the system into the so-called Mott insulator phase where there is the same number of atoms (uniform filling) in each lattice site. The mechanism for this phase transition is the same as the

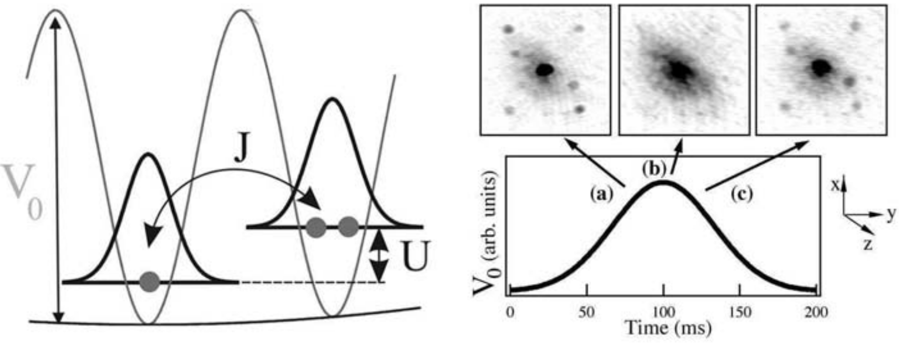


Fig. 15. – (Left) Schematic illustrating the relevant concepts for creating number (Fock) states in an optical lattice. Atoms can tunnel, at rate J, to adjacent lattice sites, however, there is an energy cost of U associated with putting two atoms in the same site. V_0 is the height of the optical lattice potential. (Right) Images of the atomic distribution (for a fixed time-of-flight period) after rapidly extinguishing the optical lattice at a particular depth. An initially delocalized state (a) exhibits diffraction peaks due to the periodic character of the optical lattice. As the depth of the optical lattice is increased to suppress tunneling (b), the system undergoes the transition to Mott insulator such that the atoms are in a Fock state and the diffraction disappears. If the optical potential is lowered (c) to allow tunneling, then the diffraction peaks reappear.

one responsible for suppression of number fluctuations in Kasevichs experiment. The atoms initially loaded in to the optical lattice from a BEC are delocalized and can move from well to well by tunneling. As the depth of the wells of the optical lattice is increased the movement of the atoms becomes constrained and the interactions between the atoms favors the filling of lattice sites with uniform numbers of atoms. The 3-D nature of the optical lattice results in a more abrupt transition from delocalized state to Fock state (uniform filling) than in the case of Kasevichs 1-D optical lattice experiment. Similar experiments have been performed at NIST [156] (see fig. 15, right).

The reduced number fluctuations of Fock states can be exploited to perform Heisenberg limited interferometry [157] in which an N-particle Fock state would have an N-fold increase in the frequency of phase evolution. To exploit this feature, a coherent superposition of Fock states (a so-called Schrdinger cat state) needs to be generated. One way to do this is to send two N-particle Fock states on a beamsplitter, simultaneously. An early example of such an experiment [158] using photons is the Hong-Ou-Mandel (HOM) effect. In this experiment, two photons produced by frequency down-conversion are sent to the two input ports of a beamsplitter, one photon in one input port and the other photon in the other port. If the twin photons arrive simultaneously on the beamsplitter, than each of the two output ports of the beamsplitter are 2-particle states. Since only two photons were sent into the beamsplitter and each output port of the beamsplitter are 2-particle states, the two particles must be in a superposition of two particles out one port and two particles out the other port, or a 2-particle cat state. This result has been confirmed in numerous experiments using photons and is often used to study the non-classical behavior of number states of photons. More recently, the HOM effect has been observed with Fock states of atoms in an optical lattice where the unit cell of can be dynamically transformed between a single and double-well structure [159].

Other measurements of the higher-order correlation functions can reveal the non-classical nature of a multi-particle field. The landmark experiment of Hanbury Brown and Twiss [160] is essentially a measurement of the 2-particle correlation function of an optical field. They observed that a thermal field (in their case, starlight) has a two-particle correlation function that starts at two and decays to zero over a time (the coherence time) inversely proportional to the spectral bandwidth. This experiment marked the beginning of the study the statistical properties of light and can arguably represent the start of the modern field of quantum optics. A similar measurement of the light emitted by a laser results in a 2-particle correlation function that starts at one and decays to zero after the coherence time. The same is true for observing three, four, five, etc. number of photons within this correlation time. This is due to the non-classical statistical properties of the state of the laser field, which can be represented by a coherent state. A measurement of the 2-particle correlation function for atoms extracted from a BEC also shows a similar behavior as a laser [161, 162]. A comparison of the 3-particle correlation at essentially zero time (measured by looking at three-body loss rates) for a thermal cloud and BEC differs by a factor of 6 (or 3!) [163], which is also the expected difference between thermal or classical source and a laser. Hence it seems even more appropriate to call the BEC a laser-like source of atoms, or atom laser. There are, however, instances where the statistical properties of atoms have no optical analog. Atoms come in two varieties bosons and fermions, while photons only exist as bosons. A recent experiment [164] measuring the 2-particle correlation function for ultracold clouds of metastable helium atoms has shown the Hanbury Brown Twiss effect for bosons (^{4}He) and the so-called anti-Hanbury Brown Twiss effect for fermions (^{3}He), because of Pauli blocking.

* * *

Much of the experimental work described here was carried out in the Laser Cooling and Trapping Group in the Atomic Physics Division of the National Institute of Standards and Technology. It is a pleasure to acknowledge the contributions to this work by the other members of the NIST staff of the Laser Cooling Group, S. Rolston and P. Lett, as well as former member C. Westbrook. This work would not have been possible without the contributions of many postdocs, students, and visitors to the group. While not attempting to be a review of all the work in this field, these notes have drawn extensively, both explicitly and implicitly, on the work of many groups throughout the world that have advanced the art and science of laser cooling. We are indebted to all of those colleagues, both cited and not, who have contributed so much to our understanding of laser cooling and to the pleasure of working in this field. We recognize that the references given in these notes are by no means complete and we apologize to those authors whose work has not been cited. The BEC experiments at NIST described in these lecture notes were carried out by present and past members of the Laser Cooling Group,Trey Porto, Steve Rolston, Jenni Sebby-Strabley, Marco Anderlini, Patty Lee, Ben Brown, Ed Hagley, Lu Deng, Mikio Kozuma, Johannes Denschlag, Jesse Simsarian, Jesse Wen, Yuri Ovchinnikov, Joerg-Helge Muller, Julien Cubizolles, Robert Lutwak, Rob Thompson, Aephraim Steinberg, Mike Gatzke and Gerhard Birkl. These experiments have also benefited greatly from direct interactions with our theoretical colleagues Paul Julienne, Carl Williams, Eite Tiesinga, Marek Trippenbach, Yehuda Band, Marya Doery, Charles Clark, David Feder, Mark Edwards, Phil Johnson and Keith Burnett.We also gratefully acknowledge the financial support not only of NIST, but of the U. S. Office of Naval Research, NASA and NSF.

REFERENCES

[1] Arimondo E., Phillips W. D. and Strumia F. (Editors), *Laser Manipulation of Atoms and Ions, Proceedings of the International School of Physics "Enrico Fermi", Varenna, 1991, Course CXVIII* (North Holland, Amsterdam) 1992.

[2] Dalibard J., Raimond J.-M. and Zinn-Justin J. (Editors), *Fundamental Systems in Quantum Optics* (North Holland, Amsterdam) 1992.

[3] Chu S. and Wieman C. (Editors), *Feature issue on laser cooling and trapping of atoms*, *J. Opt. Soc. Am. B*, **6** (1989) 2020.

[4] Meystre P. and Stenholm S. (Editors), *Feature issue on mechanical effects of light*, *J. Opt. Soc. Am. B*, **2** (1985) 1706.

[5] Metcalf H. and van der Straten P., *Phys. Rep.*, **244** (1994) 203.

[6] Kazantsev A. P., Surdutovich G. I. and Yakovlev V. P., *Mechanical Action of Light on Atoms* (World Scientific, Singapore) 1990.

[7] Minogin V. G. and Letokhov V. S., *Laser Light Pressure on Atoms* (Gordon and Breach, New York) 1987.

[8] Chu S., *Rev. Mod. Phys.*, **70** (1998) 685.

[9] Cohen-Tannoudji C., *Rev. Mod. Phys.*, **70** (1998) 707.

[10] Phillips W. D., *Rev. Mod. Phys.*, **70** (1998) 721.

[11] *Special issue on atom optics and interferometry*, *Appl. Phys. B*, **54** (1992) 321.

[12] *Special issue on atom optics and interferometry*, *J. Phys. (Paris)*, **4** (1994) 1877.

[13] *Special issue on atom optics and interferometry*, *Quantum Semiclass. Optics*, **8** (1996) 495.

[14] Dalilbard J. and Cohen-Tannoudji C., *J. Opt. Soc. Am. B*, **2** (1985) 1707.

[15] Cohen-Tannoudji C., Dupont-Roc J. and Grynberg G., *Atom-photon interactions: Basic Processes and Applications* (Wiley, New York) 1992.

[16] Gordon J. P. and Ashkin A., *Phys. Rev. A*, **21** (1980) 1606.

[17] Cook R. J., *Phys. Rev. A*, **22** (1980) 1078.

[18] Stenholm S., *Rev. Mod. Phys.*, **58** (1986) 699.

[19] Cohen-Tannoudji C., in *Fundamental Systems in Quantum Optics*, edited by J. Dalibard, J.-M. Raimond and J. Zinn-Justin (North Holland, Amsterdam) 1992, p. 1.

[20] Dalibard J., Thèse de doctorat d'état en Sciences Physique, Université de Paris VI (1986).

[21] Lett P. D., Phillips W. D., Rolston S. L., Tanner C. E., Watts R. N. and Westbrook C. I., *J. Opt. Soc. Am. B*, **6** (1989) 2084.

[22] Castin Y., Wallis H. and Dalibard J., *J. Opt. Soc. Am. B*, **6** (1989) 2046.

[23] Aspect A., Arimondo E., Kaiser R., Vansteenkiste N. and Cohen-Tannoudji C., *J. Opt. Soc. Am. B*, **6** (1989) 2112.

[24] Aspect A., Arimondo E., Kaiser R., Vansteenkiste N. and Cohen-Tannoudji C., *Phys. Rev. Lett.*, **61** (1988) 826.

[25] Pritchard D. E., Helmerson K., Bagnato V. S., Lafyatis G. P. and Martin A. G., in *Laser Spectroscopy*, edited by W. Persson and S. Svanberg, Vol. **VIII** (Springer-Verlag, Berlin) 1987, p. 68.

[26] Wallis H. and Ertmer W., *J. Opt. Soc. Am. B*, **6** (1989) 2211.

[27] Kasevich M. and Chu S., *Phys. Rev. Lett.*, **69** (1992) 1741.

[28] Phillips W. D., Prodan J. and Metcalf H., *J. Opt. Soc. Am. B*, **2** (1985) 1751.

[29] Chu S., Hollberg L., Bjorkholm J., Cable A. and Ashkin A., *Phys. Rev. Lett.*, **55** (1985) 48.

[30] Hodapp T. W., Gerz C., Furtlelehner C., Westbrook C. I. and Phillips W. D., *Appl. Phys. B*, **60** (1995) 135.

[31] Phillips W. and Metcalf H., *Phys. Rev. Lett.*, **48** (1982) 596.

[32] Cable A., Prentiss M. and Bigelow N. P., *Opt. Lett.*, **15** (1990) 507.

[33] Monroe C., Swann W., Robinson H. and Wieman C., *Phys. Rev. Lett.*, **65** (1990) 1571.

[34] Andreev S., Balykin V., Letokhov V. and Minogin V., *Pis'ma Zh. Eksp. Teor. Fiz.*, **34** (1981) 463 (*JETP Lett.*, **34** (1981) 442).

[35] Hoffnagle J., *Opt. Lett.*, **13** (1988) 102.

[36] Ketterle W., Martin A., Joffe M. A. and Pritchard D. E., *Phys. Rev. Lett.*, **69** (1992) 2483.

[37] Gaggl R., Windholz L., Umfer C. and Neureiter C., *Phys. Rev. A*, **49** (1994) 1119.

[38] Prentiss M. and Cable A., *Phys. Rev. Lett.*, **62** (1989) 1354.

[39] Letokhov V. S., Minogin V. G and Pavlik B. D., *Opt. Commun.*, **19** (1976) 72.

[40] Phillips W. D. and Prodan J. V., in *Laser-Cooled and Trapped Atoms*, edited by W. D. Phillips, *Natl. Bur. Stand. (U.S.) Spec. Publ.*, **653** (1983) 137; *Prog. Quantum Electron.*, **8** (1984) 231; in *Coherence and Quantum Optics V*, edited by L. Mandel and E. Wolf (Plenum, New York) 1984, p. 15; Phillips W., Prodan J. and Metcalf H., in *Laser Spectroscopy Laser VI*, edited by H. Weber and W. Luthy (Springer-Verlag, Berlin) 1983, p. 162.

[41] Ertmer W., Blatt R., Hall J. and Zhu M., *Phys. Rev. Lett.*, **54** (1985) 996.

[42] Salomon C. and Dalibard J., *C. R. Acad. Sci. Paris*, **306** (1988) 1319.

[43] Prodan J., Phillips W. and Metcalf H., *Phys. Rev. Lett.*, **49** (1982) 1149.

[44] Prodan J., Migdall A., Phillips W. D., So I., Metcalf H. and Dalibard J., *Phys. Rev. Lett.*, **54** (1985) 992.

[45] Migdall A., Prodan J., Phillips W., Bergeman T. and Metcalf H., *Phys. Rev. Lett.*, **54** (1985) 2596.

[46] Chu S., Bjorkholm J., Ashkin A. and Cable A., *Phys. Rev. Lett.*, **57** (1986) 314.

[47] Raab E., Prentiss M., Cable A., Chu S. and Pritchard D., *Phys. Rev. Lett.*, **59** (1987) 2631.

[48] Cornell E., Monroe C. and Wieman C., *Phys. Rev. Lett.*, **67** (1991) 3049.

[49] Spreeuw R. J. C., Gerz C., Goldner L., Phillips W. D., Rolston S. L., Westbrook C., Reynolds M. W. and Silvera I. F., *Phys. Rev. Lett.*, **72** (1994) 3162.

[50] Hofferberth S., Lesanovsky I., Fischer B., Verdu J. and Schmiedmayer J., *Nature Phys.*, **2** (2006) 710.

[51] Bethlem H. L., Berden G., Crompvoets F. M. H., Jongma R. T., van Roij A. J. A. and Meijer G., *Nature*, **406** (2000) 491.

[52] Aminoff C. G., Steane A. M., Bouyer P., Desbiolles P., Dalibard J. and Cohen-Tannoudji C., *Phys. Rev. Lett.*, **71** (1993) 3083.

[53] Petrich W., Anderson M. H., Ensher J. R. and Cornell E. A., *Phys. Rev. Lett.*, **74** (1995) 3352.

[54] Ashkin A., *Phys. Rev. Lett.*, **40** (1978) 729.

[55] Ashkin A. and Gordon J., *Opt. Lett.*, **4** (1979) 161.

[56] Dalibard J., Reynaud S. and Cohen-Tannoudji C., *Opt. Commun.*, **47** (1983) 395.

[57] Dalibard J., Reynaud S. and Cohen-Tannoudji C., *J. Phys. B*, **17** (1984) 4577.

[58] Gould P. L., Lett P. D., Phillips W. D., Julienne P. S., Thorsheim H. R. and Weiner J., in *Advances in Laser Science*, edited by A. Tam, J. Gole and W. Stwalley, Vol. **III** (American Institute of Physics, New York) 1987, p. 295.

[59] Gould P. L., Lett P. D., Julienne P. S., Phillips W. D., Thorsheim H. R. and Weiner J., *Phys. Rev. Lett.*, **60** (1988) 788.

[60] Helmerson K., *Interdisciplinary Laser Conference* (Unpublished, Monterey, CA) 1991.

[61] Rolston S. L., Gerz C., Helmerson K., Jessen P. S., Lett P. D., Phillips W. D., Spreeuw R. J. and Westbrook C. I., in *1992 Shanghai International Symposium on Quantum Optics*, edited by Yuzhu Wang, Yigui Wang and Zugeng Wang, *Proc. SPIE*, Vol. **1726** (1992), p. 205.

[62] Chu S., Bjorkholm J. E., Ashkin A., Gordon J. P. and Hollberg L. W., *Opt. Lett.*, **11** (1986) 73.

[63] Miller J. D., Cline R. A. and Heinzen D. J., *Phys. Rev. A*, **47** (1993) R4567.

[64] Miller J. D., Cline R. A. and Heinzen D. J., *Phys. Rev. Lett.*, **71** (1993) 2204.

[65] Adams C. S., Lee H. J., Davidson N., Kasevich M. and Chu S., *Phys. Rev. Lett.*, **74** (1995) 3577.

[66] Stamper-Kurn D. M., Andrews M. R., Chikkatur A. P., Inouye S., Miesner H.-J., Stenger J. and Ketterle W., *Phys. Rev. Lett.*, **80** (1998) 2027.

[67] Cook R. and Hill R., *Opt. Commun.*, **43** (1982) 258.

[68] Davidson N., Lee H. J., Adams C. S., Kasevich M. and Chu S., *Phys. Rev. Lett.*, **74** (1995) 1311.

[69] Lee H.-J., Adams C. S., Davidson N., Young B., Weitz M., Kasevich M. and Chu S., in *Atomic Physics*, edited by D. Wineland, C. Wieman and S. Smith, Vol. **14** (AIP Press, New York) 1995, p. 258.

[70] Phillips W. D., in *Laser Manipulation of Atoms and Ions, Proceedings of the International School of Physics, "Enrico Fermi", Varenna, 1991, Course CXVIII*, edited by E. Arimondo, W. D. Phillips and F. Strumia (North Holland, Amsterdam) 1992, p. 289.

[71] Ashkin A. and Gordon J., *Opt. Lett.*, **8** (1983) 511.

[72] Dalibard J., personal communication (1986).

[73] Walker T., *Laser Physics*, **4** (1994) 965.

[74] Dalibard J. and Cohen-Tannoudji C., *J. Opt. Soc. Am. B*, **6** (1989) 2023.

[75] Walhout M., Dalibard J., Rolston S. L. and Phillips W. D., *J. Opt. Soc. Am. B*, **9** (1992) 1997.

[76] Steane A. and Foot C., *Europhys. Lett.*, **14** (1991) 231.

[77] Werner J., Wallis H. and Ertmer W., *Opt. Commun.*, **94** (1992) 525.

[78] Prentiss M., Cable A., Bjorkholm J. E., Chu S., Raab E. L. and Pritchard D. E., *Opt. Lett.*, **13** (1988) 452.

[79] Sesko D., Walker T., Monroe C., Gallagher A. and Wieman C., *Phys. Rev. Lett.*, **63** (1989) 961.

[80] Walker T., Sesko D. and Wieman C., *Phys. Rev. Lett.*, **64** (1990) 408.

[81] Ketterle W., Davis K. B., Joffe M. A., Martin A. and Pritchard D. E., *Phys. Rev. Lett.*, **70** (1993) 2253.

[82] Ramsey N. F., *Molecular Beams*, in *International Series of Monographs on Physics*, edited by R. J. Elliott, J. A. Krumhansl, W. Marshall and D. H. Wilkinson (Oxford University Press, Oxford) 1956, 1985.

[83] Paul W., personal communication (1985).

[84] Heer C. V., *Rev. Sci. Instrum.*, **34** (1963) 532.

[85] Heer C. V., in *Quantum Electronics*, edited by C. H. Townes (Columbia University Press, New York) 1960, p. 17.

[86] Vladimirski V. V., *Zh. Eksp. Teor. Fiz*, **39** (1960) 1062.

[87] Wing W., *Prog. Quantum Electron.*, **8** (1984) 181.

[88] Ketterle W. and Pritchard D. E., *Appl. Phys. B*, **54** (1992) 403.

[89] Bergeman T., Erez G. and Metcalf H. J., *Phys. Rev. A*, **35** (1987) 1535.

[90] Davis K. B., Mewes M.-O., Joffe M. A., Andrews M. R. and Ketterle W., *Phys. Rev. Lett.*, **74** (1995) 5202.

[91] Petrich W., Anderson M. H., Ensher J. R. and Cornell E., in *Abstracts of Contributed Papers*, Vol. **ICAP XIV**, 1994, p. 1M-7; Davis K. B., Mewes M.-O., Joffe M. A. and Ketterle W., *ibid.*, p. 1M-3.

[92] Sesko D., Fan C. G. and Wieman C. E., *J. Opt. Soc. Am. B*, **5** (1988) 1225.

[93] Phillips W. D., Gould P. L. and Lett P. D., in *Laser Spectroscopy*, edited by W. Persson and S. Svanberg, Vol. **VIII** (Springer, Berlin) 1987, p. 64.

[94] Lett P. D., Watts R. N., Westbrook C. I., Phillips W. D., Gould P. L. and Metcalf H. J., *Phys. Rev. Lett.*, **61** (1988) 169.

[95] Phillips W. D., Westbrook C. I., Lett P. D., Watts R. N., Gould P. L. and Metcalf H. J., in *Atomic Physics*, edited by S. Haroche, J. C. Gay and G. Grynberg, Vol. **11** (World Scientific, Singapore) 1989, p. 633.

[96] Dalibard J., Salomon C., Aspect A., Arimondo E., Kaiser R., Vansteenkiste N. and Cohen-Tannoudji C., in *Atomic Physics*, edited by S. Haroche, J. C. Gay and G. Grynberg, Vol. **11** (World Scientific, Singapore) 1989, p. 199.

[97] Ungar P. J., Weiss D. S., Riis E. and Chu S., *J. Opt. Soc. Am. B*, **6** (1989) 2058.

[98] Sheehy B., Shang S.-Q., Watts R., Hatamian S. and Metcalf H., *J. Opt. Soc. Am. B*, **6** (1989) 2165.

[99] Weiss D. S., Riis E., Shevy Y., Ungar P. J. and Chu S., *J. Opt. Soc. Am. B*, **6** (1989) 2072.

[100] Sheehy B., Shang S.-Q., van der Straten P., Hatamian S. and Metcalf H., *Phys. Rev. Lett.*, **64** (1990) 858.

[101] Nienhuis G., in *Laser Manipulation of Atoms and Ions*, *Proceedings of the International School of Physics "Enrico Fermi", Varenna, 1991, Course CXVIII*, edited by E. Arimondo, W. D. Phillips and F. Strumia (North Holland, Amsterdam) 1992, p. 171.

[102] Emile O., Kaiser R., Gerz C., Wallis H., Aspect A. and Cohen-Tannoudji C., *J. Phys. II*, **3** (1993) 1709.

[103] Aspect A., Emile O., Gerz C., Kaiser R., Vansteenkiste N., Wallis H. and Cohen-Tannoudji C., in *Laser Manipulation of Atoms and Ions*, *Proceedings of the International School of Physics "Enrico Fermi", Varenna, 1991, Course CXVIII*, edited by E. Arimondo, W. D. Phillips and F. Strumia (North Holland, Amsterdam) 1992, p. 401.

[104] Cohen-Tannoudji C. and Phillips W. D., *Phys. Today*, **43** (1990) 33.

[105] Cohen-Tannoudji C., in *Laser Manipulation of Atoms and Ions, Proceedings of the International School of Physics "Enrico Fermi", Varenna, 1991, Course CXVIII*, edited by E. Arimondo, W. D. Phillips and F. Strumia (North Holland, Amsterdam) 1992, p. 99.

[106] Castin Y., Dalibard J. and Cohen-Tannoudji C., in *Light Induced Kinetic Effects on Atoms, Ions and Molecules*, edited by L. Moi, S. Gozzini, C. Gabbanini, E. Arimondo and F. Strumia (ETS Editrice, Pisa) 1991, p. 5.

[107] Castin Y., Doctoral Dissertation, Ecole Normale Supérieure (1992).

[108] Salomon C., Dalibard J., Phillips W. D., Clairon A. and Guellati S., *Europhys. Lett.*, **12** (1990) 683.

[109] Gerz C., Hodapp T. W., Jessen, Jones K. M., Phillips W. D., Westbrook C. I. and Mølmer K., *Europhys. Lett.*, **21** (1993) 661.

[110] Rolston S. and Phillips W., *Proc. IEEE*, **79** (1991) 943.

[111] Clairon A., Salomon C., Guellati S. and Phillips W., *Europhys. Lett.*, **16** (1991) 165.

[112] Kasevich M., Riis E., Chu S. and DeVoe R., *Phys. Rev. Lett.*, **63** (1989) 612.

[113] Bize S., Sortais Y., Abgral M., Zhang S., Calonico D., Mandache C, Lemonde P., Laurent Ph., Santarellie G., Salomon C., Clairon A., Luiten A. and Tobar M., in *Proc. 6th Symp. Freq. Standards and Metrology* edited by Gill P. (World Scientific) 2002, p. 53.

[114] Weyers S., Hübner U., Schröder R., Tamm Chr. and Bauch A., *Metrologia*, **38** (2001) 343.

[115] Jefferts S. R., Shirley J. H., Parker T. E., Heavner T. P., Meekhof D. M., Nelson C. W., Levi F., Costanzo G., DeMarchi A., Drullinger R. E., Hollberg L., Lee W. D. and Walls F. L., *Metrologia*, **39** (2002) 321.

[116] Gibble K. and Chu S., *Phys. Rev. Lett.*, **70** (1993) 1771.

[117] Tiesinga E. and Williams C., private communication.

[118] Kasevich M. and Chu S., *Phys. Rev. Lett.*, **69** (1992) 1741.

[119] Davidson N., Lee H. J., Kasevich M. and Chu S., *Phys. Rev. Lett.*, **72** (1994) 3158.

[120] Lemonde P., Laurent Ph., Simon E., Santarelli G., Clairon A., Salomon C., Dimarcq N. and Petit P., *Trans. Inst. Meas.*, **48** (1999) 512; see also `www.estec.esa.nl/spaceflight/map/map/atomiclock.htm`.

[121] Ashby N., *Bull. Am. Phys. Soc.*, April 2000, Q21.001; see also `www.jpl.nasa.gov/sespd/space/fphysics.html`.

[122] Peters A., Chung K., Young B., Hensley J. and Chu S., *Philos. Trans. R. Soc. London, Ser. A*, **355** (1997) 2223.

[123] Lenef A., Hammond T. D., Smith E. T., Chapman M. S., Rubinstein R. A. and Pritchard D. E., *Phys. Rev. Lett.*, **78** (1997) 760.

[124] Gustavson T. L., Bouyer P. and Kasevich M. A., *Phys. Rev. Lett.*, **78** (1997) 2046.

[125] Cladé T. L., De Mirandes E., Cadoret M., Guellati-Khélifa S., Schwob C. Nez F., Julien L. and Biraben F., *Phys. Rev. Lett.*, **96** (2006) 033001.

[126] Gupta S., Dieckmann K., Hadzibabic Z. and Pritchard D. E., *Phys. Rev. Lett.*, **89** (2002) 140401.

[127] Anderson M. H., Ensher J. R., Matthews M. R., Wieman C. E. and Cornell E. A., *Science*, **269** (1995) 198.

[128] Davis K. B., Mewes M.-O., Andrews M. R., van Druten N. J., Durfee D. S., Kurn D. M. and Ketterle W., *Phys. Rev. Lett.*, **75** (1995) 3969.

[129] Sackett C. A., Stoof H. T. C. and Hulet R. G., *Phys. Rev. Lett.*, **80** (1998) 2031; Bradley C. C., Sackett C. A., Tollett J. J. and Hulet R. G., *Phys. Rev. Lett.*, **75** (1995) 1687.

[130] Dalfovo F., Giorgini S., Pitaevskii L. P. and Stringari S., *Rev. Mod. Phys.*, **71** (1999) 463.

[131] Witte A., Kisters T., Riehle F. and Helmcke J., *J. Opt. Soc. Am. B*, **9** (1992) 1030.

[132] Barrett T., Dapore-Schwartz S., Ray M. and Lafyatis G., *Phys. Rev. Lett.*, **67** (1991) 3483.

[133] Miranda S. G., Muniz S. R., Telles G. D., Marcassa L. G., Helmerson K. and Bagnato V. S., *Phys. Rev. A*, **59** (1999) 882.

[134] Hess H. F., *Phys. Rev. B*, **34** (1986) 3476.

[135] Masuhara N., Doyle J. M., Sandberg J. C., Kleppner D., Greytak T. J., Hess H. F. and Kochanski G. P., *Phys. Rev. Lett.*, **61** (1988) 935.

[136] Moskowitz P. E., Gould P. L., Atlas S. R. and Pritchard D. E., *Phys. Rev. Lett.*, **51** (1983) 370; Gould P. L., Ruff G. E. and Pritchard D. E., *Phys. Rev. Lett.*, **56** (1986) 827.

[137] Martin P. J., Oldaker B. G., Miklich A. H. and Pritchard D. E., *Phys. Rev. Lett.*, **60** (1988) 515.

[138] Kozuma M., Deng L., Hagley E. W., Wen J., Lutwak R., Helmerson K., Rolston S. L. and Phillips W. D., *Phys. Rev. Lett.*, **82** (1999) 871.

[139] Mewes M.-O., Andrews M. R., Kurn D. M., Durfee D. S., Townsend C. G. and Ketterle W., *Phys. Rev. Lett.*, **78** (1997) 582.

[140] Hagley E. W., Deng L., Kozuma M., Wen J., Edwards M. A., Helmerson K., Rolston S. L. and Phillips W. D., *Science*, **283** (1999) 1706.

[141] Hagley E. W., Deng L., Kozuma M., Trippenbach M., Band Y. B., Edwards M. A., Doery M., Julienne P. S., Helmerson K., Rolston S. L. and Phillips W. D., *Phys. Rev. Lett.*, **83** (1999) 3112.

[142] Esslinger T., Blochr I. and Hänsch T. W., *J. Mod. Opt.*, **47** (2000) 2725.

[143] Lens G., Meystre P. and Wright E. W., *Phys. Rev. Lett.*, **71** (1993) 3271.

[144] Deng L., Hagley E. W., Wen J., Trippenbach M., Band Y. B., Julienne P. S., Simsarian J. E., Helmerson K., Rolston S. L. and Phillips W. D., *Nature*, **398** (1999) 218.

[145] Reinhardt W. P. and Clark C. W., *J. Phys. B*, **30** (1997) L785.

[146] Dobrek L., Gajda M., Lewenstein M., Sengstock K., Birkl G. and Ertmer W., *Phys. Rev. A*, **60** (1999) R3381.

[147] Feder D. L., Clark C. W. and Schneider B. I., *Phys. Rev. Lett.*, **82** (1996) 4956.

[148] Fujita J., Morinaga M., Kishimoto T., Yasuda M., Matsui S. and Shimizu F., *Nature*, **380** (1996) 691.

[149] Jackson A. D., Kavoulakis G. M. and Pethick C. J., *Phys. Rev. A*, **58** (1998) 2417.

[150] Torii Y., Suzuki Y., Kozuma M., Kuga T., Deng L. and Hagley E. W., *Phys. Rev. A*, **61** (2000) 041602.

[151] Giltner D. M., McGowan R. W. and Lee S. A., *Phys. Rev. Lett.*, **75** (1995) 2638.

[152] Denschlag J., Simsarian J. E., Feder D. L., Clark C. W., Collins L. A., Cubizolles J., Deng L., Hagley E. W., Helmerson K., Reinhardt W. P., Rolston S. L., Schneider B. I. and Phillips W. D., *Science*, **287** (2000) 97.

[153] Burger S., Bongs K., Dettmer S., Ertmer W., Sengstock K., Sanpera A., Shlyapnikov G. V. and Lewenstein M., *Phys. Rev. Lett.*, **83** (1999) 5198.

[154] Orzel C., Tuchman A. K., Feneslau M. L., Yasuda M. and Kasevich M. A., *Science*, **291** (2001) 2386.

[155] Greiner M., Mandel O., Esslinger T., Hänsch T. W. and Bloch I., *Nature*, **415** (2002) 39.

[156] Porto J. V., Rolston S., Laburthe-Tolra B., Williams C. J. and Phillips W. D., *Philos. Trans. R. Soc. London, Ser. A*, **361** (2003) 1417.

[157] Wineland D. J., Bollinger J. J., Itano W. M., Moore F. L. and Heinzen D. J., *Phys. Rev. A*, **46** (1992) R6797.

[158] Hong C. K., Ou Z. Y. and Mandel L., *Phys. Rev. Lett.*, **59** (1987) 2044.

[159] Sebby-Strabley J., Brown B. L., Anderlini M., Lee P. J., Phillips W. D., Porto J. V. and Johnson P. R., *Phys. Rev. Lett.*, **98** (2007) 200405.

[160] Hanbury Brown R. and Twiss R. Q., *Nature*, **178** (1956) 1046.

[161] Öttl A., Ritter S., Köhl M. and Esslinger T., *Phys. Rev. Lett.*, **95** (2005) 090404.

[162] Schellekens M., Hoppeler R., Perrin A., Viana Gomes J., Boiron D., Aspect A. and Westbrook C. I., *Science*, **310** (2005) 5748.

[163] Burt E. A., Ghrist R. W., Myatt C. J., Holland M. J., Cornell E. A. and Wieman C. E., *Phys. Rev. Lett.*, **97** (1997) 337.

[164] Jeltes T., McNamara J. M., Hogervorst W., Vassen W., Krachmalnicoff V., Schellekens M., Perrin A., Chang H., Boiron D., Aspect A. and Westbrook C. I., *Nature*, **445** (2007) 402.

This page intentionally left blank

Metrology with cold atoms

P. Lemonde

LNE-SYRTE, Observatoire de Paris - 61 Avenue de l'Observatoire, 75014, Paris, France

1. – Introduction

More than twenty years ago, one of the early motivations for the development of laser cooling techniques has been the idea that with cold atoms, accurate measurements could be made even more accurate [1]. The fathers of laser cooling probably did not anticipate to what extent this would be true. Cold atoms are now vastly used for the measurement of an ever increasing number of physical quantities. Time and frequency is the topical domain to which they were first applied [2-4]. In about 15 years, atomic clocks have improved by merely two orders of magnitude by using laser cooled atoms. Time and frequency measurements can now be performed with a fractional accuracy approaching 10^{-16} with atomic fountain clocks [5-7]. The basic reasons for this improvement are the following:

- *Cold atoms can be observed for an extended period of time.* This leads to very narrow resonances. In an atomic fountain, atoms coherently interact with the probe field for about one second corresponding to a Fourier limited width of the resonance of 1 Hz. Even though this has not been applied to atomic clocks so far, coherent interaction of atoms with electromagnetic waves have been extended up to almost ten seconds [8, 9]. The 1 Hz width of the resonance in a fountain should be compared to the $\sim$ 100 Hz of atomic lines in traditional thermal beam machines [10]. Reducing the width of the resonance both improves the resolution of the frequency measurements and reduces many of the systematic effects that shift the atomic frequency.

- *Frequency shifts related to the motion of the atoms are strongly reduced with cold atoms.* The first-order Doppler effect is proportional to the atomic velocity([1]), while the second-order Doppler effect (or relativistic time dilation effect) varies as the square of the atomic velocity. Both effects were two of the most severe limitations to the performances of thermal beam clocks [10]. In fountains, the atomic velocity is lowered by two orders of magnitude.

- *Experiments using cold atoms are extremely versatile.* A number of parameters can be varied independently and over a broad range: atom number, velocity at which they are launched, internal state... This proves extremely useful for the evaluation of all the systematic effects that might affect the frequency accuracy of any precision measurement performed with cold atoms. Several examples of this point will be given throughout this course. For the moment, let us just give one example which illustrates this point. As will be seen in details later on, in an atomic fountain it is possible to change the atom number by a factor of 2 without affecting their position and velocity distribution. It is straightforward to understand that by comparing the frequency delivered by the fountain with the full atom number and with half this atom number, one can evaluate the effects depending on the atom number (such as the frequency shift due to cold collisions) without affecting any other effect. By comparison, changing the atom number in an atomic beam machine can be done by modifying the temperature of the oven which is the source of the atomic beam. This in turn affects the velocity distribution of atoms via the temperature dependence of the Boltzmann distribution, but also via technical effects like a possible temperature dependence of the direction of the atomic beam. All the effects depending on the atomic velocity (like the first- and second-order Doppler effect) will therefore be entangled to this evaluation of atom number dependent frequency shift.

- *Large numbers of atoms can be used simultaneously.* All the precedent advantages of cold atoms are shared with trapped ions [11]. For precision measurements however, only one to a few ions can be used simultaneously, the Coulomb repulsion between ions leading to dramatic perturbations if more ions are used. Atoms do interact which each other but the corresponding perturbation is orders of magnitude smaller, so that to a certain extent, the problem can be dealt with. The typical number of atoms contributing to the signal in an atomic fountain is about 10^6–10^7 with a corresponding signal-to-noise ratio at detection that is orders of magnitude larger than in trapped ion experiments.

In the following, we will explore in details how one takes advantage of these fundamental features in actual devices. In sect. **2**, the operation of fountains will be explained in details. We will then discuss their performances in sect. **3** and particularly explore the physical effects that limit their performance: the quantum projection noise which

[1] In the field of time and frequency metrology this effect is often related to as the distributed cavity phase shift, see sect. **3** and [10].

sets a fundamental limit to the frequency stability of this type of clocks and the effect of collisions between cold atoms which yield a severe though controllable frequency shift of the clock transition. We will also give a brief overview of the other sources of possible frequency error in fountains and rapidly explain how they are evaluated. Altogether, with a fractional frequency accuracy of about 10^{-16} fountains allow the most accurate measurements of all physics. Atomic fountains are now the working horses of frequency metrology and have been developed in all major metrology institutes [12-18].

One would like however to do even better and a few ideas which are presently being explored by the researchers in this field will be presented in sect. **4**. The general trend is to use atomic transitions at a much higher frequency. By switching to optical transitions (instead of the microwave transitions presently used in fountains) there are good reasons to hope for accuracies at least one order of magnitude better [19]. This proposed new generation of clocks uses either cold atoms or ions confined in an electrodynamic trap. The still preliminary results obtained along this line are extremely promising [20-28]. For instance, the control of all systematic frequency shifts at a level below 10^{-16} has very recently been reported in a clock using a single trapped Hg^+ ion [24]. New ideas are emerging that in principle allow the use of confined atoms for reaching even better results [25-33].

Finally, beyond time and frequency metrology, cold atoms are now vastly used for the accurate measurement of other physical quantities. Taking advantage of the wave nature of atoms which manifests itself increasingly with the long De Broglie wavelengths of atoms at micro-Kelvin temperatures or below, the field of atom interferometry has developed impressively [34-41]. Accelerometers, gyrometers, gravimeters, gradiometers based on atom interferometry have been built with sensitivities that approach or overcome that of "traditional" devices based on totally different techniques.

Along the same line of ideas, cold atoms have been used for the measurement of the fine structure constant α with an uncertainty in the 10^{-9} range [42, 43]. These measurements are based on the following expression:

$$\alpha^2 = \frac{2R_\infty}{c}\frac{m_{\mathrm{at}}}{m_{\mathrm{e}}}\frac{h}{m_{\mathrm{at}}}\,, \tag{1}$$

where R_∞ is the Rydberg constant, c the speed of light, h the Planck constant and m_{e} and m_{at} the electron and atomic mass, respectively. R_∞ and $m_{\mathrm{at}}/m_{\mathrm{e}}$ are known with great accuracy and the goal of the measurements is to determine h/m_{at} [44]. This is done by coherently transferring the momentum of photons to cold atoms by repeating N absorption-stimulated emission cycles. The final atom velocity differs from the initial one by $2N\hbar k/m_{\mathrm{at}}$, where k is the photons' wavevector. Since the latter can be determined from a frequency measurement (using the Doppler effect) and therefore be known with great accuracy, the measurement of this velocity difference in turn yields $\hbar/m_{\mathrm{at}}$, which is the desired quantity for the determination of α.

We will not develop in the following all these experiments and will restrict this detailed description to atomic clocks. The whole subject would require a dedicated book.

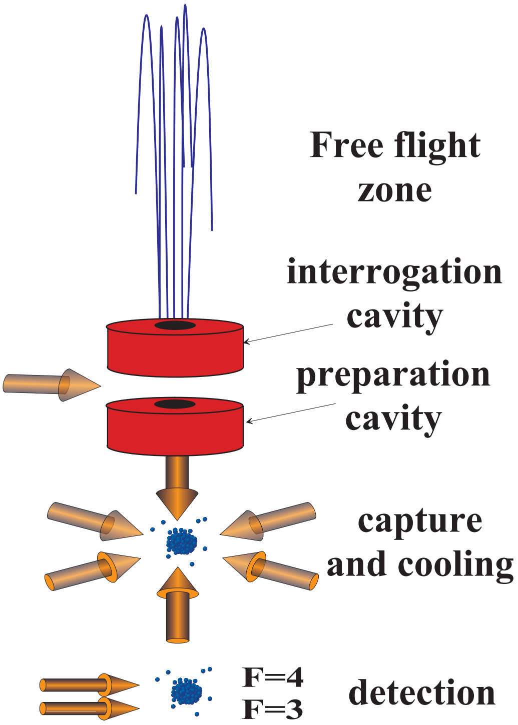


Fig. 1. – Schematic view of an atomic fountain.

Nevertheless, the reader can find in the references given above the relevant information. We would just like to stress here the deep similarities in the operation of these devices, that are comparable to atomic clocks using cold atoms.

2. – Description of atomic fountains

A schematic view describing the operation of an atomic fountain is shown in fig. 1. The device operates sequentially: each cycle of the clock starts by a phase during which the atoms are prepared, followed by an interrogation phase where the clock transition is probed. The effect of this interrogation is detected at the end of the cycle. The collected information is then used to lock the frequency of the oscillator used to probe the atoms to the atomic frequency.

2˙1. *Atomic preparation.* – The preparation phase aims at preparing both the external and internal degrees of freedom of the atoms. Using the strong D_2 transition (the energy levels of Cs are shown in fig. 2) and standard laser cooling techniques [47], atoms are first captured from a vapor or an atomic beam and cooled down to a temperature in the μK range [48]. In modern fountains, this process is extremely efficient and up to $\sim 10^9$ atoms are captured in about one hundred milliseconds [5]. The cloud of cold

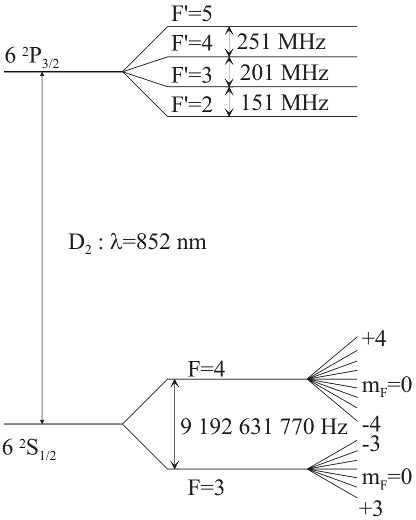


Fig. 2. – Energy levels of Cs used in an atomic fountain. Note that fountains using ^{87}Rb atoms have been and are being operated [45, 46, 5]. The energy levels of Rb exhibit a similar structure [10].

atoms is then launched upwards at a typical velocity of 3–4 m/s. This is achieved while cooling further the atoms down to a temperature of about 1 μK with the moving molasses technique: by adjusting the relative frequency of the laser beams and making use of the Doppler effect, one creates the usual optical molasses configuration but in a frame that is moving upwards at the desired velocity in the laboratory frame [3].

Once they have been launched atoms are spread among the various Zeeman states of the upper hyperfine ground state ($|F = 4\rangle$ in the case of Cs, see fig. 2). In order to cancel the first-order sensitivity to magnetic field, only the $|m = 0\rangle$ Zeeman state is used for the clock transition and atoms populating the $|m \neq 0\rangle$ states are removed from the atomic cloud. The atoms cross a first microwave cavity in which $|m = 0\rangle$ atoms are selectively transferred to $|F = 3\rangle$ (see fig. 1). A laser beam located above this first cavity and tuned to the $|F = 4\rangle - |F' = 5\rangle$ component of the D_2 line then blasts away atoms remaining in $|F = 4\rangle$ after crossing the cavity. One is therefore left with atoms in $|m = 0\rangle$ that have been shelved in $|F = 3\rangle$ and are not affected by this laser beam. This internal state selection method proves extremely efficient. In practice, the state prepared with this technique is pure to within 10^{-3}, limited by a residual optical pumping by the pushing beam of $|F = 4\rangle$ atoms down to the $|F = 3\rangle$ level. In addition, as will be seen in subsect. **3**˙2, this state selection technique can be used to vary the atomic density in an extremely well controlled way, allowing an accurate measurement of the frequency shift due to the collisions between the cold atoms.

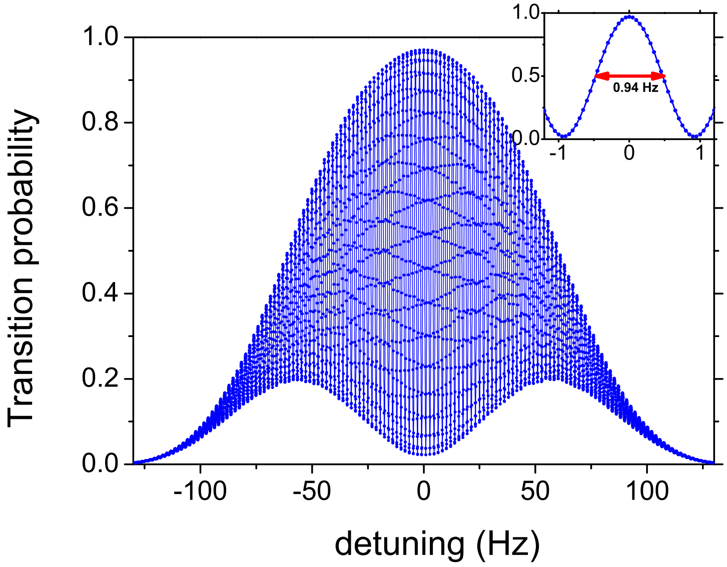


Fig. 3. – Experimental atomic resonance in a fountain (from [5]). Plotted is the transition probability from one clock state to the other as a function of the detuning of the microwave with respect to the atomic resonance. The inset is a zoom on the central fringe which is used a frequency discriminator to lock the microwave frequency to the atoms.

2˙2. *Clock transition interrogation.* – Atoms in $|F=3, m=0\rangle$ then follow their free fall trajectory in the clock apparatus. The clock transition is then interrogated: atoms cross twice the interrogation cavity, once on their way up, once on their way down. At each passage through the cavity, they undergo a microwave pulse which couples states $|F=3, m=0\rangle$ and $|F=4, m=0\rangle$. This two-pulse interrogation scheme is known as the Ramsey scheme [49-51, 10]. It can be viewed as an atom interferometer: the first pulse creates a coherent superposition of both clock states with equal weights. After a drift time T (which corresponds to the free flight of the atoms above the cavity), the effect of the second pulse is to compare the phase of the atomic coherence to the phase of the microwave field inside the cavity: since the atomic coherence evolves at the transition frequency, this relative phase is proportional to the microwave detuning with respect to the atomic resonance and to the duration T of the above-cavity free flight phase. If the phase difference is a multiple of 2π, the interference is constructive and the transfer to state $|F=4\rangle$ is total. If, on the other hand, the phase difference is an odd multiple of π ($(2p+1)\pi$), the interference is destructive and atoms end-up in $|F=3\rangle$. The transition probability from $|F=3\rangle$ to $|F=4\rangle$ therefore behaves sinusoidally as a function of the microwave detuning, as shown in fig. 3 in which an experimental atomic resonance signal is plotted. The full-width at half maximum of fringe is $1/2T$, *i.e.* typically 1 Hz in an atomic fountain. The central fringe of the pattern is zoomed in the inset of fig. 3. It is the frequency discriminator used for locking the microwave frequency to the atoms.

2˙3. *Detection.* – Finally, the atomic internal state is measured after the interrogation. The measurement is performed by laser induced fluorescence. Atoms cross a first laser beam resonant to $|F = 4\rangle - |F' = 5\rangle$: atoms in $|F = 4\rangle$ scatter about 10^4 photons before being blasted away by the laser beam. Remaining atoms (in $|F = 3\rangle$) then cross a second beam located below the first one and that repumps them to $|F = 4\rangle$. They finally cross a third beam similar to the first one. By collecting the fluorescence induced by the first and third beams, one has access to the relative population of both $|F = 4\rangle$ and $|F = 3\rangle$ states and therefore to the transition probability from one state to the other induced by the microwave interrogation. This detection scheme can be extremely efficient: by detecting both $|F = 4\rangle$ and $|F = 3\rangle$ atoms, the measurement is insensitive on the atom number fluctuations. Detection beams 1 and 3 can be issued from the same laser source, allowing for a common mode rejection of most of the noise sources due to the residual frequency or amplitude fluctuations of this laser. Finally, with a large number of detected photons per atom, the signal is large enough so that both the detector noise(2) and the photon shot noise are made negligible. The quantum efficiency of the measurement is 1 for atom numbers larger than 10^4 [52]. In a well-designed fountain, typically 10^6 atoms contribute to the signal.

2˙4. *Operation of the fountain clock.* – The operation of the fountain consists in the repetition of the above-described cycle. At each cycle the frequency of the oscillator probing the clock transition is changed:

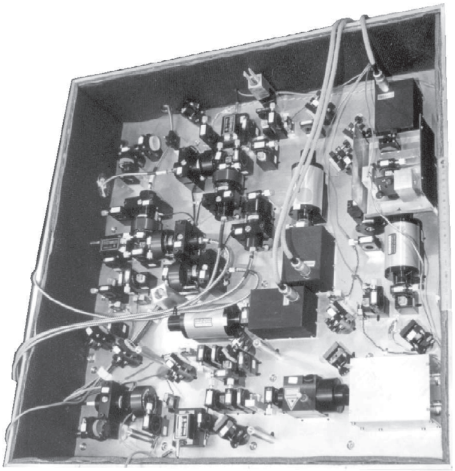
$$\nu^l(t) = \nu^f(t) + \nu^c_n \tag{2}$$

with $\nu^f(t)$ the oscillator's free-running frequency and ν^c_n the frequency correction at cycle n. It is given by

$$\nu^c_n = \nu^c_{n-1} + (-1)^n \left(\frac{\Delta}{2} + K(P_{n-1} - P_{n-2})\right). \tag{3}$$

Here $\Delta = 1/2T$ is the width of the central fringe of the resonance pattern (full width at half maximum, see fig. 3) and P_n is the transition probability measured at cycle n. The frequency modification from one cycle to the next one contains two terms: one (the Δ term) that is used to alternately probe both sides of the resonance, the other ($K(P_{n-1} - P_{n-2})$) being a frequency correction such that the frequency of the oscillator is locked to the atomic frequency. K is a gain coefficient chosen so as to optimize the servo loop. The typical time constant of the loop is a few cycles of the fountain operation.

(2) In practice a simple low noise photodiode is used.



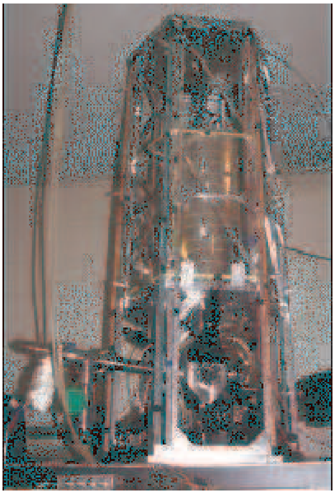


Fig. 4. – Left: atomic-fountain laser system. Right: clock tube.

Note that the loop given by eq. (3) contains a first-order integrator with a gain that keeps increasing for longer and longer averaging times [10]. More sophisticated transfer functions can be used but the simple loop described here is sufficient if a good enough local oscillator is used: this is the case with state-of-the-art quartz oscillators. Finally, photographs of the fountain main parts are shown in fig 4.

3. – Performances of fountains

Two concepts are usually used to characterize the performance of an atomic clock: its frequency stability and its frequency accuracy [10]. The stability reflects the frequency fluctuations (or frequency noise) of the clock, while the accuracy is its ability to faithfully reproduce the atomic frequency. The instantaneous frequency $\nu(t)$ of the clock can formally be written as[3]

$$\nu(t) = \nu_0(1 + \varepsilon + y(t)), \tag{4}$$

with ν_0 the atomic frequency, ε is the fractional frequency shift of the clock with respect to the atomic frequency, and $y(t)$ represent the fractional frequency fluctuations of the clock signal. The uncertainty on ε reflects the (in)accuracy of the clock, while the statistical properties of $y(t)$ reflect its frequency (in)stability.

[3] $\nu(t)$ here is the frequency of the locked oscillator $\nu^l(t)$ in eq. (2) but with the frequency modulation removed. It can be physically generated with an additional frequency synthesizer or more simply numerically processed by means of the computer that controls the operation of the fountain.

3˙1. *Frequency stability and quantum projection noise.* – From eq. (2), it is clear that $y(t)$ has two contributors: one due to the noise of the free-running oscillator, the other due to the noise of the corrections. In practice, the role of the servo-loop is precisely such that both contributions cancel if $y(t)$ is averaged over times much longer than the servo time constant. This cancelation is perfect to within the measurement noise of P_n. Several sources of noise can affect this measurement like the noise of the detection system briefly discussed in sect. **2**. Historically, two sources of noise have actually been a problem. The limitation due to the free-running oscillator frequency noise has been a problem as long as quartz oscillator have been used for that purpose [53-55]. More recently, with the use of cryogenic sapphire oscillators of extremely good short term frequency stability [56], the problem has vanished [52, 5, 57]. The main source of noise in fountains is the fundamental limit set by the quantum nature of the measurement, or quantum projection noise [58, 59, 52].

When the atoms are detected, their internal state is a coherent superposition of both clock states:

$$|\psi\rangle = c_3|F=3\rangle + c_4|F=4\rangle. \tag{5}$$

The probability P of detecting a given atom in state $|F=4\rangle$ is given by

$$P = \langle\psi|P_4|\psi\rangle = |c_4|^2, \tag{6}$$

with $P_4 = |F=4\rangle\langle F=4|$ the projection operator onto state $|F=4\rangle$. The outcome of this measurement cannot be predicted with certainty (except if c_3 or $c_4 = 0$ in which case the measurement is insensitive to the frequency of the probe oscillator) and undergoes quantum fluctuations of variance

$$\sigma_P^2 = \langle\psi|P_4^2|\psi\rangle - (\langle\psi|P_4|\psi\rangle)^2. \tag{7}$$

Since $P_4^2 = P_4$ we end up with

$$\sigma_P = \sqrt{P - P^2} = \sqrt{P(1-P)}. \tag{8}$$

The previous result is for the one-atom case. Of course, the noise on the transition probability measurement averages down with a large set of atoms.

Let us assume that N atoms are detected. If these atoms are uncorrelated, the quantum state of this ensemble can be written as a product state of each individual atom([4]):

$$|\Psi\rangle = \prod_{i=1}^{N} |\psi_i\rangle \text{ with } |\psi_i\rangle = c_3|F=3\rangle_i + c_4|F=4\rangle_i. \tag{9}$$

[4] The quantum statistics of the atoms is not taken into account here. In a fountain, the phase space density is less than 10^{-5} and the bosonic nature of the atoms does not play any role in the detection process.

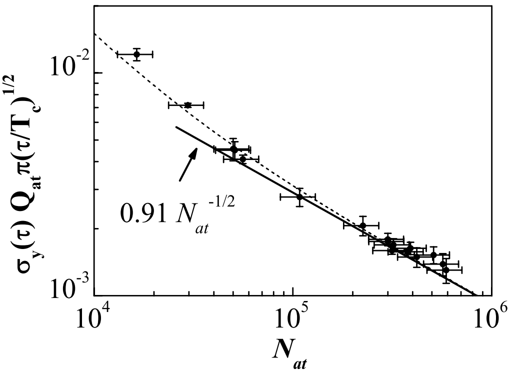


Fig. 5. – Frequency noise of the atomic fountain vs the number of detected atoms (from [52]). The $1/\sqrt{N_{at}}$ dependence is a clear signature that the measurement is quantum limited. For $N_{at} < 2 \times 10^4$ the noise is limited by the fluorescence detectors.

The average number of atoms detected in state $|F = 4\rangle$ is

$$N_4 = \langle\Psi|\sum_{i=1}^{N} P_4^{(i)}|\Psi\rangle \text{ with } P_4^{(i)} = |F = 4\rangle_i\langle F = 4|_i, \tag{10}$$

giving an average transition probability, $P = N_4/N = |c_4|^2$. P undergoes quantum fluctuations given by

$$\sigma_P^2 = \frac{1}{N^2}\left(\langle\Psi|\left(\sum_{i=1}^{N} P_4^{(i)}\right)^2|\Psi\rangle - N_4^2\right). \tag{11}$$

Using $P_4^{(i)2} = P_4^{(i)}$ and $\langle\Psi|P_4^{(i)}P_4^{(j)}|\Psi\rangle_{i\neq j} = |c_4|^4$, one ends up with

$$\sigma_P = \frac{1}{\sqrt{N}}\sqrt{P(1-P)}. \tag{12}$$

The quantum projection noise therefore exhibits a characteristic $1/\sqrt{N}$ dependence which allows to easily discriminate with other sources of noise. For instance, the limitation due to the Dick effect is independent of N, while the detection noise can scale like $1/N$ (noise of the photodiodes for instance) or be independent of N (common mode noise of the detection laser for instance). In state-of-the art fountains, the quantum projection limit is reached for N ranging from a few 10^4 up to 10^6 as illustrated in fig. 5.

This noise on the transition probability measurement is converted to frequency noise with a multiplication factor that is given by the slope of the atomic resonance. For a Ramsey interaction, the slope is proportional to $\sqrt{P(1-P)}$ so that the limitation to the frequency stability by the quantum projection noise is independent of the choice of P or,

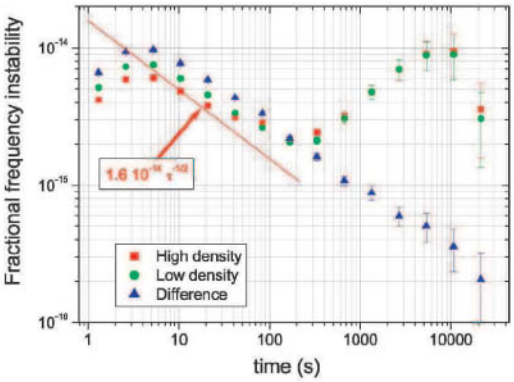


Fig. 6. – Atomic-fountain frequency stability (from [5]). The square and circles correspond to two different configurations used to measure the collisional frequency shift (see subsect. **3**˙2).

equivalently, of the choice of the depth of the modulation used to lock the oscillator to the atomic resonance.

Finally, for a linewidth of 1 Hz and about 10^6 detected atoms, the frequency stability of fountains is about $10^{-14}\,\tau^{-1/2}$ with τ the averaging time expressed in seconds. In fig. 6 the Allan deviation of the fountain FO2 of SYRTE compared to a cryogenic sapphire oscillator at that level of frequency stability is plotted [5]. In this configuration, measurements with a fractional resolution of 10^{-16} can be performed by averaging for about 10^4 seconds which is compatible with a full accuracy evaluation of the clock at that level.

Reducing further the noise of fountains would in principle be possible if one creates entangled atomic states and several schemes have been proposed towards this goal [60,61]. None of them has been successfully implemented in a fountain so far. Among other difficulties, one problem of these proposed configurations is to ensure that they do not induce unaffordable frequency shifts.

3˙2. *Frequency accuracy, collisions between cold atoms.* – A large number of systematic effects can affect the accuracy of an atomic fountain. This is illustrated in table I in which the budget of the FO2 fountain is presented.

Most of these effects are under control thanks to a careful design and realization of the clock. They are evaluated either by varying the operating parameters of the fountain, or by modeling the effect, or both. For instance, if one changes the microwave power inside the interrogation cavity, one is sensitive to a possible leakage of this cavity, to the spectral impurities of the microwave (*e.g.*, asymmetric sidebands), to the pulling by the other atomic transitions close to the clock transition (*i.e.* involving $m \neq 0$ states) which leads to the so-called Ramsey and Rabi pulling, etc. By cross-checking the effects of various parameters, one can disentangle all the contributors to the accuracy budget.

We will not discuss further most of these effects which have been extensively studied in the literature. The black-body radiation shift is now well understood and controlled

Table I. – *Accuracy budget of the FO2 fountain (from* [5]*). Note that the gravitational redshift is not included in this accuracy budget since it is not intrinsic to the clock. This effect must be evaluated if distant clocks are to be compared* [62].

Systematic effect	Frequency shift ($\times 10^{16}$)	Uncertainty ($\times 10^{16}$)
Quadratic Zeeman effect	1927.3	0.3
Black-body radiation	−168.2	2.5
Collisions and cavity pulling	−357.5	2.0
Microwave spectral purity and leakage	0	4.3
First order Doppler effect	0	3
Ramsey and Rabi pulling	0	1
Microwave recoil	0	1.4
Second order Doppler effect	0	0.1
Collisions with background gas	0	1
Total uncertainty		6.5

at a level below 10^{-16} [63-68]. The first order Doppler effect has also lead to interesting discussions and is now controlled at a level close to 10^{-16} [69,68]. Discussions on the effect of the recoil energy due to the absorption or emission of microwave photons inside the cavity can be found in ref. [70, 71], and more general considerations about the evaluation of fountains in ref. [12, 6].

Among these systematic effects, the frequency shift due to collisions between cold atoms has drawn a specific attention of the researchers in the field. Soon after the construction of the first atomic fountain, it was realized that this effect could be a true problem for the ultimate accuracy of this type of clocks [72, 73]. In ref. [72] frequency shifts due to collisions as large as 10^{-12} have been reported. In addition to being a key point in the accuracy evaluation of cold-atom clocks, this interest for cold collisions was also motivated by more general ideas: cold collisions play a key role in degenerate quantum gases and constitute a whole field of physics (see, *e.g.*, [74]). For atomic clocks, the difficulty with this frequency shift due to collisions comes from its dependence on the atomic density. The frequency shift grows linearly with the density, a parameter that is very difficult to control accurately. In a fountain, the density is inhomogeneous and varies with time: typically, the atomic density is three times larger when the atoms enter the interrogation cavity on their way up as compared to when they leave the cavity on their way down. For all these reasons this frequency shift was controlled at the 10 % level only until 2002. This lead to the development of fountains using Rb atoms, an atom for which the frequency shift due to collisions is two orders of magnitude smaller than for Cs [46, 45].

Fortunately, in 2002, a new method was proposed that allowed to evaluate and control the frequency shift due to collisions at a level better than one percent [75]. The method consists in using an adiabatic microwave pulse to prepare either a sample of density n (full adiabatic passage) or of density precisely $n/2$ (half adiabatic passage). The

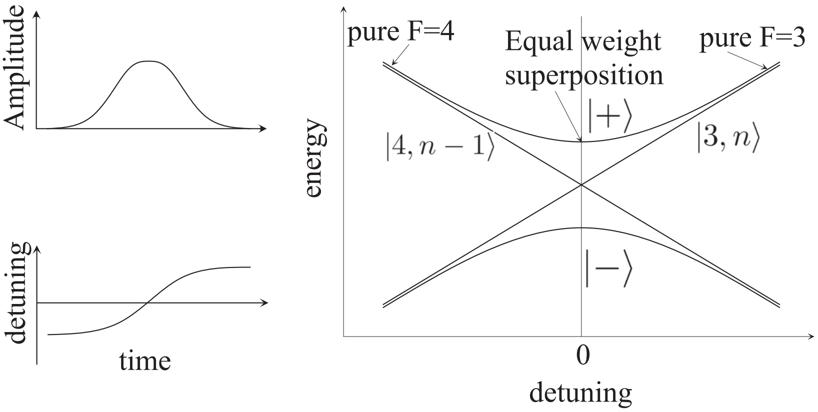


Fig. 7. – Principle of the adiabatic passage method: left amplitude and frequency of the microwave pulse used to perform the adiabatic passage. Right: atomic states in the uncoupled (lines) and coupled (curves) basis.

velocity distribution of the atomic cloud is not affected by this preparation stage so that the relative evolution of the atomic density in both cases is exactly the same. If one alternates measurements with full adiabatic passages and measurements with half adiabatic passages, one measures in real time the frequency shift due to collisions that is equal to two times the frequency difference between both configurations. The principle of operation of the technique is sketched in fig. 7: while the atoms cross the preparation microwave cavity, a microwave pulse is applied of amplitude and detuning with respect to the atomic resonance as sketched in fig. 7. The detuning is swept across resonance. If the adiabaticity condition is met during this frequency and amplitude sweep, atoms initially prepared in one of the eigenstates of the coupled Hamiltonian adiabatically follow this eigenstate during the evolution [76,77]. In the large detuning limit (compared to the Rabi frequency of the microwave coupling), the microwave hardly couples the internal atomic states and the eigenstates of the total hamiltonian (atoms+microwave) coincide with the decoupled states. For small detunings, $|F = 3, n\rangle$ and $|F = 4, n-1\rangle$ are coupled, with n the number of photons in the microwave mode. The eigenstates of the total Hamiltonian are linear superpositions of $|F = 3, n\rangle$ and $|F = 4, n-1\rangle$. For symmetry reasons, the weight of both states is equal on resonance. Remarkably, this property holds independently of the amplitude of the field (or Rabi frequency). One then sees the method used to prepare both samples. In both cases one starts with a cloud of atoms prepared in $|F = 4\rangle$. Inside the cavity, atoms follow the eigenstate $|+\rangle$ as illustrated in fig. 7: they follow the upper branch of this graph. To prepare the sample of density n the adiabatic pulse is terminated and all atoms exit the preparation cavity in state $|F = 3\rangle$. To prepare the sample of density $n/2$, the adiabatic pulse is interrupted on resonance and atoms leave the cavity in an equal weight linear superposition of states $|F = 3\rangle$ and $|F = 4\rangle$. A laser beams then selectively pushes away atoms in $|F = 4\rangle$

leaving an atomic sample with the desired properties. The only critical parameter in this preparation pulse is the frequency at which the pulse is interrupted. However, this frequency can be controlled by the fountain itself and in practice, this adiabatic passage method operates at the 10^{-3} level. In fact the limiting factor comes from the residual atoms in $|F = 3, m \neq 0\rangle$ due to imperfect repumping of the launched atoms and residual optical pumping by the pushing beam.

With this technique it is now possible to control the collisional frequency shift in Cs fountain at the 10^{-16} while maintaining a sufficiently large number of atoms to keep a frequency stability in the low $10^{-14}\,\tau^{-1/2}$ range.

Altogether, fountains now have an accuracy close to $2{-}3\times10^{-16}$ [68]. It seems however extremely difficult to go well beyond. First this would require a better frequency stability than presently achieved. Evaluating all the systematics effects at a level of 10^{-17} seems unrealistic with a frequency noise higher than a few $10^{-15}\,\tau^{-1/2}$. Reaching this level with fountains would require to increase the atoms number by roughly one order of magnitude, up to conditions where the collisional frequency shift would again be a problem. This problem seems tractable if one uses Rb atoms instead of Cs. Unfortunately, this would not be the only problem: first-order Doppler effect (or equivalently shift due to residual phase gradients inside the cavity), effect of microwave leakage, etc... seems untractable at that level. Many of these effects, which are controlled today at a level that corresponds to 10^{-6} of the width of the atomic resonance, are reduced when the atomic resonance width is decreased. Although this seems difficult with fountains (for which the resonance width scales only as $1/\sqrt{H}$ with H the height of the fountain), this is clearly feasible by using atomic transitions of much higher frequency: optical transitions. For the same experimental linewidth, the fractional width of an optical transition is four to five orders of magnitude smaller than its microwave counterpart.

4. – Beyond fountains

Optical frequency standards clearly appear as the future of the field. As said above, the possibility to significantly increase the atomic quality factors of the transitions (or equivalently reduce the fractional width of the resonance) is a strong motivation. This possibility has been foreseen decades ago and developed as a whole field of research (see, *e.g.*, ref. [19] or [78]). Historically, the field has developed along three lines addressing different problems. The first one is the quantum reference itself and will be rapidly discussed in this section. The second one is the laser used to probe the clock transition. With ever improving performances, optical atomic atomic clocks have been more and more demanding in terms of spectral purity of the probe oscillator. The best probe lasers now have a frequency stability below 10^{-15} for a one second averaging time [79-81]. Still, a two or three orders of magnitude improvement would be required in order to reach the quantum limit with neutral atom optical frequency standards operated in optimal conditions [82]. The third line of research has aimed at the measurement of optical frequencies. In the absence of fast enough electronics that would allow to compare optical clocks with each other or with their RF counterpart, optical frequency chains have been built over the last three decades [83-86]. This field has culminated with the advent of frequency chains based on femto-second frequency combs that have essentially solved the problem [87].

4·1. *Optical clocks with neutral atoms in free fall.* – Optical clocks with neutral atoms in free fall are the direct "optical versions" of atomic fountains. Their development started with the advent of laser cooling techniques [88]. Rapidly, alkaline-earth atoms were identified as possible candidates for this type of experiments due to their energy level configuration: they possess a broad transition connecting the ground state $^1S_0$ to the excited state $^1P_1$ for laser cooling. They also have narrow intercombination lines to be used as clock transitions. Experiments with Ca at PTB and NIST [21], Sr at JILA [89] and Mg at university of Hannover and Copenhagen [90,91] have been carried out leading to accuracies better than 10^{-14} for Ca and Sr [92,93,89]. Other atoms with two-photon transitions have also been investigated for optical frequency standards with free falling atoms, like H [94] at MPQ or Ag and MPQ and INM [95,96].

One of the difficulties of this type of clocks is the control of the residual first-order Doppler effect. In fractional units, this effect is independent of the frequency of the atomic transition. It is already one of the limitations of microwave fountains (see sect. **3**) and turns out to be more difficult to control for an optical clock [21]. Most of the experiments with neutrals in free fall have now been either redesigned to operate with trapped atoms (see below) or modified for other experiments of physics (*e.g.*, study of cold collisions [97]) or with the goal to build embarkable devices [98].

An efficient method to circumvent the Doppler effect problem is to use atoms confined to the Lamb-Dicke regime [99]: if the atomic motion is restricted to an area of typical dimensions smaller than the clock wavelength, the Doppler effect vanishes. The problem is then to confine atoms without inducing deleterious frequency shifts of the clock transition. This can be easily done with ions thanks to their external charge that allows very efficient trapping with fields that are small enough to not significantly alter the internal structure of the ion [11].

4·2. *Optical clocks with trapped ions.* – Due to this cancelation of the first-order Doppler effect, trapped ions have been investigated for more than thirty years as possible candidates for optical frequency standards. In most cases the choice of the ion has been guided by the possibility to have an accessible laser cooling transition together with a narrow clock transition. Yb^+ and Sr^+ have been studied and evaluated down to a level approaching the accuracy of microwave fountains, at PTB [23], NPL [20,100] and NRC [22]. At NIST this approach has been pushed one step further and an accuracy below 10^{-16} has been reported with a single Hg^+ clock [24]. The same laboratory has demonstrated the possibility to use two separate ions in the same device, one clock ion and one cooling and read-out ion: one uses quantum logic techniques to couple both ions [101]. The decisive feature of this approach is that it vastly opens the ensemble of possible ions for a clock. Indeed, most ions with interesting clock transition and metrological properties cannot be directly laser cooled either because of the absence of available cycling transitions, or because their possible cooling wavelength is in a spectral region out of reach of current laser technology. One can now choose an ion that is "optimal" with respect to spectroscopy (Al^+ in the case of [101]) on the one side and another ion that is simple to laser cool and probe on the other side (Be^+ in [101]).

The major drawback of trapped ion clocks is the fact that the number of quantum references is in practice limited to one (or at most a few units). If one tries to use a large number of ions, the Coulomb interaction between the ions leads to a large frequency shift. The major consequence is that the quantum limit to the frequency stability is orders of magnitude better in the case of optical clocks with atoms.

4·3. *Optical clocks with trapped neutral atoms.* – The difficulty with neutral atoms is that the absence of external charge imposes large trapping fields to confine the atoms. This necessarily shifts the atomic levels. To give an order of magnitude, compensating gravity with a laser trap requires shifts of typically 10 kHz [102], *i.e.* of the order of 10^{-11} of an optical frequency. If one aims at a ultimate accuracy of 10^{-17}–10^{-18}, the shift should be controlled at the 10^{-6}–10^{-7} level. This is at first sight very difficult because generally speaking this shift depends on the laser power and polarization, two parameters for which metrology is far from the required level. In 2003, a configuration has been proposed that in principle solves this problem [29]. If one uses a clock transition involving atomic states with a total electronic angular momentum $J = 0$ and a laser trap of well-chosen wavelength/frequency, it can be shown that the clock transition can be unaffected by the trap, for any laser power and polarization. This idea is extremely attractive since it in principle allows to combine the advantages of trapped ion clocks and of clocks using neutral atoms in free fall. Very rapid progress has been made along this new idea that demonstrated the validity of this new idea [103, 31, 26, 32, 33] and lead to measurements that progressively approached the accuracy of atomic fountains [30, 31, 25, 27, 104]. This scheme is applicable to alkaline-earth atoms (Mg, Ca, Sr) or atoms with similar structure like Yb or Hg [105, 106].

5. – Applications and prospects

Cold atoms have allowed a dramatic improvement of atomic clocks. In about 15 year of existence, atomic fountains have improved in such a way that their accuracy now reaches the low 10^{-16} range. This is 2 orders of magnitude better when compared to the previous generation of atomic clocks, using thermal beam of Cs atoms [10]. Fountains are now close to their ultimate performance: they operate with a frequency noise that is fundamentally limited by the quantum nature of the measurement and all frequency shifts are controlled at a level that is about 10^{-6} of the width of the atomic resonance. We have discussed briefly how it seems possible to go further by increasing the transition frequency up to optical frequencies. In a few years time frame, stabilities and accuracies at the 10^{-18} level are anticipated. Still, with the present level of performance, very interesting tests of fundamental physics can be performed.

5˙1. *Cold atom clocks in space.* – The first set of experiments can be performed by flying a cold atom clock in a spacecraft. A project lead of the European Agency, ACES([5]), has as a primary objective the goal to fly a cold atom clock in space [107]. The clock itself is being built by the French space agency, CNES, and the operation of the clock engineering model has been recently successfully demonstrated [108]. With a space clock of 10^{-16} accuracy, a new test of the gravitational red shift will be performed with a ppm uncertainty. This system will also allow clock comparisons at the global scale with 10^{-16} accuracy or better allowing for instance a direct mapping of the earth gravitational field by means of the gravitational redshift with a resolution corresponding to 1 meter of elevation. These comparisons will be a key element for the test of the stability of fundamental constants described below. Second-generation experiments with optical frequency standards are also already being investigated [109].

5˙2. *Fundamental physics.* – A large variety of fundamental tests can be performed with cold-atom clocks. For instance an improved test of Lorentz invariance has been performed by looking for a possible asymmetry of the first-order Zeeman effect [110]. Another example is a test of the stability of fundamental constants by the repeated comparison of clocks using different atoms or molecules. The central idea is that the frequency of each atomic or molecular transition specifically depends on fundamental constants like the fine-structure constant or the mass ratio of elementary particles [111-113]. Consequently, a variation of these "constants" (which would violate the equivalence principle) can be detected by the comparison of clocks based on different transitions. A first test involving Rb and Cs fountains started in the late 90's with a frequency resolution of about 10^{-15} per year [114]. With the progress of fountains and longer averaging time an improvement of this test by more than an order of magnitude is anticipated. With the advent of optical clocks, more and more different types of transitions involving an increasing number of different atoms will further enrich this test. A first step towards this goal involved clocks with an accuracy of about 10^{-14} [115-117]. This test will certainly be improved by several orders of magnitude.

REFERENCES

[1] Cohen-Tannoudj C., Chu S. and Phillips W., "Nobel Lectures," *Rev. Mod. Phys.*, **70** (1998) 685.

[2] Kasevich M., Riis E., Chu S. and de Voe R., "RF spectroscopy in an atomic fountain," *Phys. Rev. Lett.*, **63** (1989) 612.

[3] Clairon A., Salomon C., Guellati S. and Phillips W., "Ramsey resonance in a Zacharias fountain," *Europhys. Lett.*, **16** (1991) 165.

[4] Gibble K. and Chu S., "Future slow-atom frequency standards," *Metrologia*, **29** (1992) 201.

[5] Bize S. *et al.*, "Advances in atomic fountains," *C. R. Phys.*, **5** (2004) 829.

[5] Atomic Clock Ensemble in Space.

[6] Wynands R. and Weyers S., "Atomic fountain clocks," *Metrologia*, **42** (2005) S64.

[7] Bauch A. *et al.*, "Comparison between frequency standards in Europe and the USA at the 10^{-15} uncertainty level," *Metrologia*, **43** (2006) 109.

[8] Davidson N., Jin Lee H., Adams C. S., Kasevich M. and Chu S., "Long Atomic Coherence Times in an Optical Dipole Trap," *Phys. Rev. Lett.*, **74** (1995) 1311.

[9] Ferrari G., Poli N., Sorrentino F. and Tino G. M., "Long-Lived Bloch Oscillations with Bosonic Sr Atoms and Application to Gravity Measurement at the Micrometer Scale," *Phys. Rev. Lett.*, **97** (2006) 060402.

[10] Vanier J. and Audoin C., *The Quantum Physics of Atomic Frequency Standards* (Adam Hilger, Bristol and Philadelphia) 1989.

[11] Leibfried D., Blatt R., Monroe C. and Wineland D., "Quantum dynamics of single trapped ions," *Rev. Mod. Phys.*, **75** (2003) 281.

[12] Clairon A., Laurent P., Santarelli G., Ghezali S., Lea S. and Bahoura M., "A cesium fountain frequency standard: preliminary results," *IEEE Trans. Instrum. Meas.*, **44** (1995) 128.

[13] Weyers S., Schroder R. and Bauch A., "Performance of PTB's caesium fountain CSF 1," In *Proc. of the 2001 IEEE International Frequency Control Symposium* (2001).

[14] Jefferts S. *et al.*, "Accuracy evaluation of NIST-F1," *Metrologia*, **39** (2002) 321.

[15] Levi F., Lorini L., Calonico D. and Godone A., "IEN-CsF1 accuracy evaluation and two-way frequency comparison," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, **51** (2004) 1216.

[16] Kurosu T., Fukuyama Y., Koga Y. and Abe K., "Preliminary evaluation of the Cs atomic fountain frequency standard at NMIJ/AIST," *IEEE Trans. Instrum. Meas.*, **53** (2004) 466.

[17] Szymaniec K., Chalupczak W., Whibberley P., Lea S. and Henderson D., "Evaluation of the primary frequency standard NPL-CsF1," *Metrologia*, **42** (2005) 49.

[18] Li T. *et al.*, "NIM4 cesium fountain primary frequency standard: performance and evaluation," in *Proceedings of 2004 IEEE International Frequency Control Symposium* (2004).

[19] Hollberg L., Oates C. W., Wilpers G., Hoyt C. W., Barber Z. W., Diddams S. A., Oskay W. H. and Bergquist J. C., "Optical frequency/wavelength references," *J. Phys. B: At. Mol. Opt. Phys.*, **38** (2005) S469.

[20] Margolis H. S., Barwood G. P., Huang G., Klein H. A., Lea S. N., Szymaniec K. and Gill P., "Hertz-Level Measurement of the Optical Clock Frequency in a Single $^{88}\mathrm{Sr}^{+}$ Ion," *Science*, **306** (2004) 1355.

[21] Sterr U., Degenhardt C., Stoehr H., Lisdat C., Schnatz H., Helmcke J., Riehle F., Wilpers G., Oates C. and Hollberg L., "The optical calcium frequency standards of PTB and NIST," *C. R. Physique*, **5** (2004) 845.

[22] Dubé P., Madej A. A., Bernard J. E., Marmet L., Boulanger J.-S. and Cundy S., "Electric Quadrupole Shift Cancellation in Single-Ion Optical Frequency Standards," *Phys. Rev. Lett.*, **95** (2005) 033001.

[23] Peik E., Schneider T. and Tamm C., "Laser frequency stabilization to a single ion," *J. Phys. B: At. Mol. Opt. Phys.*, **39** (2006) 145.

[24] Oskay W.H. *et al.*, "Single-Atom Optical Clock with High Accuracy," *Phys. Rev. Lett.*, **97** (2006) 020801.

[25] Ludlow A. D., Boyd M. M., Zelevinsky T., Foreman S. M., Blatt S., Notcutt M., Ido T. and Ye J., "Systematic Study of the ^{87}Sr Clock Transition in an Optical Lattice," *Phys. Rev. Lett.*, **96** (2006) 033003.

[26] Barber Z., Hoyt C., Oates C., Hollberg L., Taichenachev A. and Yudin V. I., "Direct Excitation of the Forbidden Clock Transition in Neutral ^{174}Yb Atoms Confined to an Optical Lattice," *Phys. Rev. Lett.*, **96** (2006) 083002.

[27] Le Targat R., Baillard X., Fouché M., Brusch A., Tcherbakoff O., Rovera G. D. and Lemonde P., "Accurate Optical Lattice Clock with ^{87}Sr Atoms," *Phys. Rev. Lett.*, **97** (2006) 130801.

[28] Takamoto M., Hong F.-L., Higashi R., Fujii Y., Imae M. and Katori H., "Improved Frequency Measurement of a One-Dimensional Optical Lattice Clock with a Spin-Polarized Fermionic ^{87}Sr Isotope," *J. Phys. Soc. Jpn.*, **75** (2006) 104302.

[29] Katori H., Takamoto M., Pal'chikov V. G. and Ovsiannikov V. D., "Ultrastable optical clock with neutral atoms in an Engineered light shift Trap," *Phys. Rev. Lett.*, **91** (2003) 173005.

[30] Courtillot I., Quessada A., Kovacich R. P., Brusch A., Kolker D., Zondy J.-J., Rovera G. D. and Lemonde P., "Clock transition for a future frequency standard with trapped atoms," *Phys. Rev. A*, **68** (2003) 030501(R).

[31] Takamoto M., Hong F.-L., Higashi R. and Katori H., "An optical lattice clock," *Nature*, **435** (2005) 321.

[32] Brusch A., Le Targat R., Baillard X., Fouché M. and Lemonde P., "Hyperpolarizability Effects in a Sr Optical Lattice Clock," *Phys. Rev. Lett.*, **96** (2006) 103003.

[33] Boyd M. M., Zelevinsky T., Ludlow A. D., Foreman S. M., Blatt S., Ido T. and Ye J., "Optical Atomic Coherence at the 1-Second Time Scale," *Science*, **314** (2006) 1430.

[34] Berman R. P. (Editor), *Atom Interferometry* (Academic, London) 1997.

[35] Kasevich M. and Chu S., "Atomic interferometry using stimulated Raman transitions," *Phys. Rev. Lett.*, **67** (1991) 181.

[36] Gustavson T. L., Landragin A. and Kasevich M. A., "Rotation sensing with a dual atom-interferometer Sagnac gyroscope," *Class. Quantum Grav.*, **17** (2000) 2385.

[37] Peters A., Chung K. Y. and Chu S., "High-precision gravity measurements using atom interferometry," *Metrologia*, **38** (2001) 25.

[38] McGuirk J. M., Foster G. T., Fixler J. B., Snadden M. J. and Kasevich M. A., "Sensitive absolute-gravity gradiometry using atom interferometry," *Phys. Rev. A*, **65** (2002) 033608.

[39] Canuel B. *et al.*, "Six-Axis Inertial Sensor Using Cold-Atom Interferometry," *Phys. Rev. Lett.*, **97** (2006) 010402.

[40] Bertoldi A., Lamporesi G., Cacciapuoti L., de Angelis M., Fattori M., Petelski T., Peters A., Prevedelli M., Stuhler J. and Tino G., "Atom interferometry gravity-gradiometer for the determination of the Newtonian gravitational constant G," *Eur. Phys. J. D*, **40** (2006) 271.

[41] Fixler J. B., Foster G. T., McGuirk J. M. and Kasevich M. A., "Atom Interferometer Measurement of the Newtonian Constant of Gravity," *Science*, **315** (2007) 74.

[42] Wicht A., Hensley J. M., Sarajlic E. and Chu S., "A Preliminary Measurement of the Fine Structure Constant Based on Atom Interferometry," *Phys. Scr.*, **T102** (2002) 82.

[43] Cladé P., de Mirandes E., Cadoret M., Guellati-Khélifa S., Schwob C., Nez F., Julien L. and Biraben F., "Determination of the Fine Structure Constant Based on Bloch Oscillations of Ultracold Atoms in a Vertical Optical Lattice," *Phys. Rev. Lett.*, **96** (2006) 033001.

[44] Taylor B. N., "Determining the Avogadro Constant from Electrical Measurements," *Metrologia*, **31** (1994) 181.

[45] Fertig C. and Gibble K., "Measurement and Cancellation of the Cold Collision Frequency Shift in an ^{87}Rb Fountain Clock," *Phys. Rev. Lett.*, **85** (2000) 1622.

[46] Sortais Y., Bize S., Nicolas C., Clairon A., Salomon C. and Williams C., "Cold Collision Frequency Shifts in a ^{87}Rb Fountain," *Phys. Rev. Lett.*, **85** (2000) 3117.

[47] See Helmerson K. and Phillips W. D., this volume p. 211, and references therein.

[48] Metcalf H. J. and van der Straten P., *Laser Cooling and Trapping* (Springer, New-York) 1999.

[49] Ramsey N. F., "A Molecular Beam Resonance Method with Separated Oscillating Fields," *Phys. Rev.*, **78** (1950) 695.

[50] Ramsey N. F., *Molecular Beams* (Oxford University Press, Oxford) 1985.

[51] Ramsey N. F., "Experiments with Separated Oscillatory Fields and Hydrogen Masers," *Rev. Mod. Phys.*, **62** (1990) 541.

[52] Santarelli G., Laurent P., Lemonde P., Clairon A., Mann A. G., Chang S., Luiten A. N. and Salomon C., "Quantum Projection Noise in an Atomic Fountain: A High Stability Cesium Frequency Standard," *Phys. Rev. Lett.*, **82** (1999) 4619.

[53] Dick G., "Local oscillator induced instabilities in trapped ion frequency standards," in *Proceedings of Precise Time and Time Interval, Redondo Beach* (USNO) 1987, p. 133. `http://tycho.usno.navy.mil/ptti/index.html`.

[54] Santarelli G., Audoin C., Makdissi A., Laurent P., Dick G. J. and Clairon A., "Frequency Stability Degradation of an Oscillator Slaved to a Periodically Interrogated Atomic Resonator," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, **45** (1998) 887.

[55] Bize S., Sortais Y., Lemonde P., Zhang S., Laurent P., Santarelli G., Salomon C. and Clairon A., "Interrogation oscillator noise rejection in the comparison of atomic fountains," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, **47** (2000) 1253.

[56] Luiten A., Mann A., Costa M. and Blair D., "Power stabilized cryogenic sapphire oscillator," *IEEE Trans. Instrum. Meas.*, **44** (1995) 132.

[57] Chambon D., Bize S., Lours M., Narbonneau F., Marion H., Clairon A., Santarelli G., Luiten A. and Tobar M., "Design and realization of a flywheel oscillator for advanced time and frequency metrology," *Rev. Sci. Instrum.*, **76** (2005) 094704.

[58] Wineland D., Bollinger J., Itano W. and More F., "Spin squeezing and reduced quantum noise in spectroscopy," *Phys. Rev. A*, **46** (1992) R6797.

[59] Itano W., Bergquist J., Bollinger J., Gilligan J., Heinzen D., More F., Raizen M. and Wineland D., "Quantum projection noise: Population fluctuations in two-level systems," *Phys. Rev. A*, **47** (1993) 3554.

[60] Kuzmich A., Mølmer K. and Polzik E. S., "Spin Squeezing in an Ensemble of Atoms Illuminated with Squeezed Light," *Phys. Rev. Lett.*, **79** (1997) 4782.

[61] Turchette Q. A., Wood C. S., King B. E., Myatt C. J., Leibfried D., Itano W. M., Monroe C. and Wineland D. J., "Deterministic Entanglement of Two Trapped Ions," *Phys. Rev. Lett.*, **81** (1998) 3631.

[62] Wolf P. and Petit G., "Relativistic theory for clock syntonization and the realization of geocentric coordinate times.," *Astron. Astrophys.*, **304** (1995) 653.

[63] Itano W. M., Lewis L. L. and Wineland D. J., "Shift of $^2S_{1/2}$ hyperfine splittings due to blackbody radiation," *Phys. Rev. A*, **25** (1982) 1233.

[64] Simon E., Laurent P. and Clairon A., "Measurement of the Stark shift of the Cs hyperfine splitting in an atomic fountain," *Phys. Rev. A*, **57** (1998) 436.

[65] Beloy K., Safronova U. I. and Derevianko A., "High-Accuracy Calculation of the Blackbody Radiation Shift in the ^{133}Cs Primary Frequency Standard," *Phys. Rev. Lett.*, **97** (2006) 040801.

[66] Angstmann E. J., Dzuba V. A. and Flambaum V. V., "Frequency Shift of the Cesium Clock Transition due to Blackbody Radiation," *Phys. Rev. Lett.*, **97** (2006) 040802.

[67] Angstmann E. J., Dzuba V. A. and Flambaum V. V., "Frequency shift of hyperfine transitions due to blackbody radiation," *Phys. Rev. A*, **74** (2006) 023405.

[68] Bize S., "Private communication," 2006.

[69] Li R. and Gibble K., "Phase variations in microwave cavities for atomic clocks," *Metrologia*, **41** (2004) 376.

[70] Wolf P. and Bordé C. J., "Recoil Effects in Microwave Ramsey Spectroscopy," ArXiv:quant-ph/0403194 (2004).

[71] K. Gibble, "Difference between a Photon's Momentum and an Atom's Recoil," *Phys. Rev. Lett.*, **97** (2006) 073002.

[72] Gibble K. and Chu S., "A laser cooled Cs Frequency Standard and a Measurement of the Frequency Shift due to Ultra-cold Collisions," *Phys. Rev. Lett.*, **70** (1993) 1771.

[73] Ghezali S., Laurent P., Lea S. and Clairon A., "An experimental study of the spin-exchange frequency shift in a laser cooled cesium fountain frequency standard," *Europhys. Lett.*, **36** (1996) 25.

[74] Weiner J., Bagnato V. S., Zilio S. and Julienne P. S., "Experiments and theory in cold and ultracold collisions," *Rev. Mod. Phys.*, **71** (1999) 1.

[75] Pereira Dos Santos F., Marion H., Bize S., Sortais Y., Clairon A. and Salomon C., "Controlling the Cold Collision Shift in High Precision Atomic Interferometry," *Phys. Rev. Lett.*, **89** (2002) 233004.

[76] Messiah A., in *Quantum Mechanics* (1959) p. 637.

[77] Loy M. M. T., "Observation of Population Inversion by Optical Adiabatic Rapid Passage," *Phys. Rev. Lett.*, **32** (1974) 814.

[78] Bauch A. and Telle H. R., "Frequency standards and frequency measurement," *Rep. Prog. Phys.*, **65** (2002) 789.

[79] Young B., Cruz F., Itano W. and Bergquist J. C., "Visible Lasers with Subhertz Linewidths," *Phys. Rev. Lett.*, **82** (1999) 3799.

[80] Nazarova T., Riehle F. and Sterr U., "Vibration-insensitive reference cavity for an ultra-narrow-linewidth laser," *Appl. Phys. B*, **83** (2006) 531.

[81] Ludlow A. D., Huang X., Notcutt M., Zanon T., Foreman S. M., Boyd M. M., Blatt S. and Ye J., "Compact, thermal-noise-limited optical cavity for diode laser stabilization at 1×10^{-15}," ArXiv:physics/0610274 (2006).

[82] Quessada A., Kovacich R. P., Courtillot I., Clairon A., Santarelli G. and Lemonde P., "The Dick effect for an optical frequency standard," *J. Opt. B: Quantum Semiclassical Opt.*, **5** (2003) S150.

[83] Jennings D. A., Pollock C. R., Petersen F. R., Drullinger R. E., Evenson K. M., Wells J. S., Hall J. L. and Layer H. P., "Direct frequency measurement of the I_2-stabilized He-Ne 473 THz (633 nm) laser," *Opt. Lett.*, **8** (1983) 136.

[84] Clairon A., Dahmani B., Filimon A. and Rutman J., "Precise frequency measurements of CO_2-OsO_4 and HeNe-CH_4 stabilized lasers," *IEEE Trans. Instrum. Meas.*, **34** (1985) 265.

[85] Schnatz H., Lipphardt B., Helmcke J., Riehle F. and Zinner G., "First phase-coherent measurement of visible radiation," *Phys. Rev. Lett.*, **76** (1996) 18.

[86] Bernard J. E., Madej A. A., Marmet L., Whitford B. G., Siemsen K. J. and Cundy S., "Cs-Based Frequency Measurement of a Single, Trapped Ion Transition in the Visible Region of the Spectrum," *Phys. Rev. Lett.*, **82** (1999) 3228.

[87] See Udem Th. and Riehle F., this volume p. 317, and references therein.

[88] Hall J., Zhu M. and Buch P., "Prospects for using laser-prepared atomic fountains for optical frequency standards applications," *J. Opt. Soc. Am. B*, **6** (1989) 2194.

[89] Ido T., Loftus T. H., Boyd M. M., Ludlow A. D., Holman K. W. and Ye J., "Precision Spectroscopy and Density-Dependent Frequency Shifts in Ultracold Sr," *Phys. Rev. Lett.*, **94** (2005) 153001.

[90] Keupp J., Douillet A., Mehlstäubler T., Rehbein N., Rasel E. and Ertmer W., "A high-resolution Ramsey-Bordé spectrometer for optical clocks based on cold Mg atoms," *Eur. Phys. J. D*, **36** (2005) 289.

[91] Malossi N., Damkjaer S., Hansen P., Jacobsen L., Kindt L., Sauge S., Thomsen J., Cruz F., Allegrini M. and Arimondo E., "Two-photon cooling of magnesium atoms," *Phys. Rev. A*, **72** (2005) 051403.

[92] Degenhardt C. *et al.*, "Calcium optical frequency standard with ultracold atoms: Approaching 10^{-15} relative uncertainty," *Phys. Rev. A*, **72** (2005) 062111.

[93] Oates C. W., Wilpers G. and Hollberg L., "Observation of large atomic-recoil-induced asymmetries in cold atom spectroscopy," *Phys. Rev. A*, **71** (2005) 023404.

[94] Niering M. *et al.*, "Measurement of the Hydrogen 1S-2S Transition Frequency by Phase Coherent Comparison with a Microwave Cesium Fountain Clock," *Phys. Rev. Lett.*, **84** (2000) 5496.

[95] Uhlenberg G., Dirscherl J. and Walther H., "Magneto-opical trapping of silver atoms," *Phys. Rev. A*, **62** (2000) 063404.

[96] Badr T., Plimmer M., Juncar P., Himbert M., Silver J. and Rovera G., "Continuous-wave Doppler-free two-photon spectroscopy of the $4d^{10}5s\ ^2S_{1/2}$-$4d^95s^2\ ^2D_{3/2}$ transition in atomic silver," *Eur. Phys. J. D*, **31** (2004) 3.

[97] Degenhardt C., Binnewies T., Wilpers G., Sterr U., Riehle F., Lisdat C. and Tiemann E., "Photoassociation spectroscopy of cold calcium atoms," *Phys. Rev. A*, **67** (2003) 043408.

[98] Fortier T. M., Coq Y. L., Stalnaker J. E., Ortega D., Diddams S. A., Oates C. W. and Hollberg L., "Kilohertz-Resolution Spectroscopy of Cold Atoms with an Optical Frequency Comb," *Phys. Rev. Lett.*, **97** (2006) 163905.

[99] Dicke R. H., "The Effect of Collisions upon the Doppler Width of Spectral Lines," *Phys. Rev.*, **89** (1953) 472.

[100] Webster S. A., Taylor P., Roberts M., Barwood G. P., Blythe P. and Gill P., "A frequency standard using the $^2S_{1/2} - {}^2F_{7/2}$ octupole transition in ^{171}Yb$^+$," in *Sixth Symposium on frequency standards and metrology*, edited by Gill P. (World Scientific, Singapore) 2002.

[101] Schmidt P. O., Rosenband T., Langer C., Itano W. M., Bergquist J. C. and Wineland D. J., "Spectroscopy Using Quantum Logic," *Science*, **309** (2005) 749.

[102] Lemonde P. and Wolf P., "Optical lattice clock with atoms confined in a shallow trap," *Phys. Rev. A*, **72** (2005) 033409.

[103] Takamoto M. and Katori H., "Spectroscopy of the $^1S_0 - {}^3P_0$ clock transition of ^{87}Sr in an optical lattice," *Phys. Rev. Lett.*, **91** (2003) 223001.

[104] Boyd M. M., Ludlow A. D., Blatt S., Foreman S. M., Ido T., Zelevinsky T. and Ye J., "^{87}Sr lattice clock with inaccuracy below 10^{-15}," ArXiv:physics/0611067 (2006).

[105] Porsev S. G., Derevianko A. and Fortson E. N., "Possibility of an optical clock using the 6 $^1S_0 \rightarrow 6\ ^3P_0$ transition in ^{171,173}Yb atoms held in an optical lattice," *Phys. Rev. A*, **69** (2004) 021403.

[106] Ovsiannikov V., Pal'chikov V., Katori H. and Takamoto M., "Polarisation and dispersion properties of light shifts in highly stable optical frequency standards," *Quantum Electron.*, **36** (2006) 3.

[107] Salomon C. *et al.*, "Cold atoms in space and atomic clocks: ACES," *C. R. Acad. Sci.-Ser. IV*, **2** (2001) 1313.

[108] Laurent P. *et al.*, "Design of the cold atom PHARAO space clock and initial test results," *Appl. Phys. B*, **84** (2006) 683.

[109] Cacciapuoti L. *et al.*, In *Proc. of 1st ESA international workshop on optical clocks* (2005).

[110] Wolf P., Chapelet F., Bize S. and Clairon A., "Cold Atom Clock Test of Lorentz Invariance in the Matter Sector," *Phys. Rev. Lett.*, **96** (2006) 060801.

[111] Uzan J.-P., "The fundamental constants and their variation: observational and theoretical status," *Rev. Mod. Phys.*, **75** (2003) 403.

[112] Flambaum V. V., Leinweber D. B., Thomas A. W. and Young R. D., "Limits on variations of the quark masses, QCD scale, and fine structure constant," *Phys. Rev. D*, **69** (2004) 115006.

[113] Karshenboim S. G., "Precision physics of simple atoms: QED tests, nuclear structure and fundamental constants," *Phys. Rep.*, **422** (2005) 1.

[114] Marion H. *et al.*, "Search for variations of fundamental constants using Atomic fountain clocks," *Phys. Rev. Lett.*, **90** (2003) 150801.

[115] Bize S. *et al.*, "Testing the stability of fundamental constants with the ^{199}Hg$^+$ single-ion optical clock," *Phys. Rev. Lett.*, **90** (2003) 150802.

[116] Peik E., Lipphardt B., Schnatz H., Schneider T., Tamm C. and Karshenboim S. G., "Limit on the Present Temporal Variation of the Fine Structure Constant," *Phys. Rev. Lett.*, **93** (2004) 170801.

[117] Fischer M. *et al.*, "New Limits on the Drift of Fundamental Constants from Laboratory Measurements," *Phys. Rev. Lett.*, **92** (2004) 230802.

This page intentionally left blank

Atomic frequency standards, properties and applications

A. Bauch

Physikalisch-Technische Bundesanstalt - Braunschweig, Germany

1. – Introduction

Among all the standards used to implement the units of the International System of Units (SI) [1], atomic frequency standards have the best characteristics in terms of accuracy, precision and repeatability. This is due to the fact that the frequency of the standard's output signal is derived from an inherent property of free atoms. In the logic of this approach, all standards using the same atomic species should in principle deliver the same frequency. This is, admittedly, not the case since the technical capabilities to transfer atomic properties to that of macroscopic electric circuits are limited. The estimated deviation of the standard's frequency from its nominal value is expressed as the standard's uncertainty. Relative uncertainties approach one part in 10^{16} in these days for the most advanced devices. These "very best" standards are, however, not the main subject of this contribution. Instead, I wish to give a review of the function and properties of such atomic frequency standards (AFS) which are available as commercial products and thus are the basis for wide applications. My paper comes in the tradition of previous contributions at these Summer Schools during which in the past in particular Claude Audoin and Jacques Vanier have introduced this subject. They have later provided very detailed textbooks [2, 3] on the subject which can be recommended for further reading and for understanding many details which I have to skip in this contribution.

The accurate measurement of time and frequency is vital to the success of many fields of science and technology. Examples from atomic physics are atom-photon interactions, atomic collisions, and atomic interactions with static and dynamic electromagnetic fields.

Geodesy, radio-astronomy (very long baseline interferometry), and millisecond-pulsar timing rely strongly on the local availability of stable frequency standards and access to global uniform timescales. The same is true for the operation of satellite-based navigation systems. However, more common place applications, such as management of electric power networks and telecommunication networks, also require synchronization of local timing sources or syntonization of locally maintained frequency sources with national or international standards. Strictly speaking, synchronization (greek, $\chi\rho o\nu o\sigma$, time) of two clocks means setting the reading of one clock to that of the other, whereas syntonization (latin *tonus*...) means equalizing the rate between the two clocks. It is common practice to speak of synchronization even if this is not strictly correct. In almost all these fields AFS have played an important role since decades. Today, ten thousand rubidium AFS have been installed, several hundred commercial caesium atomic clocks are used in timing laboratories and in military and scientific institutes, and the number of hydrogen masers in operation surely exceeds one hundred.

A brief discussion of principles, operation and characterization of an AFS is given subsequently, followed by the description of the function of a caesium AFS. This sect. **3** contains the essence of an earlier paper [4]. Section **4** contains a quick glance at the development of primary clocks, fountain clocks and optical frequency standards. The latter are subject of separate contributions in these Proceedings. In sect. **5** I give a cursory overview of the function and performance of hydrogen masers, rubidium AFS, and modern gas cell AFS. Details can be found in various textbooks [2, 3, 5-7]. Some room is left in sect. **6** for dealing with applications. I will report on the use of AFS for the realization of International Atomic Time (TAI), in the European satellite navigation system Galileo, in sychronization of power networks, and in the quest whether fundamental constants are really constant. It is this wide spread of ambitions and requirements on accuracy, stability and reliability that makes the subject AFS so fascinating. I hope I can communicate this fascination to the reader.

2. – Atomic frequency standards: principle of operation and characterization of their performance

It is commonly assumed that atomic properties such as energy differences between atomic eigenstates and thus atomic transition frequencies are natural constants and do not depend on space and time (apart from relativistic effects). They are determined by fundamental constants which describe the interaction of elementary particles. A transition between two eigenstates differing in energy by ΔE is accompanied by absorption or emission of electromagnetic radiation of frequency $f_0 = \Delta E/h$ (h: Planck constant). The principle of operation of a *passive* AFS is illustrated in fig. 1 (left): A signal at a probing frequency f_p is used to interrogate the atomic resonator and the response of the latter is used to steer the quartz oscillator. In contrast, an *active* AFS, realized in the central part of an active hydrogen maser, emits radiation which is detected and which then steers a built-in electrical oscillator (fig. 1, right).

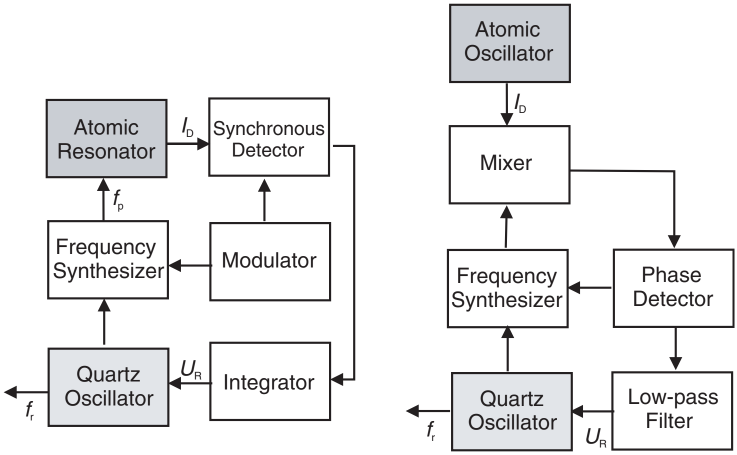


Fig. 1. – Schematic representation of a passive (left) and an active frequency standard (right); f_p characterizes the probing signal, f_r the output signal, I_D is the signal carrying the primary information from the atoms, and U_R is the control voltage to the quartz.

In general, the choice of the particular atomic transition to be used in an AFS is directed by certain requirements. These requirements refer to stability and accuracy, properties which will be discussed subsequently, but also practical constraints regarding the ease of manufacture, operation, maintenance, and the reliability. To begin with, consider the requirement to minimize random fluctuations of the output signal. This can be fulfilled if

I) the line width Γ of the resonance is small (Γ is expressed as angular frequency throughout the text);

II) the atomic resonance is observed with a high signal-to-noise ratio; this requires

III) that the interaction time T_i of the atomic absorber with the probing radiation is long;

IV) that (for passive AFS) sources of probing radiation exist which deliver a spectrally narrow radiation so that no technical broadening of the observed resonance curve occurs.

If the first three criteria are fulfilled, in principle a narrow atomic resonance can be observed. Consider atoms being irradiated with a monochromatic radiation of frequency f_p during a time interval T_i. If Γ is sufficiently small, $\Gamma/(2\pi) \ll 1/T_\mathrm{i}$, the observed line shape resembles the squared Fourier transform of the truncated sinusoidal waveform, and has a full width at half-maximum (FWHM) of approximately $1/T_\mathrm{i}$. In case of the active maser, III) translates into the need of a long undisturbed build-up of atomic coherence (see subsect. **5**˙2).

The second requirement is to minimize systematic shifts of the realized output frequency f_r from that of unperturbed atoms, which translates into

V) The energy difference of the atomic eigenstates should be insensitive to electric and magnetic fields.

VI) The velocity v of the probed atoms should be low.

I will come back to these requirements several times in the following. Now I briefly discuss the standard measures for the characterization of AFS in general. The term *frequency instability* describes the stochastic or environmentally induced fluctuations of the output frequency of a standard. The frequency instability can be expressed in the time domain as a function of the measurement time τ (averaging time) or in the frequency domain by the power spectral density. A review of the measures is included in [8,9]. Here I only introduce a widely accepted measure in the time domain for characterization of AFS output signals. It handles normalized frequency differences $y(\tau)$ of the realized frequency from its nominal value or from a suitable reference, averaged over τ. The two-sample standard deviation $\sigma_y(\tau)$, introduced by Allan [10], calculated according to

$$\sigma_y(\tau) = \left\{ \sum_{i=1}^{K-1} \left(y_{i+1}(\tau) - y_i(\tau)\right)^2 / (2K-2) \right\}^{1/2}, \tag{1}$$

is a useful measure of the relative frequency instability for an averaging time τ during the total measurement time $K \times \tau$, as long as $K \geq 10$ is valid. In a double logarithmic plot of $\sigma_y(\tau)$ *vs.* τ one can discriminate among some of the causes of instability in the clock signal because they lead to different slopes. If shot-noise of the detected atoms is the dominating noise source, the frequency noise is white and $\sigma_y(\tau)$ decreases like $\tau^{-1/2}$. In this case $\sigma_y(\tau)$ agrees with the classical standard deviation of the sample. Typically, however, one notices long-term effects due to colored noise processes which indicate that the parameters defining f_r are not sufficiently well controlled. The classical standard deviation would diverge with increasing τ in such a case whereas $\sigma_y(\tau)$ remains bounded. The slope in the log-log plot then changes to zero or even becomes positive. I will show examples for such behavior in the sections dealing with the various types of atomic frequency standards.

The observed $\sigma_y(\tau)$ can be related with operational parameters of a passive AFS through the expression

$$\sigma_y(\tau) = \frac{\eta}{Q \times (S/N)} \times \frac{1}{\sqrt{\tau/\mathrm{s}}}. \tag{2}$$

In eq. (2), η is a numerical factor of the order of unity, depending on the shape of the resonance line and of the method of frequency modulation to determine the line center. Q is the line quality factor (transition frequency/FWHM), and S/N is the signal-to-noise ratio for a 1 Hz detection bandwidth. Equation (2) reflects the requirements I)-IV) introduced before.

The term *accuracy* is generally used to express the agreement between the clock's output frequency and its nominal value conforming with the SI second definition. The manufacturers of commercial AFS state the accuracy as a range of clock frequencies to be expected, but usually without giving details about the causes of potential frequency deviations. A detailed *uncertainty* estimate deals with the quantitative knowledge of all effects which may entail that the output of the AFS does not reflect the transition frequency of unperturbed atoms. Such estimates are made for primary clocks, and examples of uncertainty budgets are given in subsect. **4**˙2. Given the lack of knowledge of the involved experimental parameters or the underlying mechanism, one estimates the *uncertainty* u due to individual effects. General rules how to do this can be found in an ISO Guide [11].

3. – Caesium clocks, classical and optically pumped

3˙1. *Caesium clocks with magnetic selection.* – Already in the early 1950s, the element caesium was identified as a very suitable candidate to fulfill many of the above-mentioned requirements. The isotope ^{133}Cs which, favorably, is the only stable isotope of this element has a nuclear spin of 7/2 and its ground state consists of two hyperfine level manifolds, designated by $F_g = 3$ and $F_g = 4$. The indices g and e are used to distinguish ground-state from excited-state energy sublevels. The ground-state manifolds comprise 7 and 9 Zeeman sublevels, respectively, as shown in fig. 2 as a function of the magnetic induction B. The energy levels relevant for discussion of optical pumping on the D_2 line are depicted in the right part of fig. 2.

The definition of the second [1] reads "*The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state*" and thus defines, in other words, that the reference transition between the $F_g = 4$ and $F_g = 3$ hyperfine ground-state energy levels occurs at a frequency of $f_0 = 9\ 192\ 631\ 770\,\mathrm{Hz}$. This definition must be understood as valid in an idealized environment with atoms at rest, without external fields present, and without disturbances due to the interaction process itself. Later I will discuss deviations from this conditions leading to the uncertainty with which the second can be realized with practical devices.

It was very important, particularly in the earlier days when the laser was still unknown, that magnetic selection of caesium atoms in the different hyperfine states was possible, and that, at the same time, efficient detection of the atoms using surface ionization could be performed. The deflection is caused by the tendency of the atoms to seek a state of low potential energy. Consequently, atoms having higher energy in high (low) magnetic field are deflected toward regions of low (high) magnetic field. State selection is thus accomplished by means of strong inhomogeneous magnetic fields (of the so-called polarizer) whose orientation is such as to force atoms in one group of levels of fig. 2 to the desired direction, whereas those in the other group are discarded. Several geometries have been realized. In commercial clocks dipole magnets are used for the purpose, which effect a deflection of the atomic beam whereas, *e.g.*, in PTB's primary clocks magnetic lenses focus or defocus the atoms. In fig. 3 the different types of magnetic deflection systems are depicted. Such magnetic lenses are also used in the hydrogen maser, see subsect. **5**˙2.

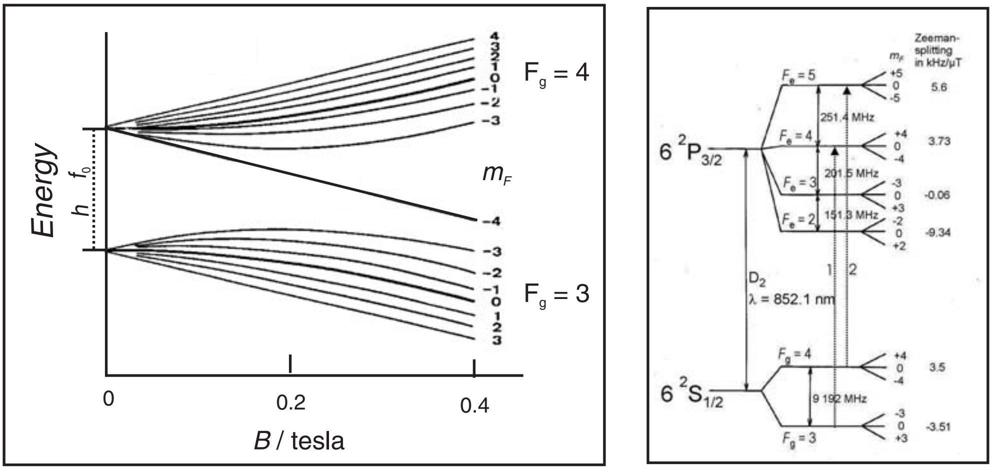


Fig. 2. – Left: ground-state energy level manifold of the caesium atom as a function of the magnetic induction B. Right: energy level diagram of the ^{133}Cs atom including the $^2P_{3/2}$ state. In high resolution, the two levels exhibit hyperfine structure (separations of sublevels not to scale). At small values of B the hyperfine levels exhibit a linear Zeeman shift proportional to m_{F} with the indicated frequency shift in kHz per μT of the $m_{\mathrm{F}} = +1$ level. The optical transition 1 is an example of a so-called pumping transition, used for state preparation. Transition 2 is a cyclic transition used for detection and for laser cooling of caesium atoms in an atomic fountain (subsect. **4**·2).

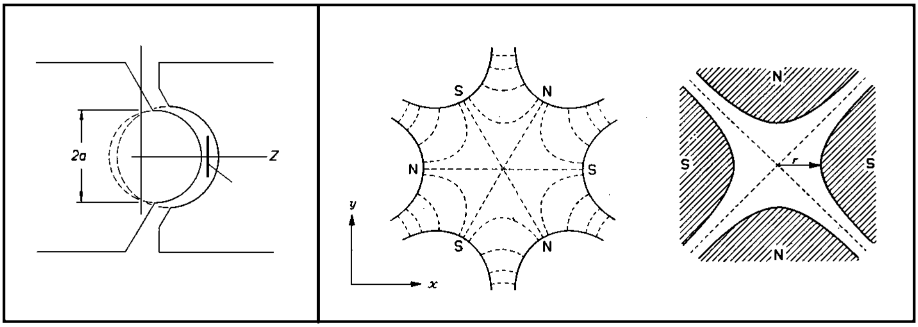


Fig. 3. – Geometry of the pole tips of permanent magnets used as state selectors in AFS. Left: dipole magnet, the atomic beam is initially collimated to a ribbon shape and the detector is made up of a ribbon-shaped filament. Right: four-poles and six-poles acting as magnetic lenses. Their use requires a tiny circular oven orifice at the focal point of the lens, a central "beam stop", and allows a small-area detector to be used.

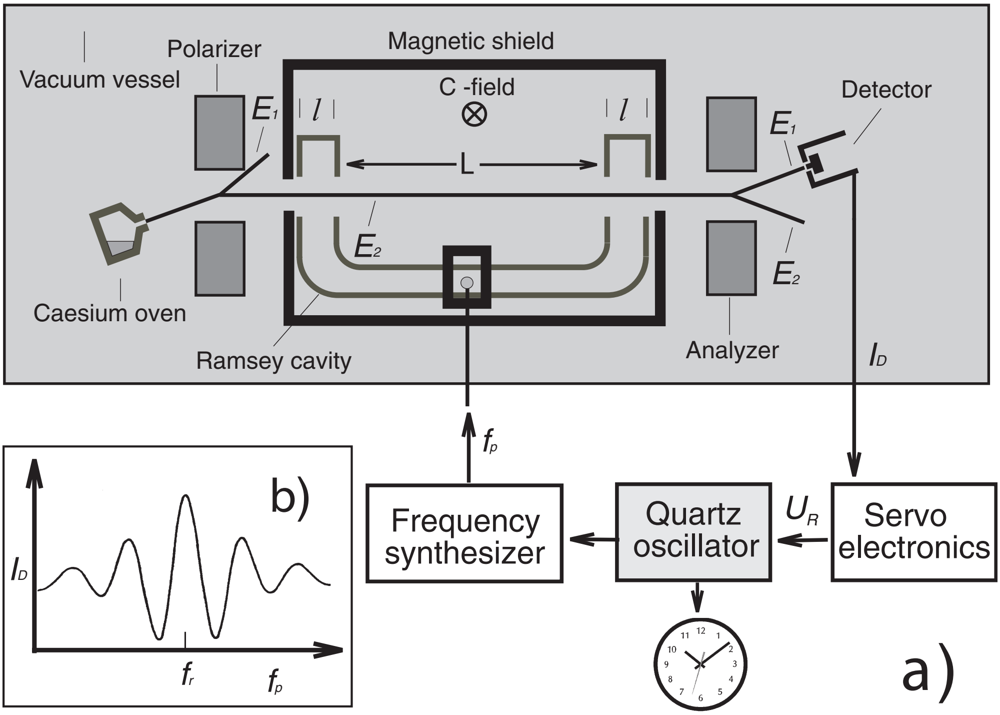


Fig. 4. – a) Schematic representation of a caesium AFS using magnetic state selection with dipole magnets; the grey-shaded rectangle is referred to as atomic resonator in fig. 1; E_1 and E_2 represent the two groups of energy states in fig. 2; the microwave magnetic field lines in the interaction region of length l and the static magentic field are perpendicular to the plane of paper. b) Schematic representation of the detected resonance signal when f_{p} is tuned around the realized resonance frequency f_{r}.

In fig. 4, the principle of a caesium atomic clock with state selection in dipole magnets is illustrated. A beam of atoms effuses from the oven and passes through the polarizer. Due to the relatively high vapor pressure at moderate temperatures T, intense thermal caesium atomic beams can be generated easily. In favor of VI), the mean thermal velocity, $v \approx (kT/M)^{1/2}$, where M is the atomic mass, is only about 200 m/s, leading to an interaction time T_{i} of a few milliseconds even in small structures. In a so-called Ramsey cavity, made up of a U-shaped waveguide, the atoms are irradiated twice with a standing microwave probing field of frequency f_{p}. Ramsey [12, 13] had shown that the effective interaction time T_{i}, as previously introduced, is equal to the time of flight between the two arms of the cavity. Transitions obeying the selection rules $\Delta F_g = \pm 1$, $\Delta m_{\mathrm{F}} = 0$ can be induced. The analyzer discriminates between atoms which have made a transition and those which have remained in the initial state and directs atoms in one of the states to a hot-wire detector. The caesium ionization energy is only 3.9 eV, thus enabling ionization on a surface with a high work function (*e.g.* on platinum) with 100% efficiency. The

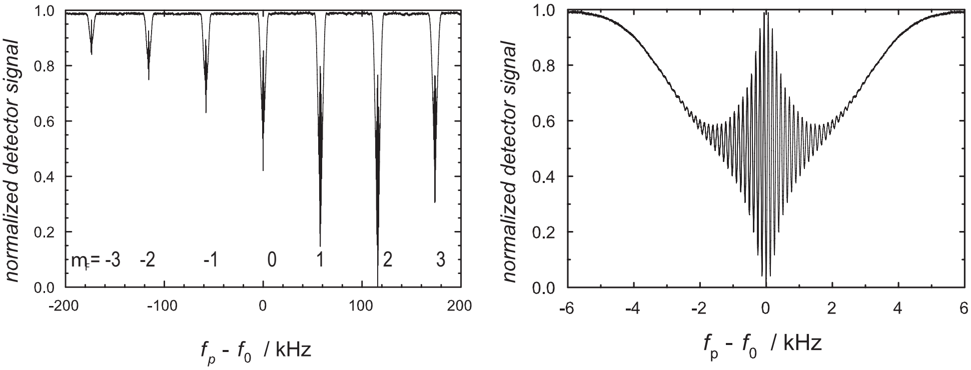


Fig. 5. – Record of the seven Zeeman components of the hyperfine transition $F_g = 4$ to $F_g = 3$, m_{F} (left) and in higher resolution that of the clock transition ($m_{\mathrm{F}} = 0$) with PTB's primary clock CS1 (right).

surface is heated so that the formed ions are evaporated. A secondary electron multiplier behind some ion optics (for acceleration, deflection and focusing) is often used to amplify the initial electric current of some pA to the range of μA for easier and faster processing. When f_{p} is tuned across f_r, I_D exhibits a resonance feature centered around f_{r}, which is shown schematically in fig. 4b. In clock operation, the probing frequency is modulated around a central value. By phase-sensitive detection of I_D and subsequent integration the control voltage, U_R is generated which tunes the voltage-controlled, temperature-stabilized quartz oscillator so that f_{p} and f_{r} agree on average.

Described by the selection rule $\Delta m_{\mathrm{F}} = 0$, seven microwave transitions exist which have a different frequency dependence on static magnetic fields. The atomic beam path is surrounded by a set of nested shields protecting against the ambient magnetic field. A coil inside the shield generates a weak static magnetic field (traditionally named C-field) which shifts the transition frequencies for $m_{\mathrm{F}} \neq 0$ proportional to m_{F} linearly by

$$f_z = 7\,\mathrm{kHz} \times (B/\mu\mathrm{T}). \tag{3}$$

It separates the resonance frequencies of the individual transitions by typically several hundred times the widths of the central fringe. In clock operation, the quartz oscillator can thus be stabilized on the well-resolved *clock transition*. This transition of interest is that from level $F_g = 4$, $m_{\mathrm{F}} = 0$ to level $F_g = 3$, $m_{\mathrm{F}} = 0$ or vice versa. Its frequency has only a weak dependence on the magnetic inductance,

$$f_{\mathrm{r}}(B) = f_0 + 0.0427\,\mathrm{Hz} \times (B/\mu\mathrm{T})^2. \tag{4}$$

In fig. 5, the spectrum of all resonances and in higher resolution the resonance pattern of the clock transition as recorded with PTB's primary clock CS1 are depicted. The frequency separation f_z between neighboring lines is a convenient measure of the mean magnetic field along the atomic trajectory, which is needed for determination of the correction according to eq. (4). Digital control systems in modern clocks allow periodic measurement of f_z during normal clock operation, and the magnetic field is either stabilized (*e.g.* in the 5071 commercial clock, subsect. **3**˙5) or the frequency correction is applied numerically in the frequency synthesizer (see fig. 1) based on real-time data.

3˙2. *Caesium AFS with optical pumping.* – It is possible to replace the twofold magnetic selection by interaction with laser fields at a wavelength, for example, of the caesium $D2$-line ($\lambda = 852.1\,\mathrm{nm}$). The relevant energy levels are depicted in fig. 2 (right). Excitation of the transition $F_g = 3[F_g = 4] \rightarrow F_e = 3$ or 4 pumps the atoms into the hyperfine state $F_g = 4[F_g = 3]$ and allows state preparation of the atoms in one of the manifolds of the hyperfine m_{F} substates. Excitation of the so-called cycling transition $F_g = 4 \rightarrow F_e = 5$ yields a larger number of fluorescence photons per atom as quantum-mechanical selection rules allow radiative decay from the excited state only back to the initial ground state. It is therefore common to use this transition in the detection process. The same optical transitions also form the basis for laser-cooling, state-preparation and detection in a fountain clock, see subsect. **4**˙2. The use of optical pumping in caesium clocks in general was motivated by the observation that a more efficient utilization of the atoms could be obtained and some systematic effects common in clocks with magnetic selection (see next section) could be avoided.

Optical pumping and detection has been employed in three primary clocks (see subsect. **4**˙1 for further explanations), NIST-7 of the National Institute of Standards and Technology, USA [14], the French JPO (Jet de Pompage Optique) [15] and NICT-01 of the Japanese National Institute of Communications Technology [16]. The technique was also studied in a small AFS at the French Laboratoire d'Horloge Atomique [17,18]. Based on these studies, a commercial AFS has been under development for quite some time [19], and industrial teams in Europe have been invited by the European Space Agency (ESA) to continue work in this direction.

3˙3. *Systematic frequency offsets.* – The magnetic field-related shift is just one example of a perturbation of the atomic energy states or of the detected line shape due to preparation, probing and detection of the atoms. I continue with a brief list of further systematic shifts which occur in a caesium clock, see [2] for a comprehensive treatment.

The total population and also the atomic velocity distribution behind the polarizer is different for the magnetic sublevels m_{F}, which partially explains the shape of the resonance lines in fig. 5. This effect is prominent in case of magnetic state selection but may to a lesser extent also exist in standards using optical pumping because of imperfect adjustment of the polarization of the laser fields. Such a population imbalance may lead to various kinds of frequency shifts, like Rabi pulling [20], Ramsey pulling [21], and those caused by Majorana transitions [22]. These shifts were found to impair the performance of commercial clocks, particularly those of older design. In fact, they can in general be suppressed easier when optical pumping is used.

Several velocity-dependent shifts are known. Intuitively one thinks here of the well-known linear Doppler effect which, however, is very effectively suppressed by the geometry of the Ramsey cavity. But a small effect, linear in v, due to imperfections in the cavity symmetry, usually named *cavity phase shift*, remains. It has determined the uncertainty of thermal beam primary clocks in most cases. The quadratic Doppler effect,

$$\delta f_d = \frac{f_{\mathrm{r}}}{2} \times \left(\frac{v}{c}\right)^2, \tag{5}$$

is a consequence of the relativistic time dilation and therefore of the atomic velocity v in the beam. To give examples, $\delta f_d/f_0$ amounts to about 2×10^{-13} for a Maxwell-Boltzmann velocity distribution in an atomic beam from an oven at 400 K and to 5×10^{-14} in PTB CS1 and CS2 due to the velocity selective effect of the magnetic lenses (subsect. **4**˙1).

Interaction of the caesium atoms with the electric field of thermal radiation emitted from the vacuum enclosure reduces the clock frequency by about 1.6×10^{-4} Hz at room temperature [23-25]. The correctness of the underlying atomic constants was unquestioned for many years. So all groups operating primary clocks derived the uncertainty of the required correction mainly from the limited ability to specify the radiation field as that of a perfect black body at a well-defined temperature. Stimulated by theoretical and experimental findings of the team at INRiM, the Italian national metrology institute [26, 27], the matter has gained renewed interest. A study of Ulzega *et al.* [28] corroborated that the numerical value given above is too large by about 15%. More recently, Beloy *et al.* [29] and Angstmann *et al.* [30] pointed to the deficiencies in that recent theoretical work and confirmed that current practice of applying the corrections is justified. This is just an example of the necessary interplay of theoretical and experimental work that has often taken place during the years and brought the uncertainty to the low levels which now have been reached.

3˙4. *Development of the accuracy with time.* – Historically [31], the measurement result of the hyperfine splitting frequency in caesium atoms made between 1956 and 1958 with the caesium AFS of the National Physical Laboratory (NPL) in the UK with reference to the ephemeris second [32] became the basis for the definition of the second in the International System of Units SI [1]. The uncertainty of the NPL device was estimated at 1 part in 10^{10} so that the measurement uncertainty was entirely that of the astronomical determination of the duration of the ephemeris second. Over the years, the insight in the causes of perturbations in caesium AFS has improved and this, together with technological advances, has entailed a reduction of the clock uncertainties by almost three orders of magnitude for the best commercial devices, another factor of 20 for the best thermal beam primary clocks (subsect. **4**˙1), and another factor of 10 in the best fountain clocks (subsect. **4**˙2).

3˙5. *Commercial caesium clocks.* – Following the principles and ideas explained in the previous sections, caesium clocks have been produced commercially since the late 1950s, starting with the so-called *Atomichron* of the National Company [33]. In de-

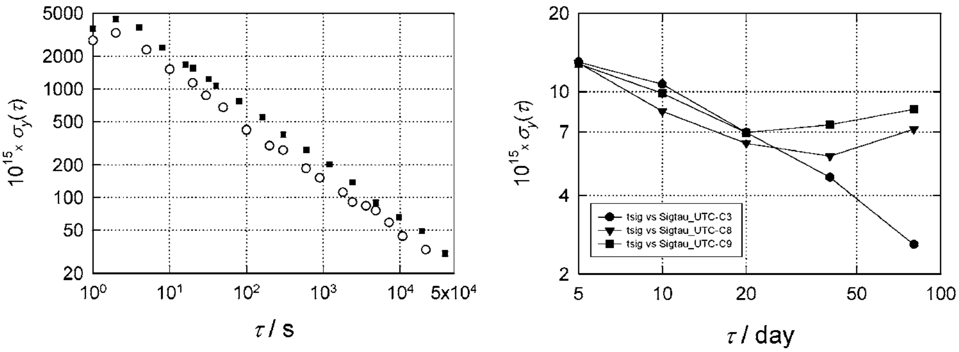


Fig. 6. – Left: relative frequency instability of the output of a 5071 high-performance caesium clock measured with reference to a hydrogen maser, data taken with a new (open circles) and an almost run out caesium beam tube (black squares). Right: relative frequency instability of three clocks of that kind operated at PTB during 2 years; the common reference is UTC.

signing commercial clocks, a compromise between weight, volume, power consumption, and performance and cost is unavoidable. Several manufacturers have participated in this business over the years, but today essentially all production of instruments for civil use is in hands of Symmetricom (`http://www.symmetricom.com`). 14 years since its first appearance on the market the model 5071, initially developed by Hewlett-Packard, later produced under the brand Agilent and now produced by Symmetricom, has gained widest acceptance in the timing community. Standard and high-performance versions of this clock are offered. Part of the improved specifications of the latter are due to a larger atomic flux employed which entails a larger S/N ratio. The price to be paid (literally) is a faster depletion of the caesium reservoir and thus the more frequent need to replace the clocks' beam tube.

I give examples of the observed frequency instability of clocks of this type operated at PTB in laboratory environment. In fig. 6 (left) records of the short-term frequency instability of one clock are shown, one record taken at an early time of operation and the other one a few months before the beam tube ran out of caesium. The new tube yielded a slightly more stable signal, the specifications, however, were fulfilled at all times. The long-term behavior of three clocks is illustrated in fig. 6 (right), using data over two years. The common reference in this stability plot is Coordinated Universal Time (UTC) whose generation is described in [34]. The clocks' frequency instability is governed by white frequency noise $\sigma_y(\tau) \propto \tau^{-1/2}$ at moderate averaging times, in the very long term a so-called flicker floor [9] is noticed for two units. The model 5071 accuracy is specified as 5×10^{-13} under good laboratory conditions. The ensemble of such clocks operated in national timing institutes [34] exhibits a scatter of monthly clock rates mostly between $\pm 2 \times 10^{-13}$ with the ensemble mean rate in very close agreement with that of TAI, which is determined with the help of primary clocks, subject of the next section. More data on long-term behavior of larger clock ensembles is included in refs. [4, 35].

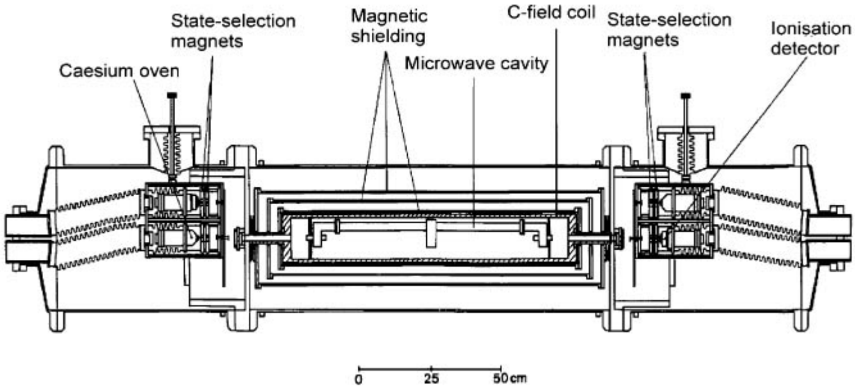


Fig. 7. – Vertical section of the vacuum chamber of PTB's primary clock CS2. The constituent elements are explained in subsect. **3**˙1. An oven and detector at each end allow alternate operation of atomic beams in opposite directions without disturbing the vacuum conditions.

4. – More accurate, more complex

4˙1. *Primary clocks.* – Several national metrology institutes during the past decades have competed in building the most accurate *primary clock*. Their basic principle is alike to that shown as fig. 4 —magnets may be replaced by laser interaction zones— but in details their construction allows the determination of all frequency shifting effects as detailed in subsect. **3**˙3 with high accuracy at all times. In 2006 operation of four primary clocks with a thermal atomic beam has been reported. These are the French JPO (Jet de Pompage Optique) [15] and NICT-01 [16], both using optical state selection and detection, and CS1 and CS2 of the Physikalisch-Technische Bundesanstalt [36]. I have been responsible for the operation of PTB's clocks for the last 15 years and will just give a few details. Magnetic lenses (see fig. 3) are used for state selection and velocity selection in these devices. As a consequence, the mean atomic velocity is more than a factor of two lower than in an effusive thermal beam from the same source, and atomic velocities are confined in a narrow interval around the mean velocity. A sketch of the CS2 design is shown in fig. 7. The CS2 interaction time amounts to about 10 ms, and the resonance curve of 60 Hz width, which looks similar to that shown in fig. 5, is recorded with a S/N of 1000 in 1 Hz bandwidth. The relative frequency instability amounts thus to 14 parts in 10^{15} at $\tau = 1$ day. From continuous comparisons over years with CS1 we could verify shot-noise limited performance for both clocks for averaging times up to 300 days. The largest systematic frequency shift, due to the magnetic field in the interaction region (eq. (4)) amounts to 2.92 Hz, but like all other shifts it can be so well determined that already in the mid-1980s u(CS2) could be estimated as 15×10^{-15} [37] and could be somewhat improved later. In subsect. **4**˙2 I present a table with uncertainty contributions for two of the primary clocks mentioned here and for two fountain clocks.

To conclude this section, primary clocks with a thermal atomic beam currently permit the realization of the SI second with a relative uncertainty of the order of one part in

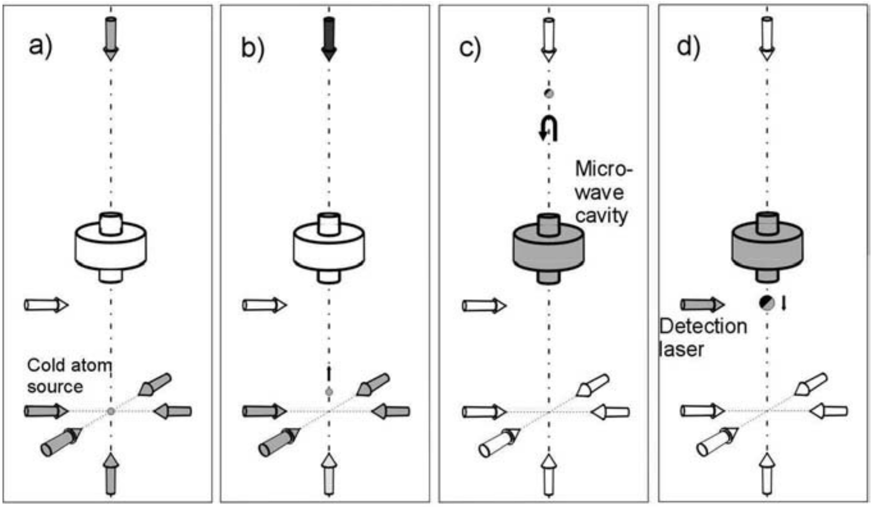


Fig. 8. – Sequence of operation of a fountain frequency standard, illustrated in a time sequence from left to right; laser beams are indicated by arrows (white if they are blocked). a) A cloud of cold atoms is loaded. b) The cloud is launched by de-tuning of the frequency of the vertical lasers. c) The cloud with an initially small volume and high density expands during ballistic flight. d) After the second passage of the atoms through the microwave cavity the state population is probed by laser irradiation and fluorescence detection.

10^{14}. They still are important in the realization of the International Atomic Time TAI (see [34]). Criteria I), III)-V) of sect. **2** are essentially fulfilled for these clocks. A substantial step toward a better fulfillment of criteria II) and VI) has been possible only by using laser-cooled atoms in a fountain design.

4˙2. *Atomic fountains.* – Already in the mid-1950s Jerome Zacharias at the Massachusetts Institute of Technology, USA, envisaged a *fountain* device in which sufficiently slow atoms from a thermal atom source, directed vertically upwards, would stop and descend under the action of gravity and would provide an interaction time T_i of more than one second [33]. While this attempt was not successful because of the lack of the very slow atoms due to collisions, laser cooling has enabled the realization of such an atomic fountain clock in the last decade. My contribution would be incomplete without at least a short description of fountain operation and properties. Detailed account is given in other contributions in these Proceedings.

Laser cooling was stressed as the key to the success of the fountain concept in the 1997 Nobel lectures [38-40]. Trapping and cooling of atoms are in general connected with strong shifts of the hyperfine energy states, and precision spectroscopy becomes impossible —unless it is performed in the dark, *i.e.* without laser fields being present. But without cooling applied, the cloud of atoms expands corresponding to its initial temperature. Instead of just letting the cloud fall under the action of gravity, in a fountain it is launched upwards with a velocity v_s and the microwave excitation is performed during the ballistic flight, as illustrated in fig. 8. The atoms come to rest under the action of

Table I. – *Uncertainty budgets of two primary clocks with a thermal beam (see subsect.* **4**˙1*) and the two fountain clocks FO2 of LNE-SYRTE, Paris* [42]*, and CSF1 of PTB* [41]*. The FO2 values were reported for May 2005, those of CSF1 were valid in Dec. 2003. All values are given in parts in* 10^{16}.

Cause of frequency shift	JPO	CS2	FO2	CSF1
Quadratic Zeeman effect	13	50	0.1	0.1
AC Stark effect caused by thermal radiation	5	10	0.6	2
AC Stark effect caused by fluorescence	24	0	<0.1	<0.1
Cavity phase	40	100	3	5
Quadratic Doppler effect	26	10	<0.1	<0.1
Asymmetric population of the m_F levels	23	1	1	1
Cold collisions	0	0	2.1	7
Electronics, microwave leakage	20	30	4.3	2
Combined uncertainty	63	120	6.1	9

gravity at a height of $H = v_s^2/(2g)$. If H is adjusted to a height of the fountain setup of 1 m ($v_s = 4.4\,\mathrm{m/s}$), then the total time of flight, back to the starting point, is about 0.9 s. On their way the atoms interact twice with the field sustained in the microwave cavity, on their way up and then on their way down. The interaction time T_i becomes typically 0.5 s. The detection comprises the determination of the number of atoms in both hyperfine levels, $F_g = 3$ and $F_g = 4$. Without cooling to kinetic energies equivalent to μK temperature, the thermal expansion of the atomic cloud would be so large that the fraction of detected atoms would be too small to obtain a useful signal-to-noise ratio. During clock operation, the transition probability is determined by changing the frequency f_p from cycle to cycle alternately on either side of the central fringe where the sensitivity to changes of f_p relative to f_r has its maximum. The difference of successive measurements is numerically integrated and a control signal is derived to steer the quartz or to adjust the output frequency of a synthesizer included in the generation of the microwave signal.

The current status of fountain properties was recently discussed in [41, 42]. In table I, the contributions to the uncertainty resulting from the most significant causes of frequency shift for two primary clocks with a thermal beam and two fountains are combined. "0" indicates that the effect is non-existent, "< 0.1" indicates that the component is by all means smaller than 10^{-17}. It is common practice to calculate the combined uncertainty u as the root-sum-of squares of the individual components, assuming that they are linearly independent. The duration of the realized second intervals should thus agree with the defined duration within $\pm u$ with about two-thirds probability. Because of the reduced linewidth, somewhat below 1 Hz, and the reduced atomic velocity, some systematic frequency shifting effects are reduced by orders of magnitude for fountains. A new effect needs consideration since ultracold atoms are used and the cross-section for frequency-shifting collisions among caesium atoms becomes much larger than that of thermal caesium atoms. The collisional shift is proportional to the density of atoms in the cloud and depends on the details of the state in which the atomic cloud has been initially prepared. As obvious from table I, the collisional frequency shift currently leads to a significant uncertainty contribution and it is still a subject of detailed studies.

Rubidium, ^{87}Rb, has been identified as another candidate species for use in a fountain. Its atomic level structure is quite similar to that of caesium (sect. **5**) and laser cooling and manipulation are technically feasible as well. The cross-section for phase-changing collisions among cold ^{87}Rb atoms is so much smaller than among cold caesium atoms that it can be hardly measured [42]. The corresponding uncertainty contribution for a Rb fountain is essentially zero. The rubidium hyperfine transition was thus recently recommended as a *secondary representation of the second*, as will be briefly explained subsequently.

4·3. *Optical clocks.* – Frequency standards in the infrared and in the visible range of the electromagnetic spectrum have been developed and used since decades. The most prominent use has been as wavelength standards in practical length metrology and for the realization of the meter. Since the new definition of the meter became effective in 1983, this SI unit should be realized according to a *mise en pratique* [43]. High-resolution laser spectroscopy both in the context of fundamental research and in the context of developing an all-optical clock has been pursued by many groups and with great success. The 2005 Nobel Prize for two prominent members of this community is further proof thereof. The performance of a laser as a frequency standard, when stabilized to a suitable narrow optical transition between a metastable state and the ground state, may surpass that of a frequency standard in the radio-frequency region. Reasons for that seem quite obvious. If a (fictive) optical transition at $\lambda = 0.5\,\mu$m, $\nu = 600\,$THz, is recorded with the same line width as a microwave transition at 10 GHz, the line Q which dictates the achievable frequency instability (eq. (2)) increases by a factor of 60000. This paves the way to explore systematically frequency-shifting effects in a conveniently short measurement time. The relative magnitude of systematic energy level shifts is currently about equal but is envisaged to become substantially lower than possibly feasible in microwave clocks.

The most precise values for optical transition frequencies have up to now been obtained for forbidden transitions in trapped and laser-cooled single ions. Ions can be localized in radiofrequency ion traps while only minimally perturbing the internal level structure. Combined with laser cooling, it is possible to reach the confinement to a spatial region smaller than the clock transition wavelength, the so-called *Lamb-Dicke regime*, so that the linear Doppler shift is eliminated and thus no line broadening occurs.

The recent improvement in optical frequency standards is visualized against that for the caesium clocks in fig. 9. During the past several years, optical standards based on cold atoms and cold trapped ions have benefitted significantly from the advent of the self-referenced femtosecond frequency comb [44]. Measurements of optical frequencies will very soon be limited by the capabilities to realize the unit of time and optical standards cannot demonstrate their full performance. They may just be found highly reproducible by comparing two optical systems directly. This situation has led to the concept of establishing *secondary representations of the second.* This formal procedure is

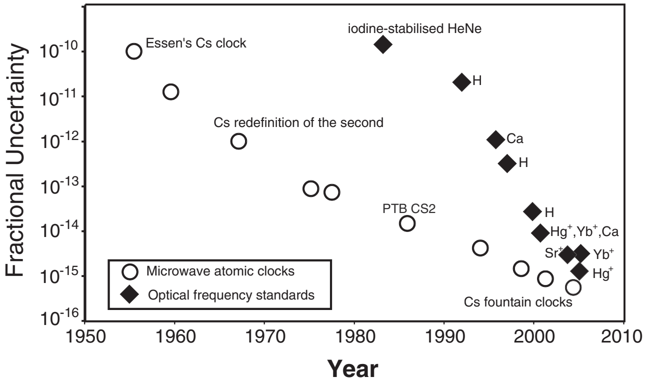


Fig. 9. – Progress in AFS uncertainty over time, adapted from fig. 1 in [45]; open symbols representing the reported uncertainty of primary clocks, full symbols that of absolute optical frequency measurements.

intended to stimulate the detailed evaluation of reproducibility of such new standards at the highest level and should significantly aid the process of comparing different standards in the preparation of a potential future redefinition of the second.

A Joint Working Group of the Consultative Committees for Length (CCL) and Time and Frequency (CCTF) was established. In 2005 it recommended that the hyperfine microwave transition in ^{87}Rb mentioned in the previous section and three optical transitions in trapped single ions of the species ^{199}Hg$^+$, ^{88}Sr$^+$, and ^{171}Yb$^+$ should be considered as secondary representations of the seconds. A description of the function of this JWG was presented in [45]. The contributions on optical frequency standards in these Proceedings and [7, 46-48] describe the properties and performance of the contemporary optical frequency standards.

5. – A survey on other AFS in practical use

5·1. *General remarks.* – The ground-state hyperfine transitions of the elements caesium, hydrogen, and rubidium (^{87}Rb) serve as references in today's commercially available AFS. I have described the function and property of the caesium clock in some detail, what follows is a cursory description of the others. More details can be found in [2, 3, 7]. The energy level manifold in the ground state is created in all three atoms by the magnetic interaction of the single outer electron with the nucleus with nuclear spin I. In table II, I summarize the relevant parameters of the three elements. A similar energy level configuration can be found in some singly ionized atoms, like ^{199}Hg$^+$ and ^{171}Yb$^+$. AFS using the ground-state hyperfine transitions in these and some other ions have been realized as laboratory standards [7]. I exclude this subject from my contribution since there has not much activity been reported recently —with one exception. Development of the microwave ion trap standard is currently pursued in the context of the NASA deep space tracking network timing system [49] and future space clocks [50].

Table II. – *Properties of the* H, Rb [42]*, and* Cs *atoms relevant for their ground-state hyperfine splitting frequencies.*

Atom	Atomic mass	I	F	ν_0/ Hz
H	1	1/2	0;1	1 420 405 751.770(3)
Rb	87	3/2	1;2	6 834 682 610.904 322 6 (16)
Cs	133	7/2	3;4	9 192 631 770

Many of the systematic effects enlisted in table I (lines 1, 2, 3, 5, 6, 8) are common to all AFS mentioned here, and similar measures are taken to circumvent them. Additional effects, sometimes actually dominating the uncertainty, have to be considered depending on the function and construction of the devices.

5˙2. *The hydrogen maser.* – The ground-state hyperfine splitting of the hydrogen atom corresponds to a transition line at a frequency of 1.4 GHz. Research at Harvard University in the 1950s proved that a frequency standard based on this transition could not be built alike to the caesium clock, but required a different realization concept. In the active maser, stimulated emission inside a high-Q cavity is used to detect the atomic transition [51, 13]. In the passive maser, the transition is probed by injecting radiation into the cavity and observing the effect of the atoms on the radiation field in the cavity [52, 2]. A brief description of both variants follows.

5˙2.1. The active hydrogen maser. A schematic diagram of the hydrogen maser is shown in fig. 10. The maser operates on the $F = 1$, $m_F = 0$ to $F = 0$, $m_F = 0$ transition. The two energy levels involved have been designated by E_1 and E_2 in fig. 10 for simplicity.

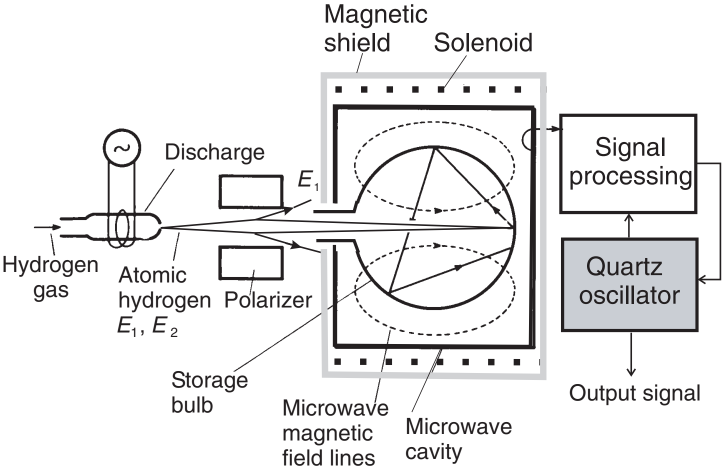


Fig. 10. – Schematic diagram of an active hydrogen maser.

The transition —like the clock transition in the caesium clock— is dependent on the magnetic-field strength only in second order. The maser works as follows. Molecular hydrogen is introduced in a dissociator consisting of an electrical discharge in a glass enclosure. The dissociation of the molecules takes place with a relatively high efficiency and a beam of atomic hydrogen is formed with the help of a small collimator. The beam is directed along the axis of a six-pole magnet (see fig. 3) in which atoms in the $F = 1$, $m_{\rm F} = 0$ and 1 are deflected towards the symmetry axis of the magnet, whereas the others are deflected away from the axis. The magnet thus focuses atoms in the upper hyperfine states into the entrance of the storage bulb. The bulb is placed inside a microwave cavity resonating at the hyperfine frequency and having the geometry which enables the highest possible Q (minimal energy dissipation in the walls). It is made of fused silica in order to reduce microwave losses and preserve the high-quality factor of the cavity. The inner surface of the bulb is coated with Teflon$^{\rm TM}$ that prevents recombination of the atoms into molecular hydrogen and relaxation of the atoms into the ground state. In a bulb having a 15 cm diameter, the lifetime of an atom in one particular level may be of the order of one second. In a cavity at ambient temperature, a weak microwave field is always present. Atoms which have entered the storage bulb in the $F = 1$, $m_{\rm F} = 0$ level, emit their energy through the process of stimulated emission of radiation. The radiation emitted is added in phase to the existing radiation, a process that results in amplification, explaining the acronym MASER (Microwave Amplification by Stimulated Emission of Radiation). The energy provided by the atoms is dissipated in the walls of the cavity, but part of the generated field is extracted via the coupling loop. If the losses are small and the relaxation time sufficiently long, a continuous oscillation at the hydrogen transition frequency occurs. The output power is of the order of 10^{-13} Watt. The signal coupled from the maser cavity is processed for phase locking a quartz crystal oscillator at a nominal frequency of 10 MHz to the maser signal, as was shown in more detail earlier in fig. 1 (right).

The solenoid shown in fig. 10 provides the axis of quantization to the atomic ensemble. The clock frequency originates from a $\Delta m_{\rm F} = 0$ transition, so quantum-mechanical selection rules require that the static magnetic field and the microwave field be parallel. This is the situation shown in fig. 10, the storage bulb confines the atoms to a region in which the static field and the microwave magnetic field are mostly parallel. The storage bulb and cavity are surrounded by concentric magnetic shields to reduce the influence of fluctuations of the ambient magnetic field. Often active field compensation is added to reduce the dependence of the output frequency on the ambient field.

This and other frequency shifting effects are common both to the active and the passive maser. The maser frequency is sensitive to a small extent to the cavity tuning which is a function of its dimensions. For this reason, the temperature of the ensemble is generally regulated to a high degree. Temperature stabilization also helps to stabilize the rather large second-order Doppler effect, $\delta\nu_{\rm D}/\nu_0 = -3kT/(2Mc^2)$, which amounts to several parts in 10^{11}. The linear Doppler effect is suppressed by the confinement of the atoms within a dimension smaller than the wavelength of the microwave radiation.

Another very substantial frequency shift is caused by the collisions of the atoms with the walls of the storage bulb, as well amounting to a few parts in 10^{11}. The shift is proportional to the inverse diameter of the storage bulb, and experimental studies tried to determine the shift using bulbs of variable diameter and extrapolation to infinite diameter. The difficulties involved in reproducing the bulb coating, however, limited the achievable uncertainty to a few parts in 10^{12}, and this is finally the uncertainty for the hydrogen hyperfine splitting frequency given in table II.

5˙2.2. The passive hydrogen maser. The component which dictates the size of the H-maser is the cavity resonator. A cavity sustaining the 1.4 GHz radiation with a minimum of electrical losses is roughly 30 cm in diameter. The size of this resonator can be reduced by operating it in a mode different from the TE011 mode (fig. 10). The resonator can also be loaded with dielectric material (*e.g.* sapphire). In this case the cavity Q is generally too low to achieve continuous oscillation. In order to relax the requirements on the quality factor of the cavity, it is possible to operate the maser in the so-called passive mode. The system may be operated with two coupling loops, one being used to inject a microwave signal close to the hyperfine frequency, while the other is used to detect the amplified signal. Another approach consists in using a circulator with a single coupling loop. All other elements of the passive maser are essentially the same as in the active model.

5˙2.3. Properties of hydrogen masers. In the early years of making AFS, the hydrogen maser won advanced laurels because of the excellent frequency stability that could be obtained. It turned out, however, that the systematic effects are so large and difficult to control that low uncertainties in reproducing the unperturbed hydrogen transition frequency could not be obtained. It became clear that the definition of the second was better based on the frequency of the caesium hyperfine transition as smaller uncertainties could be expected. The maser has thus remained the AFS of choice when a low-frequency instability is required: The active maser is by far the most stable AFS commercially available. Passive hydrogen masers have been produced in considerable quantitites by two Russian firms, KVARZ and VREMYA-CH, both located in Nizhny Novgorod. Active masers are currently produced commercially by the same Russian firms, by Symmetricom (USA) and T4Science in Switzerland, the latter firm carrying forward the long tradition of maser making at the Neuchâtel Observatory. In fig. 11, I compile the frequency instability specifications for the two types of masers. Specifications for longer averaging times are not available. The most advanced active masers exhibit a long-term frequency drift in the low 10^{-16} per day, in models without built-in control of the cavity resonance frequency and in passive masers the drift is larger by a factor of 10 typically. Nevertheless, the passive maser could be considered an alternative (regarding cost and size) for the commercial caesium clock in all applications where the long-term stability is of minor importance. According to current plans, two passive hydrogen masers will be operated on board of the satellites of the European satellite navigation system Galileo [53, 54].

5˙3. *Rubidium gas cell frequency standards.* – Techniques involving optically pumped atomic vapor in closed cells have been successfully used in numerous applications where the requirements regarding frequency instability in the short and long term cannot be fulfilled with quartz oscillators, but where constrainst on space, power consumption, or

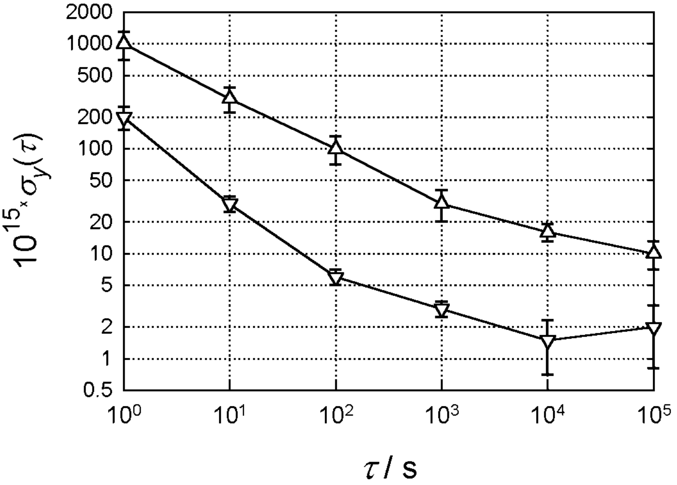


Fig. 11. – Specifications of the relative frequency instability of active (▽) and passive (△) hydrogen masers. The "error bars" reflect the spread of specifications of the models on the market in mid 2006.

cost prevent the use of other AFS. The rubidium gas cell frequency standard illustrated in fig. 12 is an example. The reference transition used is the ^{87}Rb hyperfine transition as tabulated in table II. The energy level scheme is similar to that of the caesium atom, the multiplicity of the magnetic sublevels is reduced due to the smaller nuclear spin. The simplified level scheme is given as inset a) in fig. 12. Development of this kind of frequency standard started in the late 1950s in the USA, and the first commercial model became available in the 1960s [55].

The heart of the Rb AFS is the absorption cell containing a vapor of ^{87}Rb and a proper buffer gas. Optical excitation on the D_2 line to the $^2P_{3/2}$ level requires a radiation source at 780 nm wavelength. Atoms excited to the P states in the cell relax to both levels $F_g = 1$ and $F_g = 2$ of the ground state either by spontaneous emission or by collisions with the buffer gas. Collisional relaxation is typical for gas cell standards and does not occur in the AFS dealt with so far. The light spectrum necessary to obtain population inversion is generated with a ^{87}Rb lamp and a filter containing ^{85}Rb vapour. The purpose of the filter is to clean the spectrum of the lamp from the lines at those frequencies corresponding to the transitions from the upper hyperfine level ($F_g = 2$) of the ground state to the P states in ^{87}Rb. In some buffer gases at a suitable pressure, the relaxation takes place essentially through collisions, and no radiation is emitted. This is important since fluorescence otherwise produced would optically pump from both hyperfine levels of the ground state and would reduce the signal contrast. Nitrogen was found to be a very effective buffer gas in this regard. Favourably, the buffer gas also reduces the diffusion velocity of the atoms to a few cm/s. This serves to reduce the Doppler shift and to narrow the resonance line. The effect of wall collisions is also reduced thereby.

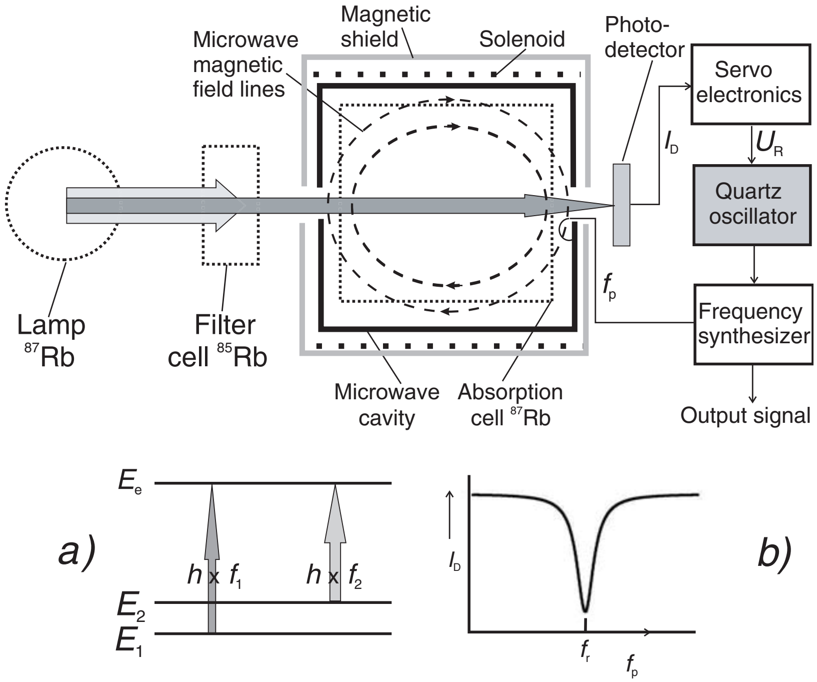


Fig. 12. – Schematic diagram of a Rb gas cell AFS; inset a) explains the two components of radiation for optical pumping, of which the component with frequency f_2 is blocked in the filter cell; inset b) shows the Lorentz-shaped absorption line observed when the probing frequency is swept over the resonance at f_r.

In a typical arrangement, as shown as fig. 12, the absorption cell is placed inside a low-Q microwave cavity having a resonance frequency equal to the hyperfine frequency of the ^{87}Rb atoms. The light transmitted through the cell is measured by means of a photodetector. Optical pumping of the atoms causes the cell to become transparent to the incident radiation since atoms are pumped out of the absorbing state. If microwave energy is fed to the cavity at the hyperfine frequency of the rubidium atoms, a field is created inside the cavity, which may stimulate transitions from the level $F_g = 2$ to $F_g = 1$. This results in a decrease of light intensity at the photodetector if the microwave field is resonant with the hyperfine transition. By sweeping the microwave frequency slowly across the resonance, a signal is observed as sketched in the inset b) of fig. 12. The signal can be used to control the frequency of the quartz oscillator according to the scheme of fig. 1 (left). The shape of the signal is a Lorentzian line broadened and shifted by several mechanisms such as buffer gas collisions, spin exchange interactions, optical pumping and saturation caused by the microwave excitation. The multitude of frequency shifting effects prevents the Rb AFS to be an accurate clock in a strict sense. The output signal at 10 MHz is calibrated with respect to some reference source, and the internal frequency synthesizer (see fig. 1, left) is adjusted accordingly. In the long term, the frequency will then typically drift away by a few parts in 10^{11} per month, relatively. The causes thereof are changes in the spectrum of the light passing through the gas cell and changes of pressure and composition of the buffer gas. The short-term frequency instability, however, is still quite favourable, and the best available commercial Rb AFS are more stable than some commercial Cs AFS for averaging times up to 10^4 seconds. This combination of properties calls for disciplining the Rb AFS through an external reference with a long time constant. Signals of the Global Positioning System (GPS), but also of long-wave standard frequency transmitters, like DCF77 in Europe, are well suited for the purpose.

In the Rb AFS, the spectral lamp can be replaced with a laser diode of the proper wavelength as light source [56, 57]. Such diodes are available at the required $D1$ and $D2$ wavelengths for Rb (780 and 794 nm) as well as for Cs (852 and 894 nm). Use of laser diodes, having much narrower line spectra than provided by spectral lamps, could improve the efficiency of the optical pumping. The background light is much reduced and a substantial gain in signal-to-noise ratio seems feasible. Up to now, however, the issue of reliability has prevented the commercial usage of laser-pumped Rb AFS.

5˙4. *Current trend in gas cell AFS*. – In this contribution I can only touch upon the two directions of development going on in the field. The Rb AFS can be made rather compact, but still usage in hand-held devices is essentially excluded. Miniaturization of AFS has been discussed since about a decade, and several research groups work on the so-called *clock on the chip* [58]. Very compact AFS could become useful in receivers for navigation systems signals and for secure communications.

Another vivid subject is the use of the coherence in an atomic sample of Rb or Cs created by irradiating it with two laser fields separated in frequency exactly by the hyperfine splitting. The sample of atoms is then excited into a non-absorbing state, called a *dark state*, and *Coherent Population Trapping* (CPT) has become the general term for this technique. AFS can be realized in several ways including CPT, and two recent reviews provide a deeper insight [59, 60]. The aim is to obtain Cs AFS performance —or even better— at reduced size, mass, and power consumption. Similar as I said before, it appears to be a great challenge to achieve comparably reliable performance and ease of maintenance as that obtained from lamp-pumped Rb gas cell standards or conventional commercial caesium clocks.

6. – Applications

6˙1. *A short note on the realization of TAI*. – The realization of the International Atomic Time TAI is in the responsibility of the Bureau International des Poids et Mesures (BIPM) and described in detail in [34]. Here I only recall the importance of the atomic clocks in that activity. The BIPM Time Section collects and processes time comparison data [9] obtained using different techniques from about 60 timing centers worldwide and data from some 300 atomic clocks operated in these centers. Commercial caesium clocks

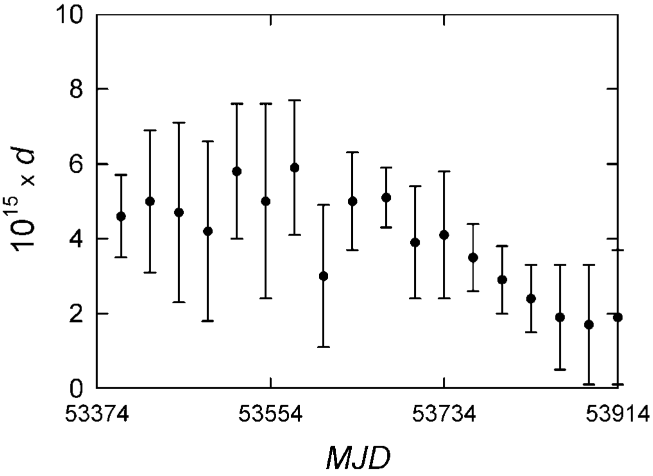


Fig. 13. – Relative departure d of the TAI scale unit from the SI second as provided by primary clocks and fountains during 18 months including June 2006. Each point represents the average value over the predeeding 30-day period. Source: BIPM Circular T, Section **4**.

to the largest part, about 60 hydrogen masers, and very few primary clocks form the clock ensemble. In a first step, a free atomic time scale, EAL (Echelle Atomique Libre), is produced using an iterative algorithm. The algorithm was designed to achieve a good long-term stability and high reliability of EAL. The duration of the scale unit of EAL is then determined with reference to primary clocks and fountain clocks which realize the SI second with a specified uncertainty. During recent years, data were available from four clocks with a thermal beam and from six fountains. A linear function of time is added to the EAL and the new scale is named TAI. The slope of this function is chosen so that the scale unit of TAI approaches the SI second as realized by the ensemble of primary clocks. In fig. 13, the estimated relative departure d of the scale unit of TAI from the SI second during 18 months is depicted. The recent discussion on the correct way to account for the frequency shift due to thermal radiation (see subsect. **3**˙3) was indeed very relevant since possibly all data points would have to be shifted by about 1 part in 10^{15}. Independent thereof, the accuracy and stability of the full clock ensemble (including masers and primary clocks) is effectively transferred to TAI and makes it an excellent reference for continuous time and frequency in scientific applications.

6˙2. *Timing aspects of the future European satellite navigation system Galileo.* – Galileo is Europe's contribution to the Global Navigation Satellite System (GNSS). The project is currently in preparation of the In-Orbit-Validation (IOV) phase, scheduled for 2008/2009. At that time four Galileo satellites will be in orbit and the ground infrastructure necessary to verify the system will have been installed. The timing system is at the heart of any GNSS. It comprises the on-board-clocks, the infrastructure on ground, and the time comparison equipment to link the various elements. Galileo's internal time reference, Galileo System Time (GST), will be realized in a Precise Timing Facility and, actually, two such facilities shall be available already at IOV. The PTF will provide GST as a physical signal with properties defined so that the navigation function of Galileo can be fulfilled according to the mission requirements.

The primary use of any GNSS relies on determining the time of arrival of satellite signals with reference to the receiver clock. Knowing the satellites' positions and also the relation between the individual satellites' clocks and the GNSS system time allows the four unknowns —three receiver coordinates and receiver clock time with respect to the GNSS system time— to be determined from four satellite observations. It is important to understand that all satellite messages are predictions. To give an example: From previous observations at monitoring sites, the current time difference between the clock time of space vehicle i, T_i, and the system time T_S are predicted in the form of a quadratic function for the current epoch t,

$$(T_i - T_S) = T_{0i} + T_{1i} \times (t - t_0) + T_{2i} \times (t - t_0)^2 , \tag{6}$$

in which the three coefficients T_{0i}, T_{1i}, and T_{2i} have to be updated regularly. It is thus required that space clocks and ground clocks generating the system time are embedded in a coordinated system of monitoring stations and uplink stations, the latter providing information generated on ground to the satellites. A prerequisite is the stability (and thus predictability) of ground and space clocks over the prediction times $t - t_0$ typically occurring. The Galileo timing system has recently been described in [61]. According to the currently published plans, each Galileo spacecraft will be equipped with two passive hydrogen masers and two Rb AFS [54]. So in total 60 of each of these clocks will be deployed in space by 2010 (or somewhat later) when the system will have reached its full operational capability. On ground, in each of the two PTFs two active hydrogen masers and four high-performance caesium clocks will be operated.

Each GNSS can also be used as a time distribution system. To facilitate this function, the rate of GPS time is maintained in close agreement with that of the reference time scale realized at the United States Naval Observatory, UTC(USNO), such that GPS Time minus UTC(USNO) equals to an integer offset of (currently) -14 s and another small offset of not more than a few ten nanoseconds. A similar function as the USNO will be fulfilled by the Galileo Time Service Provider (GTSP) for Galileo [62]. It will ensure that GST is steered towards UTC with the help of an intermediate time scale produced as the composite of about 30 atomic clocks operated in European timing institutes. The signals received from Galileo satellites will thus allow direct reference to a prediction of UTC. GST shall not deviate from UTC modulo 1 s by more than 50 ns in 95% of the time, and the prediction of the difference shall have an uncertainty of less than 28 ns (2σ). The cooperation of the various elements involved requires state-of-the art time transfer to be performed. Two-way Satellite Time and Frequency Transfer (TWSTFT) via a geostationary telecommunication satellite has been defined as the primary method [9].

6˙3. *Synchronization of the networks for electrical power distribution.* – In 2003, U.S. newspapers reported that U.S. economic losses due to unreliable electric power were 1% of the GDP, or 10^{10} \$ per year. Some readers will remember the fatal power outages that occurred in California and at the U.S. East Coast during recent years. Tracing of such power outages in extended networks requires clocks synchronized to a common time reference, preferentially UTC, to better than 1 ms at network nodes. Disciplined quartz and rubidium AFS have been deployed in large numbers at the nodes of the power grids. Some European network operators have even acquired Cs clocks providing autonomous frequency references for their network monitoring. In Europe, the members of the Union for the Coordination of Transmission of Electricity (UCTE) [63], agree on standards for network and supply properties, including synchronization issues. As of 2004, UCTE comprises 34 operators in 23 European countries, serving 500 million customers with an average annual electric energy of 2300 TWh. The power plant in Laufenburg (CH) is operated as the 50 Hz *frequency controller.* Within the UCTE service area, the network frequency 50 Hz is stabilized with very high accuracy. Under undisturbed conditions, the network frequency must be maintained within strict limits in order to ensure the full and rapid deployment of control facilities in response to a disturbance. Deviations from the set point frequency are unavoidable since they reflect the imbalance of production and consumption of electricity. The accuracy of frequency measurement required for the control process is 1–1.5 mHz. By integration of the network frequency over time, a network time is generated which must not deviate from UTC by more than 30 s. The Laufenburg control centre is responsible for the calculation of the network time and the organization of its correction. Correction involves the setting of the set point frequency at 49.99 Hz or 50.01 Hz, depending upon the direction of correction, for periods of one or more calendar days.

6˙4. *The search for the variability of the fundamental constants.* – Back to the basics. Section **2** had started saying that atomic properties such as energy differences between atomic eigenstates are natural constants, being determined by fundamental constants which describe the interaction of elementary particles.The question whether such fundamental constants are really *constants* or whether they may show temporal variations within the evolution of the universe was raised as early as 1937 by Dirac. This question has gained renewed interest today in the context of the search for a unified theory of the fundamental interactions and in cosmological models like inflation, and has been the subject of at least two recent conferences, the proceedings of which can be recommended for further reading [64, 65]. It has been *inter alia* proposed to search for variations of dimensionless quantities like Sommerfeld's fine-structure constant $\alpha \approx 1/137$. Precision laboratory experiments can look for non-zero temporal derivatives of α over times of say a few years based on precision comparisons of atomic and molecular frequency standards [66]. As the transition frequencies in different classes of transitions depend differently on α and on the Rydberg constant Ry, measuring the frequency of an electronic transition repeatedly with a caesium AFS may prove whether the product of some power of α and of the ^{133}Cs nuclear magnetic moment is constant or not. Clearly, commercial caesium AFS are not accurate and stable enough for the purpose, which explains that such kind of studies was in fact stimulated by the advent of caesium fountain clocks. In one particular study, the frequency of the ^{87}Rb hyperfine splitting realised in a rubidium fountain was measured repeatedly over a couple of years which provided a limit to the variation of *alpha* of [42]. Optical frequencies of the following transitions have been measured in SI hertz over a time interval of several years: $^2S_{1/2} \rightarrow {}^2D_{5/2}$ in Hg^+ studied at NIST (Boulder) [48,67], $^1S \rightarrow {}^2S$ in H from MPQ (Garching) [68], and $^2S_{1/2} \rightarrow {}^2D_{3/2}$ in Yb^+ from PTB [69]. All measured frequency drift rates were consistent with zero. The corresponding analyses were published in 2004 [68, 69], where the PTB group obtained the most stringent upper limits for variations of α and Ry, based on the most recent data from the Yb^+ ion [69]

$$\frac{\mathrm{d}\ln\alpha}{\mathrm{d}t} = (-0.3 \pm 2.0) \times 10^{-15}\,\mathrm{yr}^{-1}, \tag{7}$$

$$\frac{\mathrm{d}\ln Ry}{\mathrm{d}t} = (-1.5 \pm 3.2) \times 10^{-15}\,\mathrm{yr}^{-1}. \tag{8}$$

I have included this section as a demonstration how the research into atomic frequency standards can help improving our understanding of the laws of physics in general, and, assuming that the rate of improvement as depicted in fig. 9 indeed continues, there is hope that some of the open questions in cosmology can be answered based on small-scale laboratory experiments in the future.

7. – Conclusion

The development of AFS started in the late fourties of the last century, and since the first working caesium AFS became available in 1955 a variety of systems has been developed and studied. This is an ongoing process, in particular since the usage of optical frequency standards has been facilitated by the invention of new devices for measurement of optical frequencies. This subject is well covered in other contributions in these Proceedings. I have concentrated on describing the function and performance of established standards, mostly available as commercial products. Even here the progress has been breathtaking, and many applications make use of what has been achieved.

Regarding the new type of primary clocks and optical frequency standards, I notice still a gap to be bridged before they will find wider applications outside scientific institutes. While being involved in projects related to the development of the Galileo timing infrastructure, I was faced with substantially more requirements regarding reliability, availability, maintenance, mean time between failures, and mean time to repair than requirements dealing with accuracy and stability. As far as reported openly, also telecommunication systems operating agencies strive for simple to use and inexpensive components in their systems rather than for sophisticated high-performance equipment, and this attitude, driven by economical constraints, might not even compromise the further improvement of capabilities of such systems. This should, however, not discourage the persecution of further research in AFS, since nature has left us with many open questions which eventually can be answered —at least partially— with the help of such instruments.

8. – Disclaimer

The Physikalisch-Technische Bundesanstalt as a matter of policy does not endorse any commercial product. The mentioning of brands and individual models is for information only. All information provided is based on publicly available material or data taken at PTB and it will help the reader to make comparison with own observations.

* * *

This text was written in the stimulating environment of PTB's Time and Frequency Department in which research and development of atomic frequency standards is up to today a prominent part of the daily work. Many colleagues have influenced my understanding of the field. I owe thanks to E. Peik for support at various steps of producing this text.

REFERENCES

[1] *Le Système international d'unités (SI)/The International System of Units (SI)*, 8th edition (Bureau international des poids et mesures) 2006.

[2] Vanier J. and Audoin C., *The Quantum Physics of Atomic Frequency Standards* (Adam Hilger, Bristol) 1989.

[3] Audoin C. and Guinot B., *The Measurement of Time* (Cambridge University Press, Cambridge) 2001.

[4] Bauch A., *Meas. Sci. Technol.*, **14** (2003) 1159.

[5] Gerber E. A and Ballato A., *Precision Frequency Control* (Academic Press, Orlando FL) 1984.

[6] Major F. G., *The Quantum Beat* (Springer Verlag, New York) 1998.

[7] Riehle F., *Frequency Standards, Basics and Applications* (Wiley-VCH Verlag, Weinheim) 2004.

[8] International Telecommunication Union, *Handbook Selection and Use of Precise Frequency and Time Systems* (ITU, Geneva) 1997.

[9] Levin J., *Rev. Sci. Instrum.*, **70** (1999) 2567.

[10] Allan D. W., *Proc. IEEE*, **54** (1966) 221.

[11] International Organization for Standardization, *Guide to the Expression of Uncertainty in Measurement* (Geneva) 1993.

[12] Ramsey N. F., *Phys. Rev.*, **78** (1950) 695.

[13] Ramsey N. F., *Rev. Mod. Phys.*, **62** (1990) 541.

[14] Shirley J. H., Lee W. D. and Drullinger R. E., *Metrologia*, **38** (2001) 427.

[15] Makdissi A. and de Clercq E., *Metrologia*, **38** (2001) 409.

[16] Hasegawa A. *et al.*, *Metrologia*, **41** (2004) 257.

[17] Petit P. *et al.*, *Proc. 6th EFTF, Noordwijk, the Netherlands* (ESA Publication Division, ESA SP340) 1992, p. 83.

[18] Boussert B., Cérez P., Chassagne L. and Theobald G., *Proc. 11th EFTF, Neuchâtel, Switzerland* (FSRM) 1997, p. 58.

[19] Baldy M. L., *Proc. 28th Annual PTTI Systems and Application Mtg.*, (1996) 281.

[20] De Marchi A., Rovera G. D. and Premoli A., *Metrologia*, **20** (1984) 37.

[21] Cutler L., Flory C., Giffard R. P. and De Marchi A., *J. Appl. Phys.*, **69** (1991) 2780.

[22] Bauch A. and Schröder R., *Ann. Phys. (Leipzig)*, **2** (1993) 421.

[23] Itano W. M., Lewis L. L. and Wineland D. W., *Phys. Rev. A*, **25** (1982) 1233.

[24] Bauch A. and Schröder R., *Phys. Rev. Lett.*, **78** (1997) 622.

[25] Simon E., Laurent Ph. and Clairon A., *Phys. Rev. A*, **57** (1998) 436.

[26] Levi F., Calonico D., Lorini L., Micalizio S. and Godone A., *Phys. Rev. A*, **70** (033412) 2004.

[27] Micalizio S., Godone A., Calonico D., Levi F. and Lorini L., *Phys. Rev. A*, **69** (053401) 2004.

[28] Ulzega S., Hofer A., Moroshkin P. and Weis A., e-print: physics/0604233.

[29] Beloy K., Safranova U. I. and Derevianko A., *Phys. Rev. Lett.*, **97** (040801) 2006.

[30] Angstman E. J., Dzuba V. A. and Flambaum V. V., *Phys. Rev. Lett.*, **97** (040802) 2006.

[31] Nelson R. A. *et al.*, *Metrologia*, **38** (2001) 509.

[32] Markowitz W., Hall R. G., Essen L. and Parry J. V. L., *Phys. Rev*, **1** (1956) 105.

[33] Forman P., *Proc. IEEE*, **73** (1985) 1181.

[34] See Arias E. F., this volume p. 367.

[35] Kusters J. A., Cutler L. S. and Powers E. D., *Proc. 1999 Joint Mtg. EFTF and the IEEE Int. Freq. Control Symp., Besançon, France* (IEEE) 1999, p. 159.

[36] Bauch A., *Metrologia*, **42** (2005) S43.

[37] Bauch A. *et al.*, *IEEE Trans. Instrum. Meas.*, **36** (1987) 613.

[38] Chu S., *Rev. Mod. Phys*, **70** (1998) 685.

[39] Cohen-Tannoudji C., *Rev. Mod. Phys*, **70** (1998) 707.

[40] Phillips W. D., *Rev. Mod. Phys*, **70** (1998) 721.

[41] Wynands R. and Weyers S., *Metrologia*, **42** (2005) S64.

[42] Bize S. *et al.*, *J. Phys. B: At. Mol. Opt. Phys.*, **35** (2005) S449.

[43] Quinn T. J., *Metrologia*, **40** (2003) 103.

[44] Udem Th., Holzwarth R. and Hänsch, *Nature*, **416** (2002) 233.

[45] Gill P. and Riehle F., *Proc. 20th EFTF, Braunschweig, Germany* (BTP) 2006, p. 282.

[46] Madej A. A. and Bernard J. E., in *Frequency Measurement and Control: Advanced Techniques and Future Trends*, edited by Luiten A. N., *Springer Topics Appl. Res.*, Vol. **79** (Springer, Berlin, Heidelberg) 2000, p. 153.

[47] Gill P. *et al.*, *Meas. Sci. Technol.*, **14** (2003) 1174.

[48] Diddams S. A., Bergquist J. C., Jefferts S. R. and Oates C. W., *Science*, **306** (2004) 1318.

[49] Prestage J. D., Tjoelker R. L. and Maleki L., in *Frequency Measurement and Control: Advanced Techniques and Future Trends*, edited by Luiten A. N., *Springer Topics Appl. Res.*, Vol. **79** (Springer, Berlin, Heidelberg) 2000, p. 195.

[50] Maleki L. and Prestage J. D., in *Astrophyics, Clocks and Fundamental Constants*, edited by Karshenboim S. G. and Peik E., *Springer Lect. Notes Phys.*, Vol. **648** (Springer, Berlin, Heidelberg) 2004, p. 331.

[51] Goldenberg H. M., Kleppner D. and Ramsey N. F., *Phys. Rev. Lett*, **8** (1960) 361.

[52] Shirley J. H., *Am. J. Phys.*, **36** (1968) 949.

[53] Berthoud P., Pavlenko I., Wang Q and Schweda H., *Proc. 2003 IEEE Int. Frequ. Contr. Symp. and Joint 17th EFTF, Tampa, Florida* (IEEE) 2004, p. 105.

[54] Droz F. *et al.*, *Proc. 20th EFTF, Braunschweig, Germany* (BTP) 2006, p. 420.

[55] Packard M. E. and Swartz B. E., *IRE Trans. Instrum.*, **11** (1962) 215.

[56] Mileti G., Deng J. and Walls, F. L., *IEEE J. Quantum Electron.*, **34** (1998) 233.

[57] Affolderbach C., Mileti G. and Droz F., *Proc. 18th EFTF, Guilford, UK* (IEEE) 2004 on CD-ROM.

[58] See *Proceedings of the 2005 Joint IEEE Int. Frequ. Contr. Symposium and the PTTI Systems and Appl. Meeting, Vancouver Canada.* On CD-ROM, IEEE Catalog No. 05CH37664C.

[59] Godone A., Levi F. and Micalizio S., *Coherent Population Trapping Maser* (CLUT Edititrice, Torino) 2002.

[60] Vanier J., *Appl. Phys. B*, **81** (2005) 421.

[61] Hlaváč R., Lösch M., Luongo F. and Hahn, J., *Proc. 20th EFTF, Braunschweig, Germany* (PTB) 2006, p. 391.

[62] Laverty J., Bauch A. and Tavella P., *Proc. 19th EFTF, Besançon, France* (SFMC) 2005, p. 158.

[63] `http://www.ucte.org`

[64] Karshenboim S. G. and Peik E. (Editors), *Astrophysics, Clocks and Fundamental Constants*, *Springer Lect. Notes Phys.*, Vol. **648** (Springer, Heidelberg) 2004.

[65] Quinn T. and Burmett K. (Editors), *The fundamental constants of physics, precision measurement and the base units of the SI*, *Philos. Trans. R. Soc. London, Ser. A*, **363**, special issue, no. 1834 (The Royal Society, London) 2005.

[66] Karshenboim S. G., *Can. J. Phys.*, **78** (2001) 639.

[67] Oskay W. H. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 020801.

[68] Fischer M. *et al.*, *Phys. Rev. Lett.*, **92** (2004) 230802.

[69] Peik E., Lipphardt B., Schnatz H., Schneider T., Tamm Chr. and Karshenboim S. G., *Phys. Rev. Lett.*, **93** (2004) 170801.

This page intentionally left blank

Frequency combs applications and optical frequency standards

Th. Udem

Max-Planck Institute für Quantenoptik - Hans-Kopfermann Straße 1, 85748 Garching, Germany

F. Riehle

Physikalisch-Technische-Bundesanstalt - Bundesallee 100, 38116 Braunschweig, Germany

A laser frequency comb allows the conversion of the very rapid oscillations of visible light of some 100's of THz down to frequencies that can be handled with conventional electronics, say below 100 GHz. This capability has enabled the most precise laser spectroscopy experiments yet that allowed to test quantum electrodynamics, to determine fundamental constants and to search for possible slow changes of these constants. Using an optical frequency reference in combination with a laser frequency comb has made it possible to construct all optical atomic clocks, that are about to outperform the current cesium atomic clocks.

1. – Introduction

If one considers the attainable measurement accuracy of different physical quantities it turns out that time intervals and frequencies are to be determined with the utmost precision. Other physical dimensions, such as length mass or charge, can only be determined with orders of magnitude less accuracy. The intrinsic high precision comes about because counting, unlike any other physical measurement, has zero uncertainty connected to it. The only uncertainty in determining a frequency in hertz or oscillations per second, lies in the determination of the second. But this is about as good as it can be, because atomic clocks that are used to determine the second are the most precise instruments. In this sense frequency and time measurements are equivalent.

To exploit this potential it has been a top priority in metrology to convert other physical measurables into a time or frequency equivalent. The simplest example of such a conversion is to assign the speed of light c with a fixed defined value, which was done within the International System of Units (SI) in 1983. Since then the conversion of an optical wavelength λ to an optical frequency $\omega = 2\pi c/\lambda$ can be done without loss of accuracy[1]. The method requires a means to count optical frequencies, because only then one can use a precise interferometer to extract the wavelength. Therefore it is no surprise hat this redefinition had to wait until it became possible to count the frequency of light. At that time the idea was to calibrate iodine stabilized HeNe lasers with harmonic frequency chains that linked them to a cesium atomic clock. These frequency chains where complex devices that got operational only in few places and worked continuously only for short time intervals [1].

The optical frequency comb vastly simplified these efforts [2-8]. Even commercially available cesium clocks are usually more accurate than an iodine stabilized HeNe laser. Therefore these lasers are no longer required if a more accurate radio frequency (RF) source and a frequency comb is available. On the other hand, another class of optical standards based on trapped ions or atoms have been improving at a faster pace than the cesium clocks so that the frequency comb can be used to calibrate a RF source. As will be discussed in sect. **5** this led to the first all optical atomic clocks [9-11]. After agreeing on a particular ion or atom with a suitable clock transition, it seems likely that the definition of the SI second will be adjusted accordingly.

By definition an optical frequency comb consists of many continuous wave laser modes, equidistant in frequency space, that can be used like ruler to determine large frequency differences between lasers. By measuring the frequency separation between a laser at a frequency f and its second harmonic $2f$, the lasers absolute frequency $f = 2f - f$ is determined [12]. A frequency comb used for that purpose must span a complete optical octave. Then, not only f and $2f$ are known relative to the comb mode spacing, but all the modes in between, providing an octave full of calibrated laser lines at once.

Since light from lasers has been used to gather nearly all high-precision data about atoms, the analysis of this light is key to a better understanding of the microscopic world. To test quantum electrodynamics and to determine the Rydberg constant, transition frequencies in atomic hydrogen have been measured [13, 14]. In fact the spectroscopy of the narrowest line in hydrogen, the $1S$-$2S$ transition, was the motivation for setting up the first frequency comb [2-4]. From these measurements the Rydberg constant became the most accurately measured fundamental constant [14]. The frequency comb also helped to determined the fine-structure constant from atomic recoil shifts [15, 16] using precise values of the Rydberg constant and was the key for laboratory searches for possible slow variations of these constants [17-19]. Even if there is not yet a physical theory that would predict such a slow change there are some arguments, rather philosophically in nature,

[1] A second example is the utilization of the Josephson of effect to convert voltages into frequencies and vice versa.

that they should be there. The frequency combs are now providing a sharp tool for at least setting up stringent upper limits.

2. – Frequency combs from mode locked lasers

Frequency combs can be produced with fast and efficient electro-optik modulators that impose a large number of side bands on a single-frequency continuous laser [20]. The factor that limited the achievable width of the generated frequency comb was dispersion of the modulator crystal. After dispersion compensation was introduced [21], boosting the intensity with external optical amplifier allowed spectral broadening up to 30 THz bandwidth with the process of self-phase modulation [22].

Even more bandwidth can be generated with a device that already included dispersion compensation, gain and self-phase modulation: the Kerr lens mode-locked laser. Such a laser stabilizes the relative phase of many longitudinal cavity modes such that a solitary short pulse is formed. In the time domain this pulse propagates with its group velocity v_g back and forth between the end mirrors of the resonator. After each round trip a copy of the pulse is obtained at one of these mirrors that is partially transparent like in any other conventional laser. Because of its periodicity, the pulse train generated this way produces a spectrum that consists of discrete modes that can be identified with the longitudinal cavity modes. The process of Kerr lense mode locking (KLM) introduced in the early 90s [23], allows to generate pulses of 10 femtoseconds (fs) duration and below in a rather simple way. For Fourier-limited pulses the bandwidth is given by the inverse of the pulse duration, with a correction factor of order unity that depends on the pulse shape and the way spectral and temporal widths are measured [24]. A 10 fs for example will have a bandwidth of about 100 THz, which is close to the optical carrier frequency, typically at 375 THz (800 nm) for a laser operated near the gain maximum of titanium-sapphire, the most commonly used gain medium in these lasers. By virtue of the repeating pulses, this broad spectrum forms the envelope of a frequency comb. Depending on the length of the laser cavity, that determines the pulse round trip time, the pulse repetition rate ω_r is typically on the order of 100 MHz but 4 MHz through 2 GHz repetition rates have been used. In any case ω_r is a radio frequency readily measured and stabilized. The usefulness of this comb critically depends on how constant the mode spacing is across the spectrum and to what precision it agrees with the readily measurable repetition rate. These questions will be addressed in the next two section.

2˙1. *Derivation from cavity boundary conditions.* – As in any laser the modes with wave number $k(\omega_n)$ and frequencies ω_n must obey the following boundary conditions of the resonator of length L:

$$2Lk(\omega_n) = n2\pi. \tag{1}$$

Here n is a large integer number that measures the number of half wavelengths within the resonator. Besides this propagation phase shift there might be additional phase shifts, caused for example by diffraction (Guoy phase) or by the mirror coatings that are thought of being included in the above dispersion relation. This phase shift may depend on the wavelength of the mode, *i.e.* on the mode number n, and is not simple to determine accurately in practice. So we are seeking a description that lumps all the low-accuracy quantities into readily measurable radio frequencies. For this purpose dispersion is best be included by the following expansion of the wave vector about some mean frequency ω_m, not necessarily a cavity mode according to eq. (1):

$$2L\left[k(\omega_m)+k'(\omega_m)(\omega_n-\omega_m)+\frac{k''(\omega_m)}{2}(\omega_n-\omega_m)^2+\dots\right]=2\pi n. \tag{2}$$

The mode separation $\Delta\omega\equiv\omega_{n+1}-\omega_n$ is obtained by subtracting this formula from itself with n being replaced by $n+1$:

$$2L\left[k'(\omega_m)\Delta\omega+\frac{k''(\omega_m)}{2}\left((\omega_{n+1}-\omega_m)^2-(\omega_n-\omega_m)^2\right)+\dots\right]=2\pi. \tag{3}$$

To obtain a constant mode spacing, as the most important requirement for optical frequency combs, $\Delta\omega$ must be independent of n. This is the case if and only if all contributions of the expansion of $k(\omega)$ beyond the group velocity term $k'(\omega_m)=1/\overline{v}_g(\omega_m)$ exactly vanish. The unwanted perturbing terms, or "higher-order dispersion" terms, that contradict a constant mode spacing are those that deform the pulse as it travels inside the laser resonator. Therefore the mere observation of a stable undeformed pulse envelope stored in the laser cavity leads to a frequency comb with constant mode spacing given by

$$\Delta\omega=2\pi\frac{\overline{v}_g}{2L}\quad\text{with }\overline{v}_g=\frac{1}{k'(\omega_m)} \tag{4}$$

with the round trip averaged group velocity[2] given by $\overline{v}_g$. Having all derivatives beyond $k'(\omega_m)$ vanishing, means that $k'(\omega_m)$ and $\overline{v}_g$ must independent of frequency. Therefore this derivation is independent of the particular choice of ω_m, provided it resides within the laser spectrum. The group velocity determines the cavity round trip time T of the pulse and therefore the pulse repetition rate $\omega_r=\Delta\omega$:

$$T^{-1}=\frac{\overline{v}_g}{2L}=\frac{\omega_r}{2\pi}\,. \tag{5}$$

The frequencies of the modes ω_n of any frequency comb with a constant mode spacing ω_r can be expressed by [3,8,25,26]

$$\boxed{\omega_n=n\omega_r+\omega_{\mathrm{CE}}} \tag{6}$$

[2] The usual textbook approach ignores dispersion either as a whole ($\Delta\omega=2\pi c/2L$) or just includes a constant refractive index: $\Delta\omega=2\pi v_{\mathrm{ph}}/2L$ with v_{ph} and c being the phase velocities with and without dispersive material, respectively.

with a yet unknown frequency offset $\omega_{\rm CE}$ common to all modes. As a convention we will now number the modes such that $0 \le \omega_{\rm CE} \le \omega_r$. This means that $\omega_{\rm CE}$, like ω_r resides in the radio frequency domain. Using (6) to measure the optical frequencies ω_n requires the measurement of ω_r, $\omega_{\rm CE}$ and n. The pulse repetition rate can be measured anywhere in the beam of the mode-locked laser. To determine the comb offset requires some more effort as will be detailed in sect. **3**. In practice the beam of a continuous wave laser whose frequency is should be determined is superimposed with the beam containing the frequency comb on a photo detector to record a beat note with the nearest comb mode. Knowing ω_r and $\omega_{\rm CE}$ the only thing missing is the mode number n. This may be determined by a coarse and simple wavelength measurement or by repeating the same measurement with slightly different repetition rates [27].

Some insight on the nature of the frequency offset $\omega_{\rm CE}$ is obtained by resolving (6) for $\omega_{\rm CE}$ and using the cavity averaged phase velocity of the n-th mode $\overline{v}_p(\omega_n) = \omega_n / k(\omega_n)$. With the expansion of the wave vector in (2) and the fact that that it ends with the group velocity term, one derives

$$\omega_{\rm CE} = \omega_n - n\omega_r = \omega_m \left(1 - \frac{\overline{v}_g}{\overline{v}_p(\omega_m)}\right) \tag{7}$$

For this the frequency offset is independent of n, *i.e.* common to all modes, and vanishes if the cavity averaged group and phase velocities are identical. In such a case the pulse train possesses a strictly periodic field that produces a frequency comb containing only integer multiples of ω_r. In general though this condition is not fulfilled and the comb offset frequency is related to the difference of the group and phase round trip time. For an unchirped pulse, that has a well-defined carrier frequency ω_c, it makes sense to expand the wave vector about $\omega_m = \omega_c$ so that eq. (7) becomes

$$\omega_{\rm CE} T = \Delta\varphi \quad \text{with } \Delta\varphi = \omega_c \left(\frac{2L}{\overline{v}_g} - \frac{2L}{\overline{v}_p}\right). \tag{8}$$

As discussed in more detail in the next section, it follows that the pulse envelope continuously shifts relative to the carrier wave in one direction. The shift per pulse round trip $\Delta\varphi$ is given by the advance of the carrier phase during a phase round trip time $2L/\overline{v}_p$ in a frame that travels with the pulse. The shift $\Delta\varphi$ per pulse round trip time $T = 2L/\overline{v}_g$ fixes the frequency comb offset. Hence it has been dubbed carrier-envelope (CE) offset frequency.

How precise the condition of vanishing higher-order dispersion is fulfilled in practice could be estimated by knowing that an irregular phase variations between the modes on the order of 2π are sufficient to completely destroy the stored pulse in the time domain. The phase between adjacent modes of a proper frequency comb advances as $\omega_r t$, *i.e.* typically by some 10^8 times 2π a second. An extra random cycle per second would offset the modes by only 1 Hz from the perfectly regular grid, but destroy the pulse in the same time. Compared to the optical carrier frequency this corresponds to a relative uncertainty of 3 parts in 10^{15} at most, which is already close to the best cesium atomic clocks. Experimentally one observes the same pulse for a much longer time, in some lasers even for months. In fact no deviations have been detected yet at a sensitivity of a few parts in 10^{16} [28, 7, 29, 30].

It should be noted that a more appropriate derivation of the frequency comb must include non-linear shifts of the group and phase velocities because this is used to lock the modes in the first place. In fact it has been shown that initiating the mode-locking mechanism significantly shifts the cavity modes of a laser [31]. Of course the question remains how this mode-locking process forces the cancellation of all pulse deforming dispersive contributions with this precision. Even though this is beyond the scope of this article and details may be found elsewhere [32, 24], the simplest explanation is given in the time domain: Mode locking, *i.e.* synchronizing the longitudinal cavity modes to sum up for a short pulse, is most commonly achieved with the help of the Kerr effect which is expressed in the time domain through

$$n(t) = n_0 + n_2 I(t). \tag{9}$$

Here n_2 denotes a small intensity $I(t)$ dependence of the refractive index $n(t)$. Generally a Kerr coefficient with a fs response time is very small, so that it becomes noticeable only for the high peak intensity of short pulses. Kerr lens mode locking uses this effect in two ways. The radially intensity variation of a Gaussian mode produces a lens that becomes part of the laser resonator for a short pulse. This resonator is designed such that it has larger losses without such a lens, *i.e.* for a superposition of randomly phased modes. In addition self phase modulation can compensate the some de-phasing of the modes. A wave packet that does not deform as it travels because higher-order dispersive terms are compensated by self phase modulation, is called a soliton [33, 34]. The nice feature about this cancellation is that it is self-adjusting: Slightly larger peak intensity extends the pulse duration and reduces the peak intensity leaving the pulse energy constant and vice versa. So that neither the pulse intensity nor the dispersion has to be matched exactly (which would not be possible) to generate a soliton. Of course it helps to pre-compensate higher-order dispersion as good as possible before initiating mode locking. It appears that for the shortest pulses the emission bandwidth is basically given by the bandwidth this pre-compensation is achieved [35, 36]. Once mode locking is initiated, mostly by a mechanical disturbance of the laser cavity, the modes are pulled on the regular grid described by eq. (6). Another way of thinking about this process is that self phase modulation that occurs with the repetition frequency, produces side bands on each mode that will injection lock the neighboring modes by mode pulling. This pulling produces the regular grid of modes and is limited by the injection locking range [37], *i.e.* how well the pre-compensation of higher-order dispersion was done.

2˙2. *Derivation from the pulse train.* – Rather than considering intracavity dispersion, the cavity length and so on, one may simply analyze the emitted pulse train with little consideration how it is generated. For this one assumes that the electric field $E(t)$, measured for example at the output coupling mirror, can be written as the product of a periodic envelope function $A(t)$ and a carrier wave $C(t)$:

$$E(t) = A(t)C(t) + \text{c.c.} \tag{10}$$

The envelope function defines the pulse repetition time $T = 2\pi/\omega_r$ by demanding $A(t) = A(t-T)$. The only thing about dispersion that should be added for this description, is that there might be a difference between the group velocity and the phase velocity inside the laser cavity. This will shift the carrier with respect to the envelope by a certain amount after each round trip. The electric field is therefore in general not periodic with T. To obtain the spectrum of $E(t)$ the Fourier integral has to be calculated:

$$\tilde{E}(\omega) = \int_{-\infty}^{+\infty} E(t)e^{i\omega t}\mathrm{d}t. \tag{11}$$

Separate Fourier transforms of $A(t)$ and $C(t)$ are given by

$$\tilde{A}(\omega) = \sum_{n=-\infty}^{+\infty} \delta\left(\omega - n\omega_r\right)\tilde{A}_n \quad \text{and } \tilde{C}(\omega) = \int_{-\infty}^{+\infty} C(t)e^{i\omega t}\mathrm{d}t. \tag{12}$$

A periodic frequency chirp imposed on the pulses is accounted for by allowing a complex envelope function $A(t)$. Thus the "carrier" $C(t)$ is defined to be whatever part of the electric field that is non-periodic with T. The convolution theorem allows us to calculate the Fourier transform of $E(t)$ from $\tilde{A}(\omega)$ and $\tilde{C}(\omega)$:

$$\tilde{E}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{A}(\omega')\tilde{C}(\omega-\omega')\mathrm{d}\omega' + \text{c.c.} = \frac{1}{2\pi}\sum_{n=-\infty}^{+\infty} \tilde{A}_n\tilde{C}\left(\omega - n\omega_r\right) + \text{c.c.} \tag{13}$$

The sum represents a periodic spectrum in frequency space. If the spectral width of the carrier wave $\Delta\omega_c$ is much smaller than the mode separation ω_r, it represents a regularly spaced comb of laser modes just like eq. (6), with identical spectral line shapes, namely the line shape of $\tilde{C}(\omega)$ (see fig. 1). If $\tilde{C}(\omega)$ is centered at say ω_c, than the comb is shifted from containing only exact harmonics of ω_r by ω_c. The center frequencies of the mode members are calculated from the mode number n [3, 8, 25, 26]:

$$\omega_n = n\omega_r + \omega_c\,. \tag{14}$$

The measurement of the frequency offset ω_c [2-8]. as described below usually yields a value modulo ω_r, so that renumbering the modes will restrict the offset frequency to smaller values than the repetition frequency and again yields eq. (6).

The individual modes can be separated with a suitable spectrometer if the spectral width of the carrier function is narrower than the mode separation: $\Delta\omega_c \ll \omega_r$. This condition is easy to satisfy, even with a free-running titanium-sapphire laser. If a single mode is selected from the frequency comb, one obtains a continuous wave. However it

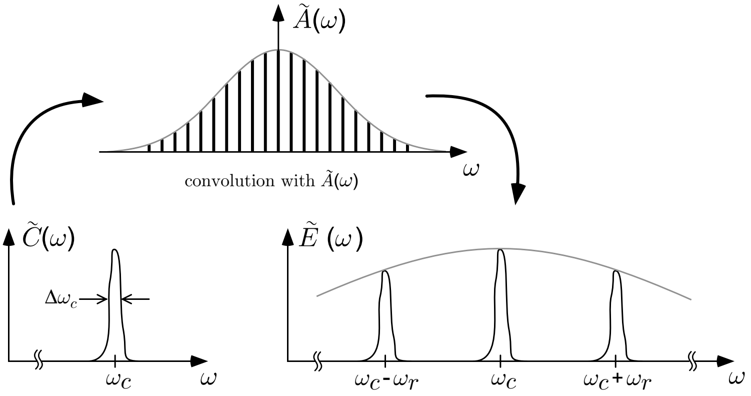


Fig. 1. – The spectral shape of the carrier function (left), assumed to be narrower than the pulse repetition frequency ($\Delta\omega_c \ll \omega_r$), and the resulting spectrum according to eq. (13) after modulation by the envelope function (right).

is easy to show that a grating with sufficient resolution would be at least as large as the laser cavity, which appears unrealistic for a typical laser with a 2 m cavity length. Fortunately for experiments performed so far it has never been necessary to resolve a single mode in the optical domain as explained in sect. **3**.

Now let us consider two instructive examples of possible carrier functions. If the carrier wave is monochromatic $C(t) = e^{-i\omega_c t - i\varphi}$, its spectrum will be δ-shaped and centered at the carrier frequency ω_c. The individual modes are also δ-functions $\tilde{C}(\omega) = \delta(\omega - \omega_c)e^{-i\varphi}$. The frequency offset (14) is identified with the carrier frequency. According to eq. (10) each round trip will shift the carrier wave with respect to the envelope by $\Delta\varphi = \arg(C(t-T)) - \arg(C(t)) = \omega_c T$ so that the frequency offset is given by $\omega_{\mathrm{CE}} = \Delta\varphi/T$ [3,8,25,26]. In a typical laser cavity this pulse-to-pulse carrier-envelope phase shift is much larger than 2π, but measurements usually yield a value modulo 2π. The restriction $0 \le \Delta\varphi \le 2\pi$ is synonymous with the restriction $0 \le \omega_{\mathrm{CE}} \le \omega_r$ introduced earlier. Figure 2 sketches this situation in the time domain for a chirp free-pulse train.

As a second example consider a train of half-cycle pulses like

$$E(t) = E_0 \sum_k e^{-\left(\frac{t-kT}{\tau}\right)^2}. \tag{15}$$

In this case the electric field would be repetitive with the round trip time. Therefore $C(t)$ is a constant and its Fourier transform is a delta-function centered as $\omega_c = 0$. If it becomes possible to build a laser able to produce a stable pulse train of that kind, all the comb frequencies would become exact harmonics of the pulse repetition rate. Obviously, this would be an ideal situation for optical frequency metrology([3]).

[3] It should be noted though that this is a rather academic example because such a pulse would be deformed quickly upon propagation since it contains vastly distinct frequency components with different diffraction. For example, the DC component that this carrier certainly has, would not propagate at all.

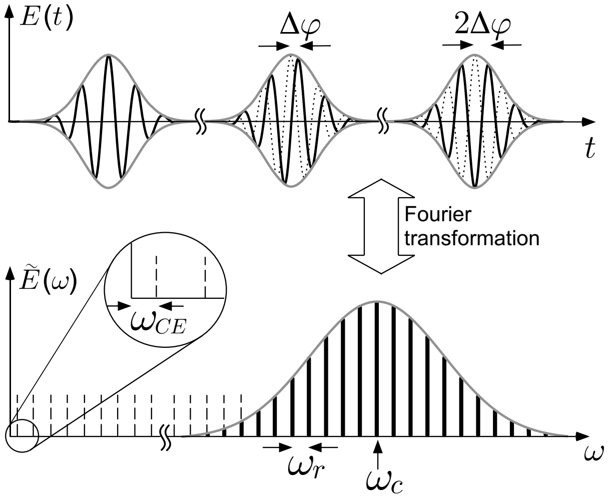


Fig. 2. – Consecutive un-chirped pulses ($A(t)$ real) with carrier frequency ω_c and the corresponding spectrum (not to scale). Because the carrier propagates with a different velocity within the laser cavity than the envelope (phase- and group velocity), the electric field does not repeat itself after one round trip. A pulse-to-pulse phase shift $\Delta\varphi$ results in an offset frequency of $\omega_{\mathrm{CE}} = \Delta\varphi/T$. The mode spacing is given by the repetition rate ω_r. The width of the spectral envelope is given by the inverse pulse duration up to a factor order unity that depends on the pulse shape (the time bandwidth product of a Gaussian pulse for example is 0.441 [24]).

As these examples are instructive it is important to note that one neither relies on assuming a strictly periodic electric field nor that the pulses are unchirped. The strict periodicity of the spectrum as stated in eq. (13), and the possibility to generate beat notes between continuous lasers and single modes [38], are the only requirement that enables precise optical to radio frequency conversions.

In a real laser the carrier wave will not be a clean sine wave as in the above example. The mere periodicity of the field, allowing a pulse-to-pulse carrier envelope phase shift, already guarantees the comb-like spectrum. Very few effects can disturb that property. In particular, for an operational frequency comb, both ω_r and ω_{CE} will be servo controlled so that slow drifts are compensated. The property that the comb method really relies on, is the mode spacing being constant across the spectrum. As explained above, even a small deviation from this condition will have very quick and devastating effects on the pulse envelope. Not even an indefinitely increasing chirp could disturb the mode spacing constancy, as this can be seen as a constantly drifting carrier frequency that does not perturb the spectral periodicity but shifts the comb as a whole. However, the phase of individual modes can fluctuate about an average value required for staying in lock with the rest of the comb. This will cause noise that can broaden individual modes as discussed in the next section.

2·3. *Linewidth of a single mode.* – The modes of a frequency comb have to be understood as continuous laser modes. As such they possess a linewidth which is of interest here. Of course as usual in such a case, several limiting factors are effective at the same time. It is instructive to derive the Fourier limited linewidth that is due to observing the pulse train for a limited number of pulses only. Following a derivation by Siegman [37]([4]) the linewidth of a train of N pulses can be derived. In accordance with the previous section we assume that the pulse train consists of identical pulses $\mathcal{E}(t)$ separated in time by T and subjected to a pulse-to-pulse phase shift of $e^{i\Delta\varphi}$:

$$E(t) = \frac{E_0}{\sqrt{N}} \sum_{m=0}^{N-1} e^{im\Delta\varphi} \mathcal{E}(t - mT). \tag{16}$$

For decent pulse shapes, that fall off at least as $\propto 1/t$ from the maximum, this series converges even as N goes to infinity. Using the shift theorem

$$\mathcal{FT}\left\{\mathcal{E}(t-\tau)\right\} = e^{-i\omega\tau} \mathcal{FT}\left\{\mathcal{E}(t)\right\} \tag{17}$$

one can relate the Fourier transform of the pulse train $E(t)$ to the Fourier transform of a single pulse $\tilde{\mathcal{E}}(\omega)$. With the sum formula for the geometric series this becomes

$$\tilde{E}(\omega) = \frac{E_0 \tilde{\mathcal{E}}(\omega)}{\sqrt{N}} \sum_{m=0}^{N-1} e^{-im(\omega T - \Delta\varphi)} = \frac{E_0 \tilde{\mathcal{E}}(\omega)}{\sqrt{N}} \frac{1 - e^{-iN(\omega T - \Delta\varphi)}}{1 - e^{-i(\omega T - \Delta\varphi)}}\,. \tag{18}$$

Now the intensity spectrum for N pulses $I_N(\omega)$ may be calculated from the spectrum of a single pulse $I(\omega) \propto |\tilde{\mathcal{E}}(\omega)|^2$:

$$I_N(\omega) = \frac{1 - \cos(N(\omega T - \Delta\varphi))}{N(1 - \cos(\omega T - \Delta\varphi))} I(\omega). \tag{19}$$

Figure 3 sketches this result for several N. As the spectrum of single pulse is truly a continuum, the modes are emerging and becoming sharper as more pulses are added. This is similar to the diffraction from a grating that becomes sharper as more grating lines are illuminated. The spectral width of a single mode can now be calculated from (19) and may be approximated by $\Delta\omega \approx \sqrt{24}/TN$ for the pulse train observation time NT. From this the Fourier limited line width is reduced after one second of observation time NT to $\sqrt{24}/2\pi\,\text{Hz} \approx 0.78\,\text{Hz}$ and further decreases as the inverse observation time. In the limit of an infinite number of pulses pulse the spectral shape of the modes approximate delta-functions with $x = \omega T - \Delta\varphi \approx 2\pi n$

$$\frac{1}{2\pi} \lim_{N\to\infty} \frac{1 - \cos(Nx)}{N(1 - \cos(x))} \approx \frac{1}{\pi} \lim_{N\to\infty} \frac{1 - \cos(Nx)}{Nx^2} = \delta(x). \tag{20}$$

[4] The carrier envelope phase shift was ignored in [37] but can easily be accounted for here.

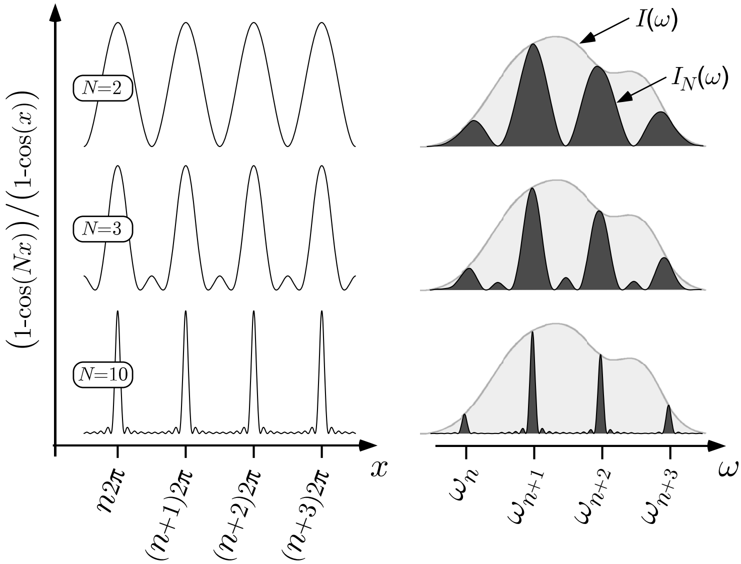


Fig. 3. – Left: The function $(1-\cos(Nx))/(1-\cos(x))$ normalized to the peak for $N = 2, 3$ and 10 pulses. The maxima are at $x = \omega_n T - \Delta\varphi = 2\pi n$ with integer n. From that and redefining $\omega_r = 2\pi/T$ and $\omega_{\text{CE}} = \Delta\varphi/T$ we derive again the frequency comb equation (6). Right: The resulting frequency comb spectrum is given by the spectrum of a single pulse multiplied with the comb function at the left-hand side.

The whole frequency comb becomes an equidistant array of delta-functions:

$$I_N(\omega) \to I(\omega) \sum_n \delta(\omega T - \Delta\varphi - 2\pi n). \tag{21}$$

The Fourier limit to the linewidth derived here is important when only a number of pulses can be used for example for direct comb spectroscopy as will be discussed in subsect. **4**˙5. On the other hand, when a large number of pulses contribute to the signal, for example when measuring the carrier envelope beat note, other limits enter. In most of these cases acoustic vibrations seem to set the limit as can be seen by observing that the noise of the repetition rate of an unstabilized laser dies off very steeply for frequencies above typical ambient acoustic vibrations around 1 kHz [39]. Even if these fluctuations are controlled as they can be with the best continuous wave lasers, a limit set by quantum mechanics in terms of the power-dependent Schawlow-Townes formula applies. Remarkably the total power of *all* modes enters this formula to determine the line width of a *single* mode [40]. In fact subhertz linewidths have been measured across the entire frequency comb when it is stabilized appropriately [41, 42].

2˙4. *Generating an octave spanning comb.* – The spectral width of a pulse train emitted by a fs laser can be significantly broadened in a single-mode fiber [34] by self-phase modulation. According to eq. (9) assuming a single-mode carrier wave, a pulse that has propagated the length l acquires a self-induced phase shift of

$$\Phi_{\mathrm{NL}}(t) = -n_2 I(t)\omega_c l/c \quad \text{with } I(t) = \frac{1}{2}c\varepsilon_0|A(t)|^2 \,. \tag{22}$$

For fused silica the non-linear coefficient is comparatively small but almost instantaneous even on the time scale of fs pulses. This means that different parts of the pulse travel at different speed. The result is a frequency chirp across the pulse without affecting its duration. The pulse is no longer at the Fourier limit so that the spectrum is much broader than the inverse pulse duration where the extra frequencies are determined by the time derivative of the self-induced phase shift $\dot{\Phi}_{\mathrm{NL}}(t)$. Self-phase modulation modifies the envelope function in eq. (10) according to

$$A(t) \longrightarrow A(t)e^{i\Phi_{\mathrm{NL}}(t)} \,. \tag{23}$$

Because $\Phi_{\mathrm{NL}}(t)$ has the same periodicity as $A(t)$ the comb structure of the spectrum is maintained and the derivations of subsect. **2**˙2 remain valid because periodicity of $A(t)$ was the only assumption made. An optical fiber is most appropriate for this process because it can maintain the necessary small focus area over a virtually unlimited length. In practice, however, other pulse reshaping mechanisms, both linear and non-linear, are present so that the above explanation is too simple.

Higher-order dispersion is usually limiting the effectiveness of self-phase modulation as it increases the pulse duration and therefore lowers the peak intensity after a propagation length of a few mm or cm for fs pulses. On can get a better picture if pulse broadening due to group velocity dispersion $k''(\omega_c)$ is included. To measure the relative importance of the two processes, the dispersion length L_{D} (the length that broadens the pulse by a factor $\sqrt{2}$) and the non-linear length L_{NL} (the length that corresponds to the peak phase shift $\Phi_{\mathrm{NL}}(t = 0) = 1$) are used [34]:

$$L_{\mathrm{D}} = \frac{4\ln(2)\tau^2}{|k''(\omega_c)|} \qquad L_{\mathrm{NL}} = \frac{cA_f}{n_2\omega_c P_0}\,, \tag{24}$$

where τ_0, A_f and $P_0 = (1/2)A_f c\varepsilon_0|A(t = 0)|^2$ are the initial pulse duration, the effective fiber core area and the pulse peak power. In the dispersion dominant regime $L_{\mathrm{D}} \ll L_{\mathrm{NL}}$ the pulses will disperse before any significant non-linear interaction can take place. For $L_{\mathrm{D}} > L_{\mathrm{NL}}$ spectral broadening could be thought as effectively taking place for a length L_{D} even though the details are more involved. The total non-linear phase shift can therefore be approximated by the number of non-linear lengths within one dispersion length. As this phase shift occurs roughly within one pulse duration τ, the spectral broadening is estimated to be $\Delta\omega_{\mathrm{NL}} = L_{\mathrm{NL}}/L_{\mathrm{D}}\tau$. As an example consider a silica single mode fiber (Newport F-SF) with $A_f = 26\,\mu\mathrm{m}^2$, $k''(\omega_c) = 281\,\mathrm{fs/cm}^2$ and $n_2 = 3.2 \times 10^{-16}\,\mathrm{cm}^2/\mathrm{W}$ that is seeded with $\tau = 73\,\mathrm{fs}$ Gaussian pulses (FWHM intensity) at 905 nm with 225 mW average power and a repetition rate of 76 MHz [3, 4]. In this case the dispersion length becomes 6.1 cm and the non-linear length 35 mm. The expected spectral broadening of $L_{\mathrm{NL}}/L_{\mathrm{D}}\tau = 2\pi \times 44\,\mathrm{THz}$ is indeed very close to the observed value [3].

It turns out that within this model the spectral broadening is independent of the pulse duration τ because $P_0 \propto \tau$. Therefore using shorter pulses may not be effective for extending the spectral bandwidth beyond an optical octave as required for simple self-referencing (see sect. **3**). However, very efficient spectral broadening can be obtained in microstructure fiber([5]) that can be manufactured with $k''(\omega_c) \approx 0$ around a design wavelength [43-45]. In this case the pulses are broadened by other processes (linear and non-linear) than group velocity dispersion as they propagate along the fiber. Eventually this will also terminate self-phase modulation and the dispersive length has to be replaced appropriately in the above analysis. At this point a whole set of effects enter such as Raman and Brillouin scattering, optical wave breaking and modulation instability [34]. Some of these processes even spoil the usefulness of the broadened frequency combs as the amplify noise.

A microstructure fiber uses an array of submicron-sized air holes that surround the fiber core and run the length of a silica fiber to obtain a desired effective dispersion. This can be used to maintain the high peak power over an extended propagation length and to significantly increase the spectral broadening. With these fibers it became possible to broaden low peak power, high repetition rate lasers to beyond one optical octave as fig. 4 shows.

A variant of the microstructure fibers are regular single-mode fibers that have been pulled in a flame to form a tapered section of a few cm lengths [46]. When the diameter of the taper becomes comparable to the core diameter of the microstructure fibers, pretty much the same properties are observed. In the tapered section the action of the fiber core is taken over by the whole fiber. The original fiber core then is much too small to have any influence on the light propagation. The fraction of evanescent field around the taper and along with it the dispersion characteristics can be adjusted by choosing a suitable taper diameter.

The peak intensity that can be reached with a given mode-locked laser does not only depend on the pulse duration but critically on the repetition rate. Because most laser have pretty much the same average output power, a lower repetition rate concentrates this available power into fewer pulses per second. Comparing different laser systems the

([5]) Some authors refer to these fibers as photonic crystal fibers that need to be distinguished form photonic bandgap fibers. The latter use Bragg diffraction to guide the light, while the fibers discussed here use the traditional index step, with the refractive index determined by the air filling factor.

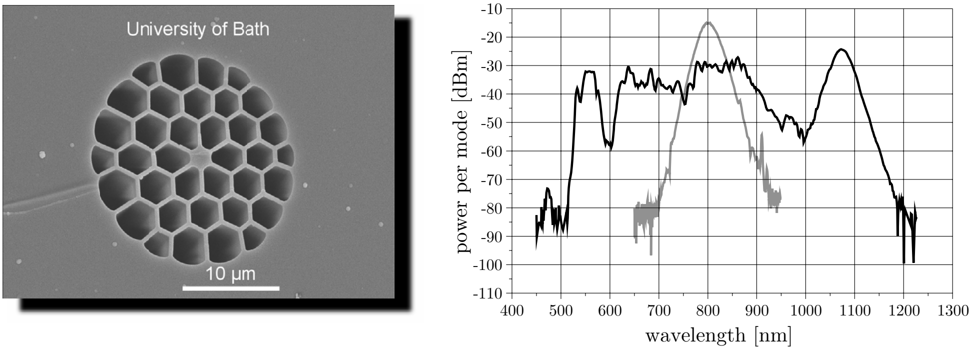


Fig. 4. – Left: SEM image of a the core of a microstructure fiber made at the University of Bath, UK [43]. The light is guided in the central part but the evanescent part of the wave penetrates into the air holes that run parallel to the fiber core and lower the effective refractive index without any doping. The guiding mechanism is the same as in a conventional single mode fiber. Right: Power per mode on a logarithmic scale (0 dBm = 1 mW). The lighter 30 nm (14 THz–3 dB) wide spectrum displays the laser intensity and the darker octave spanning spectrum (532 nm through 1064 nm) is observed after the microstructure fiber that was 30 cm long. The laser was operated at $\omega_r = 2\pi \times 750\,\mathrm{MHz}$ (modes not resolved) with 25 fs pulse duration. An average power of 180 mW was coupled through the microstructure fiber [47].

repetition rates in use cover more than 12 orders of magnitude from the most powerful lasers (one laser shot per 1000 s) to highly repetitive lasers at $\omega_r = 2\pi \times 2\,\mathrm{GHz}$ [48]. It has been long known that with enough peak intensity one can produce very wide spectra that where initially called "white light continuum". Unfortunately for a long time this was only possible at a repetition rate of around 1 kHz that indeed justifies the name: The generated spectrum could be called a "continuum" as there was not much hope to resolve the modes in any way for self-referencing or by a beat note with another cw laser. Because of its high efficiency the microstructure fiber allowed to generate an octave wide spectra with repetition rates up to 1 GHz that conveniently allowed the beat notes to be separated as described in sect. **3**. In addition a large mode spacing puts more power in each mode improving the signal to noise of the beat notes.

The laser that quickly became the working horse in the field for this reason was a rather compact titanium-sapphire ring laser with typical repetition rates of 500 MHz to 1 GHz [48]. The ring design solved another problem that is frequently encountered when coupling a laser into an optical fiber. Optical feedback from the fiber may disturb the laser operation and even prevent mode locking in some cases. The standard solution to this problem would be to place an optical isolator between the fiber and and the laser. In this case however such a device would have almost certainly enough group velocity dispersion to prevent any subsequent spectral broadening if this is not compensated for. In a ring laser the pulses reflected back from the fiber travel in the opposite direction and do not talk to the laser pulses unless they meet inside the laser crystal. The latter can be prevented by observing the distance of the fiber from the laser. A disadvantage of these lasers is that they are not easy to align and have so far not become turn key systems that can be operated unattended for a long time say in an all optical atomic clock (see below).

Even though microstructure fibers have allowed the simple $f - 2f$ self-referencing for the first time, they also have some drawbacks. To achieve the desired properties, the microstructure fibers need to have a rather tiny core. The coupling to this core causes problems due to mechanical instabilities and temperature drifts even with low level and stable mounts. Another problem is the observed but not fully understood strong polarization dependence of the fibers broadening action. The possibility of long-term continuous operation is not so much of an issue for spectroscopy, because data taking in these experiments usually do not last very long and they need some attention on there own. However, this possibility seem to be a key requirement operating an all optical atomic clock. So far the only way to operate such a system unattended for hours was to use a set of additional servo systems that continuously measure and correct deviations from the fiber coupling and polarization [49].

Another problem with spectral broadening by self phase modulation in general is a excess noise level of the beat notes well above the shot noise limit [50, 51]. In fact using the 73 fs laser mentioned above with a microstructure fiber, a two-octave–spanning spectrum is generated within a few centimeters of fiber. However this spectrum does not separate into modes that could be detected by a beat note measurement with a cw laser but consist of noise [52]. One possible explanation is the Raman effect that produces strong gain about 13 THz to the red from the pump wavelength. If this gain is not seeded, it may trigger an avalanche of photons from the vacuum, that bear no phase relationship to the carrier wave and coherence is lost([6]). For sufficiently broad input spectra the Raman gain is seeded coherently with modes from the frequency comb amplifying the low-frequency modes at the cost of the high-frequency modes. For longer pulses, say below $0.441/(13\,\mathrm{THz}) = 34\,\mathrm{fs}$ for Gaussian pulse shape, less seeding occurs. In fact a more detailed calculation predicts that enough coherence is maintained by self-phase modulation if the seeding the pulses are shorter than 50 fs [53].

By going to shorter pulses for the the seed laser this problem can be handled but the alignment issue remains. On the other hand, lasers that reach an octave-spanning spectrum without using any external self-phase modulation can solve this problem [54-57]. So far however, these lasers seem to be rather delicate to handle so that one alignment problem is replaced with another. An interesting alternative are lasers that avoid the use of microstructure fibers in another way. For wide-band but not quite octave-spanning lasers, a $2f - 3f$ self-referencing becomes possible by doubling the blue wing of the spectrum and beat it with the tripled red wing [58, 59]. Such a system can remain phase

([6]) This process is called "stimulated Raman scattering" because all but the first photon is produced by stimulated emission. It should be noted though that the first spontaneous photon destroys the coherence of the whole process.

locked unattended for several hours without the burden of having extra servo systems. Related but somewhat simpler seems to be a laser that produces pulses short enough so that a little bit of self-phase modulation generated in single pass through a difference frequency generating crystal provides sufficient spectral broadening [60].

Yet another class of frequency combs that can stay in lock for even longer times are fs fiber lasers [61]. The most common type is the erbium-doped fiber laser that emits within the telecom band around 1500 nm. For this reason advanced and cheap optical components are available to build such a laser. The mode-locking mechanism is similar to the Kerr lens method, except that non-linear polarization rotation is used to favor the pulsed high peak intensity operation. Up to a short free-space section that can be build very stable, these lasers have no adjustable parts. Bulk fused silica has its zero group velocity dispersion at around $1.2\,\mu$m but this can be shifted to $1.5\,\mu$m in an optical fiber. If, in addition, the radial dependence of the refractive index is designed to obtain a small core area A_f, the fiber becomes what is called a highly non-linear fiber (HNLF) without any microstructure. These HNLF's are commercially available and can be spliced directly to a fs fiber laser. This virtually eliminates the remaining alignment sensitive parts as the free space frequency doubling stage and bat note detection can be build rather robust. Continuous stabilized operation for many hours [62, 63] have been reported. The Max-Planck Institute für Quantenoptik in Garching/Germany operates a fiber based self-referenced frequency comb that stays locked without interruption for months. A significantly large jitter of the observed CE-beat note has been observed in these lasers and can either be suppressed by using low noise pump lasers [64] or eliminated with a fast servo system.

3. – Self-referencing

The measurement of ω_{CE} fixes the position of the whole frequency comb and is called self-referencing. The method relies on measuring the frequency gap between *different* harmonics derived from the *same* laser or frequency comb. The first crude demonstration [2] employed the 4th and the 3.5th harmonic of a $f = 88.4\,\mathrm{THz}$ ($3.39\,\mu$m) laser to determine ω_{CE} according to $4\omega_n - 3.5\omega_{n'} = (4n - 3.5n')\omega_r + 0.5\omega_{\mathrm{CE}} = 0.5\omega_{\mathrm{CE}}$ with $4n - 3.5n' = 0$. To achieve the condition of the latter equation, both n and n' have to be active modes of the frequency comb. The required bandwidth is $0.5f = 44.2\,\mathrm{THz}$ which is what the 73 fs laser together with a single-mode fiber as discussed in the previous section can generate.

A much simpler approach is to fix the absolute position of the frequency comb by measuring the gap between ω_n and ω_{2n} of modes taken directly from the frequency comb [4-8]. In this case the carrier-envelope offset frequency ω_{CE} is directly produced by beating the frequency doubled([7]) red wing of the comb $2\omega_n$ with the blue side of the

[7] It should be noted that this does not simply mean the doubling of each individual mode, but the general sum frequencies generation of all modes. Otherwise the mode spacing, and therefore the repetition rate, would be doubled as well.

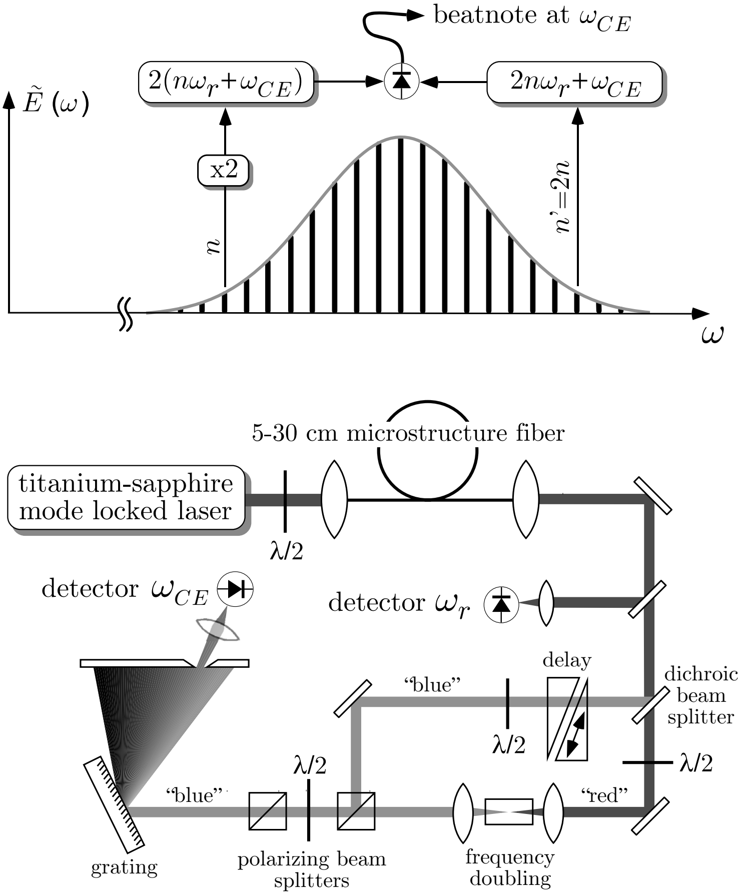


Fig. 5. – Top: $f-2f$ self-referencing by detecting a beat note at ω_{CE} between the frequency doubled "red" wing $2(n\omega_r+\omega_{\mathrm{CE}})$ of the frequency comb and the "blue" modes at $2n\omega_r+\omega_{\mathrm{CE}}$. Bottom: Layout of the self-referencing scheme. See text for details.

comb at ω_{2n}: $2\omega_n-\omega_{n'}=(2n-n')\omega_r+\omega_{\mathrm{CE}}=\omega_{\mathrm{CE}}$ where again the mode numbers n and n' are chosen such that $(2n-n')=0$. This approach requires an octave spanning comb, *i.e.* a bandwidth of 375 THz if centered at the titanium-sapphire gain maximum at 800 nm.

Figure 5 sketches the $f-2f$ self-referencing method. The spectrum of a titanium-sapphire mode-locked laser is first broadened to more than one optical octave with a microstructure fiber. A broad-band $\lambda/2$ wave plate allows to choose the polarization with the most efficient spectral broadening. After the fiber a dichroic mirror separates the infrared ("red") part from the green ("blue"). The former is frequency doubled in a non-linear crystal and reunited with the green part to create a wealth of beat notes, all at ω_{CE}. These beat notes emerge as frequency difference between $2\omega_n - \omega_{2n}$ according to eq. (6) for various values of n. The number of contributing modes is given by the phase matching bandwidth $\Delta\nu_{pm}$ of the doubling crystal and can easily exceed 1 THz. To bring all these beat notes at ω_{CE} in phase, so that they all add constructively an adjustable delay in form of a pair of glass wedges or corner cubes is used. It is straightforward to show that the condition for a common phase of all these beat notes is that the green and the doubled infrared pulse reach the photo detector at the same time. The adjustable delay allows to compensate for different group delays, including the fiber. In practice the delay needs to be correct within $c\Delta\nu_{pm}$ which is 300 μm for $\Delta\nu_{pm} = 1$ THz. Outside this range a beat note at ω_{CE} is usually not detectable.

Whereas the half wave plates in the two interferometer arms are used to adjust for whatever polarization exits the microstructure fiber, the half wave plate between the two polarizing beam splitters helps to find the optimum relative intensity of the two beating pulses. It can be shown that the maximum signal to noise ratio is obtained for equal intensities reaching the detector within the optical bandwidth that contributes to the beat note [3]. In practice this condition is most conveniently adjusted by observing the signal-to-noise ratio of the ω_{CE} beat note with a radio frequency spectrum analyzer. For this purpose a low-cost analog device that operates up to ω_r is usually sufficient.

A grating is used to prevent the extra optical power, that does not contribute to the signal but adds to the noise level, from reaching the detector. Typically only a large relative bandwidth of say 1 THz/375 THz needs to be selected so that a very moderate resolution illuminating 375 lines is sufficient. For this reason it is usually not necessary to use a slit between the grating and the photo detector. Sufficient resolution can be reached with a small low-cost 1200 lines per mm grating illuminated with a beam collimated with ×10 microscope objective out of the microstructured fiber.

When detecting the beat note as described above, more than one frequency component is obtained for two reasons. First of all any beat note, even between two cw lasers, generates two components because the radio frequency domain cannot decide which of the two optical frequencies is larger than the other. Secondly, observing the beat notes between frequency combs, not only the desired component $k = 2n - n' = 0$ is registered, but all integer values of k, positive and negative contribute, up to the bandwidth of the photo detector. This leads to a set of radio frequency beat notes at $k\omega_r \pm \omega_{\mathrm{CE}}$ for $k = \ldots -1, 0, +1 \ldots$. In addition the repetition rate, including its harmonics will most likely give the strongest components. After carefully adjusting the nonlinear interferometer, spatially and spectrally, and scanning the delay line for the proper pulse arrival times, the radio frequency spectrum may look like the one shown in fig. 6. A low-pass filter with a cut-off frequency of $0.5\omega_r$ selects exactly one beat note at $\pm\omega_{\mathrm{CE}}$. The design of such a filter may be tricky, mostly depending on how much stronger the repetition rate signal exceeds the beat note at ω_{CE}. The sketch in fig. 6 gives a feeling on how steep this filter needs to be at the cut-off in order to suppress the unwanted components below the noise level. Such a suppression is required for taking the full advantage of the signal-to-noise ratio. For this reason it is desirable to work at higher repetition rates. At ω_r around $2\pi \times 800$ MHz, as used mostly for the ring titanium-sapphire lasers described above, the

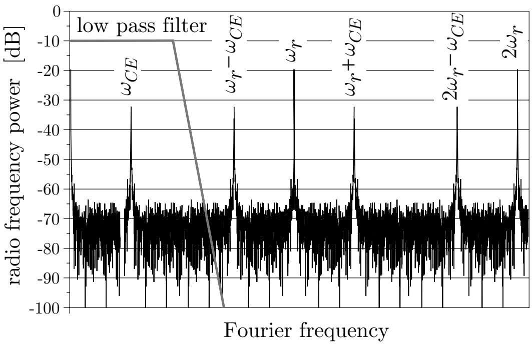


Fig. 6. – Radio frequency spectrum produced by a self-referencing non-linear interferometer such as the $f-2f$ interferometer shown in fig. 5. A low-pass filter with a cut-off at $0.5\omega_r$ selects the component at $\pm\omega_{\mathrm{CE}}$.

filter requirements are much more relaxed than say at 80 MHz. In addition, a larger repetition rate concentrates more power in each mode further improving the beat notes with the frequency comb. It should be noted though, that currently higher repetition rates cannot be used because the associated lower peak power will make it difficult to achieve spectral broadening beyond one optical octave as detailed in subsect. **2**˙4.

As described, both degrees of freedom ω_r and ω_{CE} of the frequency comb can be measured up to a sign in ω_{CE} that will be discussed below. For stabilization of these frequencies, say relative to a radio frequency reference, it is necessary to be able to control them. Again the repetition rate turns out to be simpler. By mounting one of the lasers cavity mirrors on a piezo electric transducer allows to control the pulse round trip time. Another option is offered by mode-locked lasers that use prism pairs to compensate the intracavity group velocity dispersion. In this case tipping the mirror at the dispersive end where the cavity modes are spatially separated, changes the relative cavity lengths of the individual modes and thereby the mode spacing in frequency space [7]. In practice the detected repetition frequency is mixed with the radio frequency reference, *i.e.* the frequency difference is generated, low-pass filtered and with appropriate gain send back to the piezo electric transducer. When properly designed such a phase-locked loop forces one oscillator, the repetition rate, to stay in phase with another, the radio frequency reference. Because these servo systems are standard components in many electronic devices such as FM radio receivers, a large amount of literature exists on their design and stability analysis [65].

Setting up a phase-locked loop for the repetition rate therefore seems rather straightforward. However, some caution concerning the servo bandwidth needs to be observed. It turns out that the large frequency multiplication factor n in eq. (6) may also multiplies the noise of the reference oscillator. The phase noise power for direct frequency multiplication by n increases proportional to n^2 [66], so that a factor of $n = 10^6$, that would take us from a 100 MHz radio frequency signal to a 100 THz optical signal, increases the noise by 120 dB. On this basis it has been predicted that, using even the best available reference oscillator, it is impossible to multiply in a single step from the radio frequency domain into the optical [67]. The frequency comb does just that but avoids the predicted carrier collapse. In this case the laser acts as a flywheel in the optical that does not follow the fast phase fluctuations of the reference oscillator but averages them out. In this sense the n^2 multiplication law does not apply, because it assumes a phase stiff frequency multiplication that would correspond to an infinite servo bandwidth. Fortunately a typical free-running titanium-sapphire mode-locked laser shows very good phase stability of the pulse train on its own. For averaging times shorter than typical acoustic vibrations of several ms period, such a laser shows better phase stability than a high-quality synthesizer. It is therefore essential to use a moderate servo bandwidth for phase locking the repetition rate of a few 100 Hz at most. A small servo bandwidth may be implemented electronically by appropriate filtering or mechanically by using larger masses than the usual tiny mirrors mounted on piezo transducers for high servo speed. In some case a complete one inch mirror mount has been moved for controlling the repetition rate [15].

Controlling the carrier envelope frequency requires some effort. Experimentally it turned out that the energy of the pulse stored inside the mode locked laser has a strong influence on ω_{CE}. After initial explanations of this effects turned out to be too crude, more appropriate mechanisms have been found [68, 69]. Conventional soliton theory [33] predicts a dependence of the phase velocity but no dependence of the group velocity on the pulse peak intensity. Any difference in the cavity round trip phase delay and the cavity round trip group delay results in a pulse to pulse carrier envelope phase shift and therefore a non-vanishing ω_{CE}. However, the intensity dependence of that effect may turn out to have the wrong sign [70]. The reason is that higher-order effects, usually neglected in the conventional soliton theory, play an important role. The Raman effect in the titanium-sapphire crystal produces an intensity-dependent redshift that in turn affects the group round trip time. In general this leads to an extra term in eq. (8) for the pulse to pulse carrier envelope phase shift per round trip [69]:

$$\Delta\varphi = \omega_c \left(\frac{2L}{\overline{v}_g} - \frac{2L}{\overline{v}_p} + BI_p \right). \tag{25}$$

Because this phase shift is directly proportional to ω_{CE} this equation also describes its dependence on the pulse peak power I_p. The magnitude of the parameter B may best be determined experimentally, as it turns out to depend on the operating parameters of the mode-locked laser. In some cases it even changes in sign as the pump laser intensity is changed [51].

To phase lock the carrier envelope offset frequency ω_{CE}, one uses an actuator, in most cases an acousto-optic modulator, that drains an adjustable part of the pump laser power. Electro-optic modulators have also been used, but they have the disadvantage that they need to a bias voltage that wastes some of the pump energy to work in the linear regime. To servo control the phase of the ω_{CE} component usually requires much more servo bandwidth than locking the repetition rate. How much is needed in practice depends on the type of laser, the intensity and beam pointing stability of the pump laser and the phase detector in use. Mode-locked lasers that use pairs of prisms to compensate for group velocity dispersion generally show a much larger carrier envelope frequency noise. This is because intensity fluctuations slightly change the pulse round trip path because of the intensity-dependent refraction of the titanium-sapphire crystal [71]. Even small variations of the beam pointing result in a varying prism intersection. It should be noted that already $50\,\mu$m of extra BK7 glass in the path, shifts the carrier envelope phase by 2π and the carrier envelope frequency by ω_r. In most cases the carrier envelope frequency fluctuations seem to be dominated by the pump laser noise, so that stabilizing ω_{CE} with a modulator as described above even reduces this noise. Today titanium-sapphire lasers are mostly pumped by frequency doubled solid-state lasers that seem to show some differences between the models currently on the market [72]. Fiber-based mode-locked lasers used to have significantly larger noise in the carrier envelope beat note than titanium-sapphire lasers before the semiconductor pump lasers have been stabilized carefully [64].

In most cases a simple mixer is not sufficient to detect the phase of ω_{CE} relative to a reference oscillator as the expected in-loop phase fluctuations are usually much larger as for the ω_r servo. Prescalers or forward-backward counting digital phase detectors may be used to allow for larger phase fluctuations, that in turn allow the use of moderate speed (several 10 kHz) electronics. A complete circuit that has been used for that purpose very successfully is published in [73]. Stabilizing the carrier envelope frequency, even though it generally requires faster electronics, does not have the stability and accuracy issues that enter via the repetition rate due to the large factor n in eq. (6). Any fluctuation or inaccuracy in ω_{CE} just adds to the optical frequencies rather than in the radio frequency domain where it is subsequently multiplied by n.

None of the controls discussed here acts solely on either frequency ω_{CE} and ω_r. In general a linear combination of the two is affected. In practice this turns out to be not important because the different speeds of the two servo systems ensure that they do not influence each other.

Measuring the frequency of an unknown cw laser at ω_L with a stabilized frequency comb, involves the creation of yet another beat note ω_b with the comb. For this purpose the beam of the cw laser is matched with the beam that contains the frequency comb, say with similar optics components as used for creating the carrier envelope beat note. A dichroic beam splitter, just before the grating in fig. 5, could be used to reflect out the spectral region of the frequency comb around ω_L without effecting the beat note at ω_{CE}. This beam would then be fed into another set-up consisting of two polarizing beam splitters, one half wave plate, a grating and a photo detector for an optimum signal-to-noise ratio. The frequency of the cw laser is then given by

$$\omega_L = n\omega_r \pm \omega_{\mathrm{CE}} \pm \omega_b\,, \tag{26}$$

where the same considerations as above apply for the sign of the beat note ω_b. These signs may be determined by introducing small changes to one of the frequencies with a known sign while observing the sign of changes in another frequency. For example the repetition rate cold be increased by picking a slightly different frequency of the reference oscillator. If ω_L stays constant we expect ω_b to decrease (increase) if the "+" sign ("−" sign) is correct.

The last quantity that needs to be determined is the mode number n. If the optical frequency ω_L is already known to a precision better than the mode spacing, the mode number can simply be determined by solving the corresponding equation (26) for n and allowing for an integer solution only. A coarse measurement could be provided by a wave meter for example if its resolution and accuracy is trusted to be better than the mode spacing of the frequency comb. If this is not possible, at least two measurements of ω_L with two different and properly chosen repetition rates may leave only one physically meaningful value for ω_L [27].

4. – Scientific applications

In this section a few experiments in fundamental research where optical frequency measurements have been applied are discussed. Before the introduction of optical frequency combs only a few measurements of visible light could be carried out. Since then not only the available data has multiplied but also its accuracy has improved significantly. Maybe even more important, the frequency combs have enabled the construction of all optical atomic clocks that are treated in sect. **5**.

4˙1. *Hydrogen and drifting constants.* – The possibility to readily count optical frequencies has opened up new experimental possibilities. High-precision measurements on hydrogen have allowed for improved tests of the predictions of quantum electrodynamics and the determination of the Rydberg constant [14]. As the simplest of all stable atomic systems, the hydrogen atom, provides the unique possibility to confront theoretical predictions with experimental results. To explore the full capacity of such a fundamental test, one should aim for the highest possible accuracy so that measuring a frequency is imperative as explained in the introduction. At the same time one should use a narrow transition line that can be well controlled in terms of systematic frequency shifts. The narrowest line starting from the $1S$ ground state in hydrogen is the $1S$-$2S$ two-photon transition with a natural line width of 1.3 Hz and a line Q of 2×10^{15}.

At the Max-Planck-Institute für Quantenoptik (MPQ) in Garching measurements of this transitions frequency around 2466 THz has been improved over many years [13, 19, 74]. The hydrogen spectrometer used for that purpose is sketched in fig. 7. It consists of a highly stable frequency-doubled dye laser whose 243 nm radiation is enhanced in a linear cavity located in a vacuum vessel. The emission linewidth of the dye laser is narrowed to about 60 Hz by stabilizing it to an external reference cavity. This stabilization also reduces the drift rate below 1 Hz per second. The linear enhancement cavity ensures that the exciting light field is made of two counterpropagating laser fields, so that the Doppler

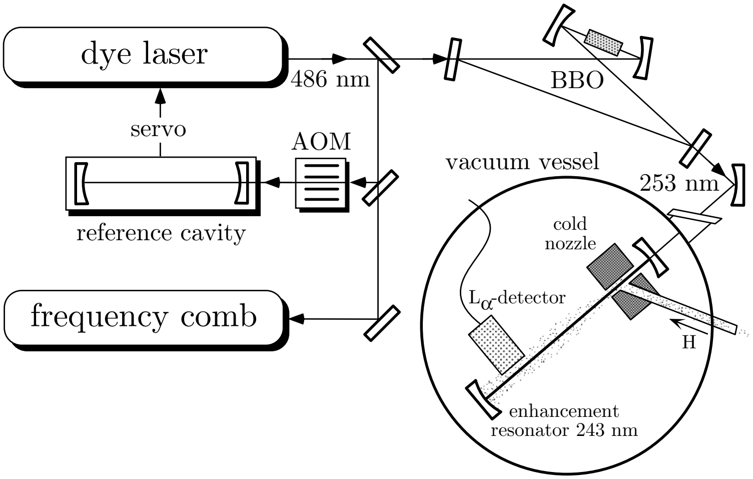


Fig. 7. – Exciting the hydrogen $1S$-$2S$ transition with two counterpropagating photons in a standing-wave field at 243 nm. This radiation is obtained from a dye laser frequency doubled in a BBO crystal and stabilized to a reference cavity. While scanning the hydrogen resonance the frequency of this laser is measured with a frequency comb to be 2 466 061 413 187 074 (34) Hz for the hyperfine centroid [74].

effect is cancelled to first order. Hydrogen atoms are produced in a gas discharge and ejected from a copper nozzle kept at a temperature of 6 K. After propagating the length of 13 cm, the excited atoms are detected by quenching them to the ground state in an electric field. The Lyman-α photon at 121 nm released in this process is then detected with a photomultiplier.

The optical transition frequency was determined in 1999 [13] and in 2003 [74] with a frequency comb that was referenced to a transportable cesium fountain clock from LNE-SYRTE, Paris [75]. At this time the repeated measurement did not yield an improved value for the $1S$-$2S$ transition frequency. Lacking a suitable laser cooling method is a particular problem for the light hydrogen atom. Even after thermalizing with the cold nozzle to 6 Kelvin, the average atomic velocity is $v = 360\,\mathrm{m/s}$ causing a second-order Doppler effect of $0.5(v/c)^2 = 7 \times 10^{-13}$ that needs to be accounted for. In addition the brief interaction of the atoms with the laser cause a variety of problems and can distort and shift the lineshape in an unpredictable way on the 10^{-14} level. In addition, because of the limited interaction time, larger laser power has to be used for a sufficient excitation rate. This increases to ac Stark shift and cause problems when extrapolating to zero laser power to find the unperturbed transition frequency. Nevertheless the inaccuracy is only about an order of magnitude larger than the best atomic clocks. A historic summary of hydrogen spectroscopy is given at the left side of fig. 8. As a remarkable aspect it should be noted that before the introduction of quantum electrodynamics basically every order of magnitude that measurement improved, required a new theory or at least some

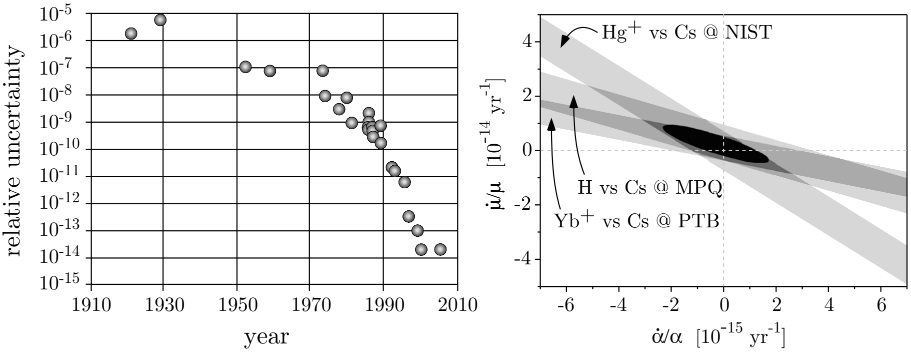


Fig. 8. – Left: a century of hydrogen data. Since the early 1950's when quantum electrodynamics was developed, measurements have been improved by almost 7 orders of magnitude and still no serious discrepancy has been discovered. Right: from the observation of transitions in single trapped ytterbium and mercury ions and the $1S$-$2S$ transition over several years, upper limits on small possible variations of the electromagnetic and the strong interaction can be found. The latter appears here as a variation of the cesium nuclear magnetic moment μ measured in units of Bohrs magneton.

refinement. Quantum electrodynamics by now has resisted a gain of almost 7 orders of magnitude without such a refinement. This is probably unprecedented for any physical theory. The development is summarized at the left panel of fig. 8.

The two comparisons of the hydrogen $1S$-$2S$ frequency with the LNE-SYRTE fountain clock may also be used to derive upper limits of possible slow variations of the fundamental interactions. The question of a possible time variation of fundamental constants([8]) was first raised in 1937 by P. A. M. Dirac, where he speculated that fundamental constants could change their values during the lifetime of the universe [77]. The traditional way to search for such a phenomenon is to determine the value of say the fine-structure constant as it was effective billions of years ago. For this purpose atomic absorption lines of interstellar clouds back-illuminated by distant quasars have been analyzed [78-81]. In a related method, the only known natural nuclear fission reactor that became critical some 2 billion years ago at Oklo, Gabon has been investigated [82-84]. Analyzing the fission products the responsible cross-sections and from that the fine-structure constant at the time this reactor was active can be deduced.

Now using frequency combs, high-precision optical standards can be compared with the best cesium clocks on a regular basis. Besides the hydrogen transition, the best monitoring data of this type so far derives from comparisons of narrow lines in single trapped mercury [17] and ytterbium [18] ions with the best cesium atomic clocks. Remarkably, the precision of these laboratory measurement makes it possible to reach the same sen-

([8]) For a review on this subject see ref. [76].

sitivity within a few years of monitoring that astronomical and geological observations require billions of years of look back time. The laboratory comparisons address some additional issues: Geological and astronomical observations may be affected by systematic effects that lead to partially contradicting results (see refs. [78,81,85,86] for contradicting results on quasar absorption and [83,84] for Oklo phenomenon data).

In the laboratory systematics can be investigated or challenged, and if in doubt, experiments may be repeated given the relative short time intervals. So far atomic transition frequencies have been compared with the cesium ground-state hyperfine splitting, which is proportional to its nuclear magnetic moment. The latter is determined by the strong interaction, but unlike the electronic structure, this interaction cannot easily be expressed in terms of the coupling constant. Lacking an accepted model describing the drift of the fundamental constants, it is a good advice trying to analyze the drift data with as few assumptions as possible. In previous analysis of the Oklo phenomenon, for example, it was assumed that all coupling constants but the fine-structure constant are real constant in time. However, if grand unification is a valid theory, at least at some energy scale, all coupling constants should merge and drift in a coordinated way [87]. Therefore the possibility that the cesium ground-state hyperfine splitting, *i.e.* the pace of the fountain clocks may have changed, should not be ruled out by the analysis. In this sense any of the optical frequencies monitored relative to the cesium clock can only put limits on the relative drift of the electromagnetic and the strong coupling constant. Fortunately the three mentioned comparisons show different functional dependences on the fine-structure constant leading to different slopes in a two-dimensional plot that displays the relative drift rates of the coupling constants (see right panel of fig. 8). The region compatible with this data is consistent with no drift at all, at a sensitivity level only a factor 2 away from the best astronomical observations that have been detecting a statistically significant variation [19]. It should be noted though that without a model for the drift, linearity in time cannot be assumed, so that the laboratory measurements do not even compare with the astronomical observations as they probe on different epochs. In the near future direct comparisons between different optical transitions will provide much better data because the cesium clock drops out and some optical transitions are getting more accurate than even the best cesium fountain clocks [9,88]. The frequency comb also allows this type of frequency comparison by locking one of the optical modes ω_n to the first optical reference and measuring the second with another mode $\omega_{n'}$. Ideally the carrier envelope offset frequency is stabilized to an integer fraction $1/m$ of the repetition rate, so that the ratio of the two optical frequencies derives as [89].

$$\frac{\omega_{n'}}{\omega_n} = \frac{n + 1/m}{n' + 1/m}\,. \tag{27}$$

No radio frequency enters in this comparison and possible beat notes can also be referenced to the repetition rate. Probably the most advanced experiment of this type would be the comparison of narrow transitions in aluminum and mercury ions operated in the group of J. Bergquist at NIST, Boulder. These standards are now reaching a reproducible within a few parts in 10^{17} [9]. To put this in perspective, it should be noted that the gravitational red shift at the Earth's surface is $g/c^2 = 1.1 \times 10^{-16}\,\mathrm{m}^{-1}$.

4·2. *Fine structure constant.* – Besides helping to detect a possible variation of the fine-structure constant α, the frequency combs are also useful to determine its actual value. All experiments to determine α have in common that a quantity that depends on it, is measured. The fine-structure constant is then determined by inverting the theoretical expression which usually comes as a power series in α.

As mentioned above, precision spectroscopy of hydrogen has led to an accurate experimental value for the Rydberg constant. Using all the available hydrogen data, *i.e.* the $1S$-$2S$ transition frequency and other less precise measurements [14], a value with an uncertainty of only 7 parts in 10^{12} is obtained [14, 90]. The Rydberg constant R_∞ can be traced back to other constants according to

$$R_\infty = \frac{\alpha^2 m_e c}{2h}\,, \tag{28}$$

so that the fine-structure constant α is derived as precise as m_e/h, the electron mass divided by Planck's constant, is known. This is because the speed of light c has a defined value within the SI units and enters with no uncertainty.

Currently the most precise measurement of the fine-structure constant has an uncertainty of 7 parts in 10^{10} [91]. This measurement is based on an experimental value of the electrons gyromagnetic ratio or more precisely its deviation from 2. This quantity can be calculated with quantum electrodynamics with comparable accuracy in terms of a power series in α. By comparison with the measured value, the fine-structure constant is determined.

Because α scales the strength of all electromagnetic interactions, it can in principle be determined with a large number of different experiments. Currently the second best method is based on the recoil an atom, such as cesium [92] or rubidium [93], experiences when absorbing an optical photon. Momentum conservation requires that the transition frequency is shifted by the kinetic energy associated with the photon momentum $\hbar k$ and the atomic mass M:

$$\Delta\omega = \frac{\Delta E}{\hbar} = \frac{\hbar k^2}{2M} = \frac{\hbar}{M}\frac{\omega^2}{2c^2}\,. \tag{29}$$

Measuring this recoil shift $\Delta\omega$ and the optical transition frequency ω yields a value for M/h. The recoil shift is typically only a few kHz on top of the transition frequency of several 100 THz. Therefore high-resolution optical frequency measurements are mandatory. To obtain α from (28) one additionally needs to know the mass ration m_e/M. Atomic mass rations can be measured very precise by comparing their cyclotron frequencies in Penning traps [94]. In fact such an effort was the motivation for the first frequency comb measurement performed with a femtosecond laser [15]. Note that all ingredients for deriving α via eq. (29) are obtained from precision frequency measurements, two of which optical frequencies.

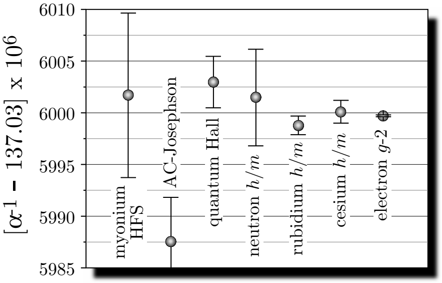


Fig. 9. – The agreement of various values, derived from different experiments is a crucial test for quantum electrodynamics. Currently the values nicely agree within their assigned uncertainty, except the one derived from the Josephson effect. For further details see [91, 90].

Even though less precise than the gyromagnetic ratio, the recoil measurements are important when it comes to testing of quantum electrodynamics. In general any theory that uses N parameters can only be said to have a predictive character if at least $N+1$ different experimental outcomes can be verified. The first N measurements are only fixing the parameters. In this context here the available data may be interpreted in a way where the gyromagnetic ratio fixes α without any verification of the theory. The recoil measurements are then interpreted as a test quantum electrodynamics, or the other way around. Currently the gyromagnetic ratio fits very well with recoil measurements in cesium and rubidium and other measurements. As fig. 9 shows, the only exemption may be a value derived from the Josephson effect, which does not seem to fit within its assigned error bar.

Finally it should be mentioned that of course quantum electrodynamics could be correct while calculations and/or measurements have some undetected errors. Calculating the electron gyromagnetic ratio as a function of α has so far used 891 Feynman diagrams. One might think that the recoil measurements does not show this problem, but deriving the Rydberg constant from hydrogen data involves a similar complex evaluation [95].

4˙3. *Optical frequencies in astronomy.* – In connection with cosmological search for a variation of what we believe are fundamental constants, optical frequency measurements are required on samples in the sky and on Earth as reference. Yet another type of observation relies on precise optical frequency measurements. To detect extrasolar planets the most powerful method has been to measure the changing recoil velocity of its star during the orbital period. These recoils velocities are rather small unless a massive planet in close orbit is considered. This is the reason why mostly "hot Jupiters" are among the roughly 200 extrasolar planets detected so far. The lightest of those planets possesses about 10 Earth masses. The wobble that our planet imposes on the motion of our Sun has a velocity amplitude of only $v_E = 9\,\text{cm/s}$ with a period of one year of course. Because Earth and Sun maintain their distance as they go around their common center of mass, this motion is invisible from Earth. On the other hand, it can be detected at other stars, where it is superimposed with the center-of-mass motion of that system of typically 100's of km/s and the motion of the Earth around the Sun. To detect Earth-like planets that orbit sun-like suns with the recoil velocity method, a relative Doppler shift of $v_E/c = 3\times10^{-10}$ needs to be measurable. Converted to visible radiation of say 500 THz this requires a resolution of 150 kHz and the same reproducibility after half the orbital time.

Spectral lines from atoms and ions from interstellar clouds and the surface of stars are subject to strong line broadening due to collisions and the Doppler broadening of typically several GHz due to their thermal motion. They are measured with telescopes like the Very Large Telescope operated by the European Southern Observatory which can be connected to an Echelle-type spectrometer for high resolution. Given these rather broad lines, the required spectral or velocity resolution can be obtained only by using the statistics of many lines observed simultaneously. It requires a spectrometer with very small irregularities in the calibration curve. So far the rather irregular line spectrum Tr-Ar lamps have been used for that purpose, with large spectral gaps. Using a frequency comb for this purpose appears to be the optimum tool, both in terms of providing an equidistant dense calibration and for allowing long-term reproducibility that goes well beyond the typical life time of an individual spectrometer [96]. The latter property derives from the possibility to reference to a precise clock. In this case even a simple GPS disciplined rubidium clock suffices for the required 3×10^{-10} reproducibility to detect Earth-like extrasolar planets.

In the long run one may even think about direct heterodyne detection with a frequency comb. In this scheme the star light is mixed with the frequency comb on a fast photo detector producing a radio frequency spectrum identical to the optical spectrum but shifted by the optical frequency. Signal processing can then be done with a radio frequency spectrum analyzer rather than a with an optical spectrometer. This type of detection is known for producing shot noise limited signals and is used to demonstrate noise levels below the shot noise limit with squeezed light. The frequency comb can provide a large number of optical local oscillators to shift any optical component within its span to the radio frequency domain.

4·4. *Reconstructing pulse transients and generating attosecond pulses.* – The application of the frequency combs has also enabled advances in another field as it allows the possibility for stabilizing the carrier envelope phase [3, 6]. According to fig. 2 and eq. (8), a pulse train with a vanishing carrier envelop frequency ω_{CE} has a fixed phase of the carrier with respect to the envelope. This means that all the pulses have the same electric field. With the technique of self-referencing this can be readily accomplished. However, even though the electric-field transients of the pulses are identical in this case, it is unknown what value the carrier envelope phase actually assumes upon stopping its pulse-to-pulse slippage. In order to figure out at what position the carrier envelope phase has actually stopped, all delays in the system including the optical path delays have to be known. Taking into account that the latter progresses by 2π for each wavelength propagation distance such an attempt seems unrealistic.

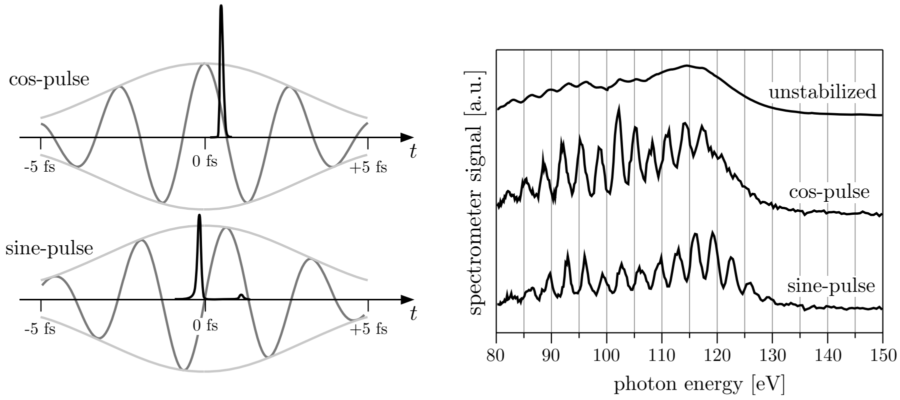


Fig. 10. – Left: infrared driving pulses of high intensity and the calculated intensity of high harmonic radiation around the short wavelength cut-off spectral region for two different values of the carrier envelope phase [97]. Whereas a "sine" pulse creates two high harmonic pulses, a "cos" produces a single isolated attosecond pulse [98]. Right: measured spectra (different intensity) for the two values of the carrier envelope phase. It is seen that the cut-off region looses its periodicity with the carrier wave as expected for single isolated attosecond pulses. Figures adapted from ref. [97, 99].

To detect the value of the carrier envelope phase very short pulses that drive processes which depend on the electric field in high order are used. The left-hand side of fig. 10 shows two extreme values of the carrier envelope phase that correspond to a "cos" and a "sine" pulse. Close observation of the field transients reveals that the peak electric field slightly depends on the carrier envelope phase. In addition the duration of the optical carrier cycle changes slightly due to the steep pulse envelope when measured say between two field maxima. These effects are largely enhanced if the pulses are short and can be detected using highly non-linear process such as "high harmonic generation" [100-102] or "above threshold ionization" [103]. High harmonics are generated if pulses of high intensity are focused into a noble-gas jet and are emitted in nicely collimated laser-like beam. Very short wavelengths up to the soft–X-ray regime have been produced this way [104]. Added to the left-hand side of fig. 10 is the calculated high harmonic radiation at a narrow bandwidth around the high-frequency cut-off. Two important aspects are realized from this: For "cos" driving pulse the high harmonic pulse is much shorter than one cycle of the generating field. So by stabilizing the carrier envelope phase to the proper value ("cos" drive pulses) single isolated attosecond could be produced for the first time [97, 99]. Secondly the high-harmonic spectrum reveals whether the carrier envelope phase has been fixed at the proper value or not. A "sine" drive pulse creates two high-harmonic pulses, that have almost but not quite the time separation given by the carrier frequency. For this reason the spectrum shows well-separated peaks even in the cut-off spectral region. However the positions of these peaks are not exactly harmonics of the infrared driving field. In contrast to that, a "cos" produces a shifted harmonic

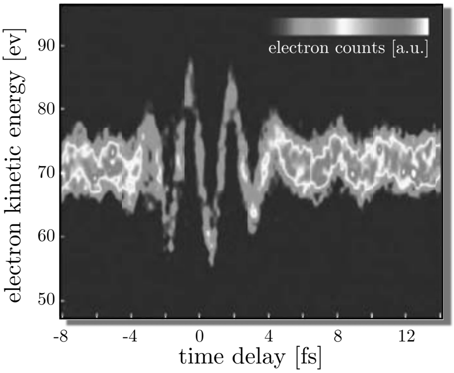


Fig. 11. – Cross-correlating attosecond pulses with the driving fundamental infrared laser pulses. The attosecond pulses are used to ionize atoms and the liberated electrons are accelerated by the instantaneous field of the infrared pulses. By changing the delay between the pulses, the electric-field transient can be sampled with sufficient temporal resolution. Figure courtesy of E. Goulielmakis *et al.* [105].

spectrum that looses its periodicity in the cut-off region. This is because it belongs to a single isolated attosecond pulse [98], as show in the upper left part of fig. 10. It should be noted that all of this applies only when the high-harmonic radiation is properly filtered. This should be obvious for two reasons: Only the highest harmonics posses enough non-linearity to distinguish the carrier envelope phase settings. Secondly when attosecond pulses are generated, their carrier frequency must be significantly larger than the inverse pulse duration.

Not only that the detection and stabilization of the carrier envelope phase allowed the production of attosecond pulses for the first time, it also allowed to completely recover the electric-field transients of ultrashort pulses. The work in that direction relied on measuring the pulses autocorrelation which, together with the determination of the carrier envelope phase via spectral analysis of high harmonic allows the calculation of the field transient [99]. A more direct measurement uses cross correlation between attosecond pulses generated the way described above with the driving pulses. In this way the field transient of the latter can be sampled with a temporal resolution significantly shorter than one optical cycle [105]. Figure 11 shows the result of such a measurement.

4·5. *Frequency comb spectroscopy.* – While for high-resolution spectroscopy of transitions such as the hydrogen $1S$-$2S$ single-mode lasers are currently employed, many transitions of fundamental interest occur at wavelengths too short for state-of-the-art continuous wave lasers. The boundary is set by the transparency range of the existing non-linear crystals. The crystal material that is useful for the shortest wavelength is BBO (β-barium-borate), with a transparency cut-off at about 190 nm. Only with pulsed lasers it is possible to efficiently convert into new regimes in the vacuum UV (200–10 nm), extreme UV (30–1 nm) and possible even the soft X-ray ($< 10\,\mathrm{nm}$)([9]). The process that allows this is high-harmonic generation [100-102], that was already discussed in the last section. For a long-time–pulsed lasers and high-resolution spectroscopy seemed to exclude each other because of the large bandwidth associated with short pulses. However, coherent train of pulses as compared to isolated or non-coherent pulses (say from a Q-switched laser) have quite distinct spectra. The coherent pulse train generated from a mode locked laser consists of narrow modes. It was shown in subsect. **2**˙3 that the Fourier and the Schawlow-Townes limit of the linewidth of an isolated mode is identical to the linewidth of a single-mode laser of the same type with the same average power. This opens the possibility to combine the high peak powers of mode-locked lasers, suitable for frequency conversion, with the properties of a sharp laser line produced by a single-mode laser.

One obvious problem with this approach is the limited power per mode in a broad frequency comb. Given an octave-spanning comb, as required for simple self-referencing, this can easily drop below 100 nW in practice. In addition the unused modes may cause excess ac Stark shift that poses a problem in high-accuracy measurements. Still these problems can be handled as demonstrated in ref. [106]. Another solution, at least for the UV spectral region, is to combine a fs mode-locked laser as a precise frequency reference with a ps mode-locked laser for spectroscopy [107]. The latter type of lasers can be converted in wavelength with efficiencies approaching unity even in single pass.

Yet an even better idea is to employ a two-photon transition as initially proposed by Ye. F. Baklanov and V. P. Chebotayev [108]. At first glance it seems that a two-photon transition does require even more power. However it is straightforward to see that in this case the modes can sum up pairwise such that the full power of the frequency comb contributes to the transition rate: suppose the frequency comb is tuned such that one particular mode $n\omega_r + \omega_{\mathrm{CE}}$, say near the center, is resonant with the two-photon transition. This means that two photons from this mode provide the necessary transition energy of $\omega_{eg} = 2(n\omega_r + \omega_{\mathrm{CE}})$. In this case the combination of modes with mode numbers $\{(n-1, n+1), (n-2, n+2), (n-3, n+3), \ldots\}$ are also resonant as they sum up to the same transition frequency. In fact all modes contribute to the transition rate in this way. The same applies if the two-photon resonance occurs exactly half ways between two modes. Figure 12 gives more details and a sample curve obtained from the $6S$-$8S$ two-photon resonance in cesium at 822 nm. It can be shown for unchirped pulses that the total two-photon transition rate is the same as if one would use a continuous laser with the same average power [108]. If the pulses are chirped, transition amplitudes corresponding to the various combinations of modes do no longer add up in phase so that the total transition rate is lower [109].

Experimentally frequency comb spectroscopy has been pioneered by J. Eckstein [110] and M. J. Snadden [111] and coworkers on sodium and rubidium, respectively, directly with a mode-locked laser. While the former experiment was still a factor 2.5 short

[9] These spectral ranges are not used in a unified way and partially overlap.

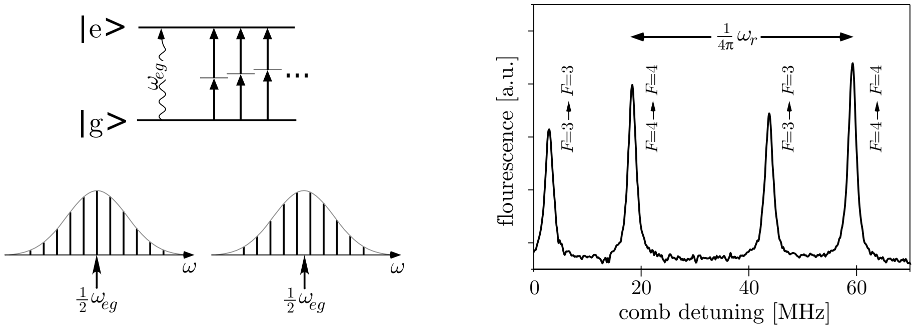


Fig. 12. – Left: the modes of a frequency comb add up to largely enhance the transition amplitude between levels $|g\rangle$ and $|e\rangle$ if only one of the central modes is resonant with a two-photon transition or that resonance occurs exactly half ways between two modes. Right: as an example, frequency comb excitation of the $6S$-$8S$ two-photon resonance is shown as the frequency comb is tuned across it. The hyperfine doublet with angular momentum F repeats with half the repetition rate of $\omega_r/2 = 2\pi \times 41\,\text{MHz}$ as expected from the resonance conditions [107].

of the 1.6 MHz natural linewidth, the latter reached the natural linewidth of 300 kHz. The main difference between the two measurements is that Snadden and coworkers used laser-cooled atoms in a magneto-optical trap.

This comparison reveals one disadvantage of the method. There may be a significant time-of-flight broadening because the Doppler free signal only emerges from atoms within the pulse collision volume for counterpropagating pulse trains. Atoms that fly through this volume can only absorb a limited number of pulses that cause the line broadening effects described in subsect. **2**˙3. To obtain a narrow resonance, it is therefore important to apply many pulses, ideally for a time that exceeds the inverse natural linewidth. When using fs pulses the collision volume may be smaller than 1 mm so that the atoms must be laser cooled and/or trapped as in ref. [111] to reach the natural linewidth. Note that the time-of-flight broadening may also be understood as a residual first-order Doppler effect. This comes about because, unlike the single-mode case, the Doppler shift $k_1 v$ and $-k_2 v$ for counterpropagating beams with wave numbers k_1 and $-k_2$ does not exactly compensate for the motion of the atom with velocity v. However for any residual first-order Doppler shift there exists a component obtained by exchanging the wave numbers $k_1 \rightleftharpoons k_2$ that possess the opposite shift. Hence a pure broadening and no shift results just as in the time domain description. This is certainly an advantage.

In the meantime several other groups have used laser-cooled atoms for frequency comb spectroscopy [112, 113]. S. Witte and coworkers used UV pulses at 212 nm generated in conventional non-linear crystals to excite a two-photon transition in krypton [114]. In this work only 3 pulses could be applied to the atoms so that the expected linewidth was about one third of the repetition rate or 23 MHz. Indeed that is about the observed linewidth which compares to the natural linewidth of 6.9 MHz. This experiment demonstrates the principle, knowing that improvements by many orders of magnitude should be possible. However, realizing these disadvantages of frequency comb spectroscopy, it appears advisable to use a continuous wave laser whenever possible. On the other hand, if there is no laser of this kind, frequency comb spectroscopy can become a very powerful method. One of the first demonstration in the short, so far inaccessible wavelength range, uses the ninth harmonic of a titanium-sapphire laser generated in a xenon gas jet to perform Ramsey-type spectroscopy on krypton at 88 nm [115]. The Ramsey method is identical to frequency comb spectroscopy with two pulses and produces narrow lines only when the two pulses come at a long delay (low repetition rate). In the meantime the same Ramsey-type spectroscopy has also been demonstrated at a one-photon transition in xenon at 125 nm [116]. The required radiation was obtained as the third harmonic produced in a gas cell. In this experiment the "repetition rate", *i.e.* the time separation between the Ramsey pulses was varied to increase the resolution.

Using high harmonics from titanium-sapphire lasers several 100 octaves of laser radiation may be addressed without gaps. Given the large tunability of these lasers allows to shift between the harmonics seamlessly. Therefore it might become possible to use a single laser system to cover everything from the near infrared to soft X-rays with atomic clock resolution. A possibility that seems out of reach for single-mode lasers. For such a general laser system there is yet another obstacle that needs to be solved. So far long pulse trains of many pulses at very short wavelengths with a high repetition rate necessary to resolve the modes have been contradicting requirements. This is because all methods up to very recently, that allowed to reach the necessary intensity for high-harmonic generation of typically $5 \times 10^{13}\,\mathrm{W/cm^2}$, effectively concentrate the available average power in fewer pulses per second. For this the repetition rate is typically reduced to the kHz regime. The modes of the resulting dense frequency comb would be very difficult to resolve. Even if the time-of-flight broadening could be reduced, say by using trapped ions and transitions with narrow natural linewidths, the requirements on the laser system would be difficult to achieve. For this reason a method allowing the production of high-harmonic radiation with MHz repetition rates was sought. The solution to this problem was to use an enhancement resonator for the driving pulses with an intracavity gas jet for high-harmonic generation [117, 118]. This method is similar to resonantly enhanced second-harmonic generation that has been used for many years. However, there are several extra requirements that need to be fulfilled in order to resonantly enhance fs pulses. First of all the pulse round trip time in that cavity has to be matched to repetition rate of the laser. In fact this is not difficult to accomplish and standard continuous wave locking schemes can be applied to stabilize the frequency comb to the modes of the enhancement resonator. Somewhat more difficult is it to prevent the pulse stored in the enhancement resonator from reshaping, or more precisely to reshape it such that after one round trip it will match the next pulse from the laser. As discussed in subsect. **2**˙1 this means that the carrier envelope phase shift of the cavity has to be the same as for the laser and higher-order dispersion has to be suppressed. In the frequency domain this simply means that the modes of the enhancement resonator must be equidistant in frequency just as the frequency comb is. In this case all modes can resonate at the same time. So far this problem has not been solved completely, so that the high-harmonic intensity generated this way still seems too low for a reasonable transition rate in stored ions. Fortunately this process scales very favorably with the driving intensity I. The so-called plateau harmonics scale with number of fundamental photons required for ionization, which is $\propto I^9$ for Xe driven with a titanium-sapphire laser around 800 nm.

5. – All optical clocks

Most likely the first time piece of mankind was the periodic movement of a shadow cast by a fixed object. The daily period would measure the Earth rotation and the superimposed annual period could be used to find the solstice and measure the length of the year in units of days. Later other periodic phenomena have been used. For an operational clock the periodic phenomena or oscillator has to be completed by a counter that keeps track of the number of periods. The early sundials had a human operator for counting the days. Later pendulum clocks used mechanical counters and todays precise clocks such as quartz or atomic clocks have electronic counters. If we look at the history of time keeping, it becomes obvious that clocks got more accurate as the oscillation frequency increased. One simple reason for that is that higher oscillation frequencies can slice time into finer intervals. Similar to a ruler that improves with the density of length marks.

The two most important properties of clocks are accuracy and stability. The former describes at what level two identically constructed clocks agree. The clocks stability, on the other hand, measures how well a particular clock maintains its pace. Both, accuracy and stability are determined by comparing at least two clocks and are limited by many factors such as the oscillators sensitivity to external perturbations. An increased oscillator frequency helps to improve stability. As an example consider the Earth rotation represented by a sundial and a good quartz oscillator. Whereas the two oscillators may be comparable in terms of accuracy, it would be much more difficult to measure a period of one second with a sundial than with a quartz oscillator that typically vibrates 32 768 times a second. Even faster than the vibration of a quartz standard is the precession of the Cs nuclear magnetic moment in the magnetic field of its electrons. This frequency has been defined within the International Systems of Units (SI) to be exactly 9 192 631 770 Hz.

To quantify the mutual stability of two oscillators a statistical analysis of a set of subsequent phase comparisons, obtained by averaging over time intervals τ, is used. For this purpose one may not use the standard deviation because it does not converge if the two clocks have slightly different frequencies, which is almost always the case. Instead the statistical analysis is done in terms of the Allan variance $\sigma(\tau)$ [119]. Having eliminated all other sources of instability an atomic oscillator, that obeys the laws of quantum mechanics, possesses the following expression for the Allan variance:

$$\sigma(\tau) = \frac{\Delta\omega_0}{\pi\omega_0}\sqrt{\frac{T_c}{2N\tau}}\,. \tag{30}$$

Here the frequency and the linewidth of the transition are given by ω_0 and $\Delta\omega_0$, respectively and the number of oscillators (atoms or ions) and the interrogation time is given by N and T_c. Modern Cs atomic clock can be operated close to that limit [75]. Of course the Allan variance will not reduce indefinitely for longer averaging times as suggested by the quantum limit. At some point varying systematic effects will set in and prevent further reduction. Current state-of-the-art Cs clocks need to be averaged for hours to reach that limit at around one part in 10^{15}. If one sets out to reach an accuracy of one part in 10^{18}, which is the predicted systematic uncertainty for some optical transitions, the necessary averaging times would extend by 6 orders of magnitude according to eq. (30). This of course is no longer useful for any practical purpose.

Fortunately the same equation suggests that one can make up for this by increasing the transition frequency ω_0 as illustrated above. At first glance it may seem that reducing the linewidth $\Delta\omega_0$ by choosing a narrow transition would serve the same purpose. In principle one could pick an almost arbitrary narrow transition, such as between ground-state hyperfine levels that basically live forever or highly forbidden transition such as the Yb^+ $^2\mathrm{S}_{1/2} \to {}^2\mathrm{F}_{7/2}$ transition, whose upper level lives for about ten years [120]. However this means that there will be one photon of information per lifetime available at most, assuming 100% detection efficiency. For this reason a much better strategy is to use a larger transition frequency such as an optical transition leaving the linewidth, *i.e.* the inverse transition rate fixed. Ideally one would choose an optical transition that represents an oscillator at frequency of several hundreds of terahertz. Operating at these high frequencies would have been possible after tremendous advances in laser spectroscopy in the 1970s that ultimately resulted in trapped atom and ion standards [121] in the 1980s. What was missing though was an efficient counter that keeps track of these fast oscillations. When it became possible to count these oscillations with the so-called harmonic frequency chains in the early 1970s [122], physicists started to seriously thinking of running an optical clock. However, working with these counters was so tedious that most of the harmonic frequency chains never reached the stage where they could operate continuously even for minutes. So it was decided to use them only to calibrate some chosen frequencies, like iodine or methane stabilized HeNe lasers, that could then be reproduced in other labs that could not provide the tremendous resources required for setting up a harmonic frequency chain. The calibrated lasers where then mostly used as wavelength references in interferometers for the realization of the meter and in some scientific experiments as frequency references.

With the introduction of femtosecond frequency combs a reliable running optical clock became reality. In particular the fiber-based frequency combs can now run for months with out touching them [63]. In addition excess noise observed in these lasers has recently been suppressed to the level of titanium-sapphire laser [64]. One of the first set-ups that deserved the term “optical clock” used a transition at 1064 THz in a trapped single mercury ion operated at NIST [89]. For a device deserving this name one would ask for some requirements on its accuracy and its ability to be operated long enough to calibrate a hydrogen maser for example. The NIST Hg^+ clock can be operated repeatedly with an uncertainty low enough such that comparison with an ensemble of very stable

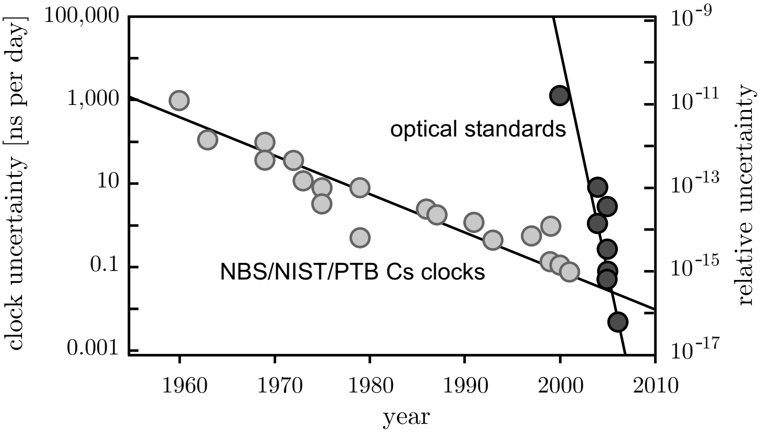


Fig. 13. – The historic comparison of radio frequency standards reveals that just about now the optical standards seem to take over the lead. The current lowest uncertainty is reached with the Hg^+ clock at NIST [9] with an estimated standard uncertainty of 7 parts in 10^{17}.

hydrogen masers resulted in an Allan variance of the latter [9]. By now many national standard institutes are working on optical atomic clocks. The Physikalisch-Technische Bundesanstalt, Germany (PTB) on Yb^+ [123], the National Physics Laboratory, UK (NPL) on Sr^+ and Yb^+ [11] and the Laboratoire national de métrologie et d'essais—Système de Références Temps-Espace, France (LNE-SYRTE) on neutral Sr [124] to name a few. The tremendous progress of the optical standards made since frequency combs where introduced is summarized in fig. 13.

To operate a frequency comb as a clockwork mechanism one might stabilize its nearest mode ω_k to the clock transition. Similar to the frequency ratio scheme of eq. (27), the carrier envelope offset frequency and possibly any local oscillator used in the stabilization, can be locked to an integer fraction m of the repetition rate [89].

$$\omega_k = k\omega_r + \omega_{\mathrm{CE}} = \omega_r(k + 1/m). \tag{31}$$

The countable clock output is then obtained in form of the pulse repetition rate that is given by the optical reference ω_k divided by $k + 1/m$. No other reference frequency besides the optical reference is required at this point as it should be for a real clock. In particular there is no radio frequency source required other than the repetition rate. To use the repetition rate in this way as a radio frequency generator is advantageous as it can be shown that, on a short time scale, this frequency is more stable than a good synthesizer even for a free-running laser [39].

6. – Optical frequency standards

Since the redefinition of the unit of length, the meter, in the International Systems of Units (SI) in 1983 by the 17th General Conference on Weights and Measures (CGPM) [125] the speed of light is fixed by definition to $c = 299\ 792\ 458\,\text{m/s}$. With the general relation $c = \lambda \times \nu$ any radiation emitting device whose frequency ν and hence its wavelength λ is stable and can be traced back to the frequency of the Cs clock can be used as a frequency standard or a wavelength standard. Such frequency standards have been used for decades for interferometric length metrology. The International Committee of Weights and Measures (CIPM) has therefore recommended a list of evaluated frequency standards, *i.e.* the so-called *Mise en Pratique* for the definition of the Metre [126] which has been updated on several occasions [125, 127, 128].

The first frequency standards in this list where mainly gas lasers whose frequencies were stabilized to the saturated absorption of suitable transitions in molecules. The selected molecules by natural coincidence have transitions whose frequencies are located in the tuning range of the laser. The most prominent examples are the CH_4 stabilized He-Ne laser at $3.39\,\mu$m, the OsO_4 stabilized CO_2 laser at $10.3\,\mu$m and the I_2 stabilized Ne-He or Nd:YAG lasers at 632 nm and 532 nm, respectively. While the first two of these lasers probably have the best reproducibility of all lasers stabilized this way, the much more compact and transportable iodine stabilized laser has been the working horse in length metrology for quite some time. The best of these lasers are reproducible within a few parts in 10^{13} and therefore cannot compete with the current Cs atomic clocks.

The possibility to cool atoms and ions by laser radiation has led to optical frequency standards that could rival the best microwave clocks. In these laser-cooled quantum absorbers suitable narrow transitions can be interrogated that are no longer excessively broadened and shifted by effects associated with the velocity of the absorbers, *e.g.*, the short interaction time and the influence of the second-order Doppler effect.

Optical frequency standards employing ions or neutral atoms as the frequency references show distinctive differences. Ions can be stored in a so-called radio frequency trap where an alternating electric voltage is applied to a suitable arrangement of electrodes that results in a net force onto the ions directed towards the field-free center of the ion trap. Since alike charged ions repel each other only a single ion can be kept in the center whereas the non-vanishing field outside the center in general shifts the transition frequency. Most accurate ion frequency standards therefore store only a single ion which, however, leads to a very weak signal.

To produce a much stronger signal and an associated much higher stability as compared to single ions ($N = 1$ in eq. (30)) the second type of laser-cooled optical frequency standards uses clouds of millions of neutral atoms. In this case though collisions between the atoms may result in uncontrollable shifts in frequency. In other words, so far (but not necessarily in the long run) single-ion standards have high statistical but low systematic uncertainties whereas neutral-atom standards have excellent short-term stability but poorer systematic properties.

Table I. – *Optical clock transitions recommended by the CIPM.*

Ion/Atom	Transition	Frequency/ wavelength	Fractional uncertainty
^{88}Sr$^+$	$5s\ ^2S_{1/2}$-$4d\ ^2D_{5/2}$	444 779 044 095 484 Hz 674 nm	7×10^{-15}
^{171}Yb$^+$	$6s\ ^2S_{1/2}$-$5d\ ^2D_{3/2}$	688 358 979 309 308 Hz 435 nm	9×10^{-15}
^{199}Hg$^+$	$5d^{10}\ 6s\ ^2S_{1/2}$-$5d^9\ 6s^2\ ^2D_{5/2}$	1 064 721 609 899 145 Hz 282 nm	3×10^{-15}
^{87}Sr	$5s\ ^2S_{1/2}$-$4d\ ^2D_{5/2}$	429 228 004 229 877 Hz 698 nm	1.5×10^{-14}

The recent progress with optical frequency standards in the mean time motivated the CIPM to recommend in autumn 2006 four optical frequency standards (see table I) that can be used as "secondary representations of the second" and be included into the new list of "Recommended frequency standard values for applications including the practical realisation of the metre and secondary representations of the second". The phrase "secondary representations" takes into account that the frequency of such an optical frequency standard can never surpass the accuracy of the primary standard of time and frequency, the Cs clock, but that it might be very important also for time keeping to use optical clocks because of their better stability. We will discuss the frequency standards of table I in more detail in the following subsections.

The uncertainties given in table I are in general larger than the uncertainties presented in the relevant publications since additional contributions have been taken into account caused for example by the uncertainty in linking to different Cs clocks.

6˙1. *Optical frequency standards based on single ions.* – The trapping and preparation of ions and their use for frequency standards has been described in detail in a number of reviews and textbooks (see, *e.g.*, [129-132]). An important feature however is the special kind of interrogation of single ions that will be discussed briefly in the following.

6˙1.1. Clock interrogation by use of quantum jumps. A properly chosen ion can get close to the ideal situation where a single unperturbed particle at rest is probed. The drawback though is that a single ion does not produce a strong signal by scattering many photons per second on the clock transition. However, the method of quantum jumps or electron shelving proposed by Dehmelt [133,134] and demonstrated later in Wineland's [135] and Dehmelt's group [136] is capable to detect a transition with unity probability. The technique is often applied to so-called V-systems where, as in fig. 14, a strong (cooling) transition and a weak (clock) transition are connected at the ground state. When the ion is irradiated by radiation whose frequency is in resonance with the strong transition up to about 10^8 absorption-emission cycles lead to the same number of emitted photons. Even

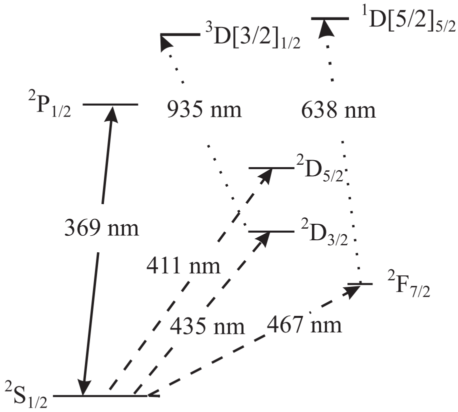


Fig. 14. – Excerpt from the energy level diagram of Yb$^+$. The transition at 369 nm is used for cooling the ion and for detecting quantum jumps (see text). Dashed lines (411 nm, 435 nm and 467 nm) represent transitions that were proposed for optical clocks.

with a limited detection probability of the order of 10^{-3} about 10^5 photons per second can be detected. If, however, the ion is excited on the weak clock transition by resonant radiation, the excited electron is "shelved" in the excited state. Consequently the fluorescence on the strong transition cannot be excited before the ion makes a transition to the ground state again. Thus the fluorescence light from the strong transition monitors the "quantum jumps" of the ion into the excited state and back as sudden intensity changes.

6˙1.2. Yb$^+$ and Hg$^+$ single-ion standards. The clock transitions of the ^{171}Yb$^+$ and ^{199}Hg$^+$ ions given in table I are quadrupole transitions with natural linewidths of 3.1 Hz and 1.1 Hz, respectively. The level scheme of both ions is basically very similar and hence both systems will be treated here at the same time. In Yb$^+$ the transition at 369 nm is used for cooling the ion and for detecting quantum jumps. The $^2S_{1/2}$-$^2D_{3/2}$ transition at 435 nm is preferred, rather than the $^2S_{1/2}$-$^2D_{5/2}$ transition at 411 nm since the $^2D_{5/2}$ level decay may occur into the long-lived $^2F_{7/2}$ that has a natural lifetime of about ten years. This level is connected to the ground state by an 467 nm octupole clock transition with a natural linewidth of the transition in the nanohertz range [120] and is also investigated as an optical frequency standard.

Two independent ^{171}Yb$^+$ single-ion optical frequency standards operating at the 435 nm line were compared at PTB to look for systematic frequency shifts. A significant contribution to the systematic uncertainty could result from the interaction of the atomic electric quadrupole moment with the gradient of an electric stray field in the trap. Two methods can be used to eliminate the quadrupole shift. One uses the fact that the quadrupole shift averages to zero in three mutually orthogonal directions of the magnetic quantization field [137]. A second method uses the fact that the average over all Zeeman and quadrupole shifts of all magnetic sublevels is zero [138].

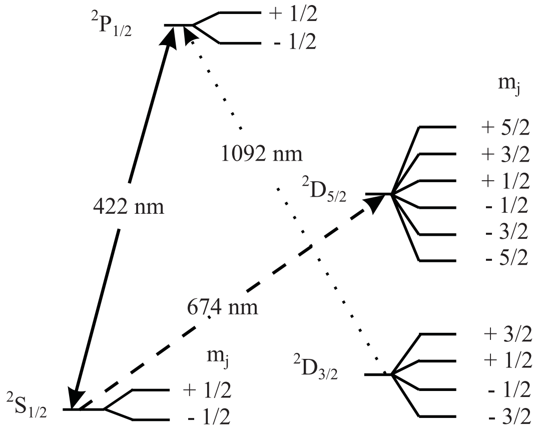


Fig. 15. – ${}^{40}\mathrm{Sr}^+$ energy level scheme.

Another possible systematic shift can result from the quadratic Stark shift. In the case of the ${}^{171}\mathrm{Yb}^+$ clock transition the scalar polarizability of the ground state, $\alpha_S({}^2S_{1/2})$, and from the scalar and tensorial polarizabilities of the ${}^2D_{3/2}$ state, $\alpha_S({}^2D_{3/2})$ and $\alpha_T({}^2D_{3/2})$ contribute to this effect. The contribution has been measured and is corrected for.

For the unperturbed clock transition a mean relative frequency difference of the frequencies between both standards of $(3.8 \pm 6.1) \times 10^{-16}$ was observed [88]. This agreement is similar to that of the most accurate comparisons between cesium fountain clocks. The frequency of the ${}^{171}\mathrm{Yb}^+$ standard was measured with a relative systematic uncertainty of 3×10^{-15} and a statistical uncertainty of 0.6×10^{-15} using a femtosecond frequency comb generator based on an Er^{3+}-doped fiber laser [139].

A similar level scheme of the cooling and clock transitions is present in the optical frequency standard based on the 1.06×10^{15} Hz (282 nm) electric quadrupole transition in a single trapped ${}^{199}\mathrm{Hg}^+$ ion that was developed at NIST. This frequency standard uses a cryogenic spherical electromagnetic (Paul trap) to trap a single Hg ion for up to 100 days. The frequency measurements over a period of more than five years between the ${}^{199}\mathrm{Hg}^+$ standard and the Cs primary clock transition agreed to better than 1×10^{-15} over three years [9]. A fractional frequency instability of about 7×10^{-15} at 1 s was demonstrated that follows the $\tau^{-1/2}$ dependence down to the 10^{-17} regime. The uncertainty budget presented from the NIST groups leads to a total fractional uncertainty of 7.2×10^{-17} and represents the smallest uncertainty to find the unperturbed line center reported for an optical clock. This uncertainty is an order of magnitude smaller as the uncertainty of the best cesium atomic clocks. As a result the frequency (in SI Hz) of such an optical clock can never surpass the uncertainty of the latter one.

6˙1.3. ${}^{88}\mathrm{Sr}^+$ single-ion standard. In contrast to the Yb^+ and Hg^+ single-ion standards discussed so far the ${}^{88}\mathrm{Sr}^+$ $5s\ {}^2S_{1/2}$-$4d\ {}^2D_{5/2}$ at 674 nm 444 779 044 095 484 Hz (fig. 15) of the even isotope is lacking hyperfine structure. Thus there is no magnetic-field insensitive $m_{F=0} \to m_{F=0}$ transition and a small external field splits the clock transition with a natural linewidth of 0.4 Hz into five pairs of Zeeman components.

This standard is investigated at the British National Physical Laboratory (NPL) [140] and at the National Research Council (NRC) [138] of Canada. In order to eliminate the first-order Zeeman shift, these groups probe alternately a symmetric pair of Zeeman components.

6˙2. *Neutral atom optical frequency standards.* – Neutral-atom–based frequency standards can be divided in three groups: the first one is based on free atoms in an atomic beam effusing from a nozzle and collimated by diaphragms.

6˙2.1. Atomic beam standards. The most advanced beam standard in the optical regime are the hydrogen standard based on the $1S$-$2S$ transition in atomic hydrogen [13, 19, 74] described above and the Ca standard based on the $^1S_0$-$^3P_1$ transition (see fig. 16) in atomic Ca [141-143]. The large velocity of the atoms leads to a large second-order Doppler effect which leads to a considerable broadening and shift of the interrogated line. Even though techniques that suppress the first-order Doppler effect were applied, residuals resulting from non-ideal alignment or curvature of the interrogating optical beams (residual first-order Doppler effect) occur that seem to prevent the use of atomic beam for use in optical frequency standards with the highest accuracy.

6˙2.2. Neutral atom standards based on ballistic atoms. The large influence of the first- and second-order Doppler effects can be reduced very efficiently by laser cooling the atoms, *e.g.*, in a magneto-optical trap. After shutting off the magneto-optical trap the released atoms have an initial velocity between a few cm/s and a m/s and follow ballistic paths in the gravitational field. Due to the free fall the interrogation of these atoms is limited to a few milliseconds in typical laser beams. The broadening of the interrogated line associated with such an interaction time corresponds to a few hundred hertz. In general, fountain geometries could be utilized to increase the interaction time to about a second [144]. Such a technique has been proposed a long time ago but has never been realized up to now.

For neutral ballistic atom standards therefore mostly alkaline earth atoms like Mg [145], Ca [146-149], Sr [150] have been investigated since there the intercombination transitions $^1S_0$-$^3P_1$ with their natural linewidths between a few tens of hertz and a few kilohertz are well adapted to the achievable interaction time-limited linewidth.

The most investigations and probably the best results have been obtained with calcium. The Ca intercombination transition $^1S_0$-$^3P_1$ (see fig. 16) has a natural width of 0.37 kHz which has been resolved to below the natural linewidth. In [148] a thermal beam of Ca atoms from an oven at 900 K is decelerated using a Zeeman slower from a velocity of 600 m/s to a velocity of about 40 m/s. Then the slow atoms are collimated and deflected by 10 degrees towards a magneto-optical trap (MOT) with a 2D molasses. This set-up avoids the black-body radiation from the hot oven which would lead to a so-called black-body shift [151]. At the first cooling step Ca atoms are cooled on the al-

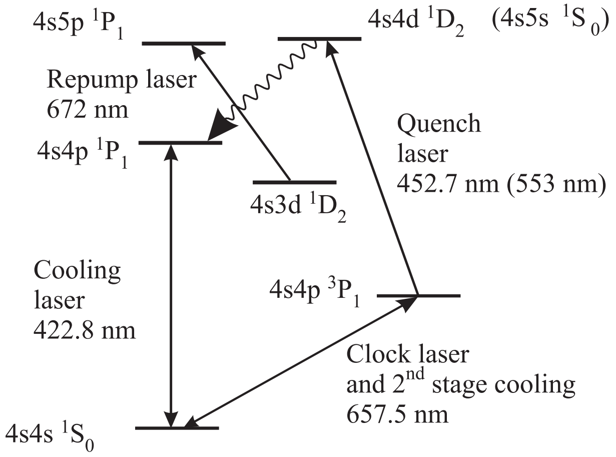


Fig. 16. – Excerpt of the ^{40}Ca energy level scheme. The transition at 423 nm is used for cooling the atoms. The clock transition (657 nm) is also used for second-stage cooling in combination with the quench laser at 423 nm or 553 nm.

lowed 423 nm transition $^1S_0$-$^3P_1$ (see fig. 16) and captured in a MOT. In 10 ms about 10^7 atoms are collected in a cloud at a temperature close to the Doppler limit of 0.8 mK for this cooling transition. A second cooling stage is performed using the forbidden 657 nm transition (see fig. 16). For efficient cooling the scattering rate is increased by quenching the upper metastable $^3P_1$ state by excitation of the 453 nm transition to the $4s4d\ ^1D_2$ state. With this quench cooling method the atoms are further cooled down to 12 μK, corresponding to an r.m.s. velocity of 10 cm/s. Then the clock transition is interrogated using a sequence of four pulses cut from a single cw laser beam. The first pair of pulses and the second one are antiparallel to allow for a first-order Doppler free excitation. The resolution is determined by the time between the first and the second and the third and the fourth pulse. This sequence represents the time domain analogue to the well-known optical separated field excitation [152] or a Ramsey-Bordé atom interferometer [153]. The necessary radiation to interrogate the clock transition with a spectral linewidth of approximately 1 Hz was provided by a master-slave diode laser system [154].

The frequency of the clock transition at 657 nm was referenced to the caesium fountain clock utilizing a femtosecond comb generator to be 455 986 240 494 144 (5.3) Hz in PTB with a fractional uncertainty of 1.2×10^{-14} and later at NIST with 6.6×10^{-15} [149], one of the lowest uncertainties reported to date for a neutral atom optical standard. These authors also concluded that this type of standard could be capable to achieve a fractional uncertainty in the 10^{-16} regime.

A fractional frequency instability of 4×10^{-15} at 1 s averaging time was achieved. As has been pointed out [39, 146] the quantum projection-noise limit [151] for the Ca standard corresponds to a fractional instability or Allan deviation of $\sigma_y(\tau) < 10^{-16}$ for

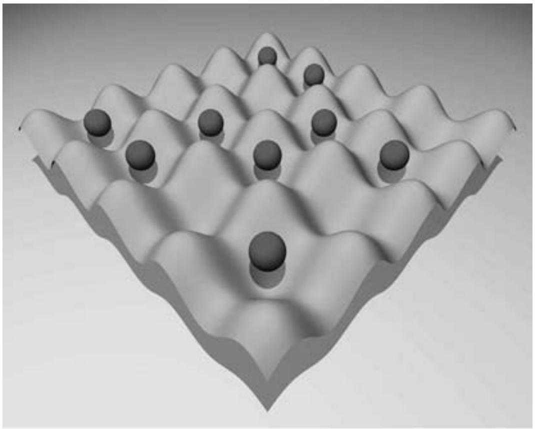


Fig. 17. – Atoms are trapped in the potential minima located at the maxima of a standing wave of the lattice laser.

$\tau = 1\,\mathrm{s}$. To approach this limit, a frequency instability of the spectroscopy laser of better than 3×10^{-16} for the duration of the atom interferometry $2T \approx 1.3\,\mathrm{ms}$ is necessary since the so-called Dick effect [155] degrades the stability through the aliasing of the laser frequency noise in a discontinuous interrogation. The Dick effect for an optical frequency standard with Ramsey-Bordé interrogation was first addressed by Quessada *et al.* [156]. Hence, the relatively short interaction time achievable with neutral atoms in free flight limits the accuracy and the stability.

6˙2.3. Optical lattice clocks. A possible way to combine the advantages of trapped single ions and the large number of neutral atoms, *i.e.* the long storage times and the good signal-to-noise ratio, respectively, was pointed out by Katori [157, 158]. A large array of micro traps for atoms (see fig. 17) can be generated in a standing wave laser field capable of holding maybe a million of atoms. Even though these optical lattices have been investigated for a long time they did not seem to be appropriate to be used in an optical frequency standard because of the large ac Stark shifts associated with the laser field. The remedy for this effect is to tune the lattice laser to a so-called "magic wavelength" where the upper and lower clock levels are shifted by the same amount, leaving the clock transition frequency unaltered (see fig. 18). The big advantage here is that the ac Stark shift can be controlled by controlling the lattice laser frequency rather than by measuring the lattice laser intensity. The former can be determined very accurately, in particular if one has an optical clock at hand.

Several atomic species are investigated for use as lattice clocks, *e.g.*, ^{87}Sr [10,159-161], Yb [162], and in the future also Hg. The best results have been obtained with ^{87}Sr which will be discussed in the following. The strongly forbidden $J = 0 \rightarrow J = 0$ transition $^1S_0$-

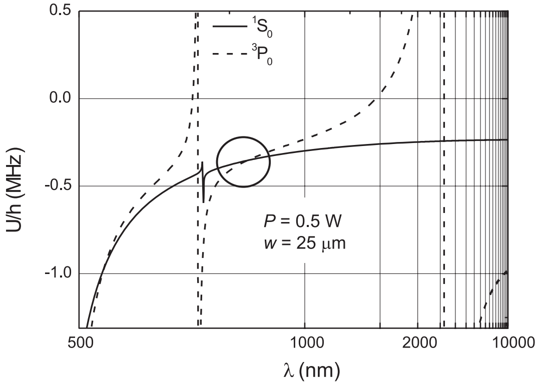


Fig. 18. – The calculated ac Stark shift potential energies U for a given power P and a given waist w of the lattice laser beam for the ground-state $^1S_0$ and the excited state $^3P_0$ are equal at the so-called "magic wavelength" (circle).

$^3P_0$ in Sr (see fig. 19) at 689 nm becomes weakly allowed through hyperfine mixing leading to a natural linewidth of about 1 mHz. The atoms are cooled on the 461 nm transition to several hundred microkelvin. To further reduce the temperature cooling on the $^1S_0$-$^3P_1$ transition is performed before the atoms are transferred into a dipole trap operated at the magic wavelength near 813.4 nm. In a one-dimensional lattice Boyd *et al.* [160] stored typically 2×10^4 atoms distributed across about 80 lattice sites. Interrogating the atoms several effects contribute to the uncertainty to find the unperturbed line center. These authors gave an uncertainty budget where the perturbations due to the residual ac Stark

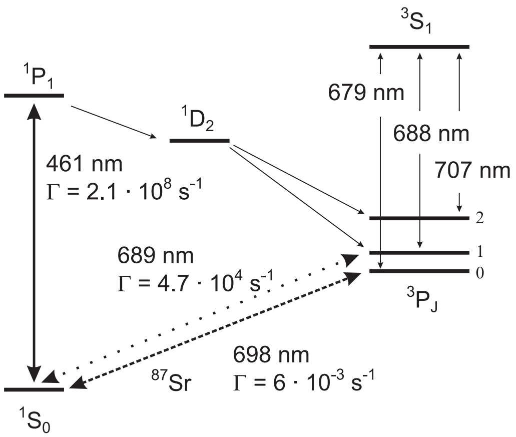


Fig. 19. – Excerpt from the energy level scheme of ^{40}Sr. The transitions at 461 nm and 689 nm are used for cooling the atoms. The dashed line (698 nm) is the clock transition.

shift of the lattice laser, the effect of the magnetic field necessary to separate the Zeeman levels, and possible collisions contribute the largest effects ending up with a fractional uncertainty below 1×10^{-15}. The frequency of this transition has been determined now by three independent groups [159-161]. The measurement of Boyd *et al.* [160] gave 429 228 004 229 874.0 (1.1) Hz. The associated uncertainty is already much smaller as the one recommended very recently by the CIPM (see table I) indicating the current speed of the evolution of optical clocks and frequency standards.

7. – Conclusions

After only a few years the measurement of optical frequencies by optical frequency combs is a mature technology that has led to a variety of novel applications. We are now in a situation where optical frequency standards and clocks are beginning to outperform traditional microwave clocks and most likely will lead in the future to a new definition of the SI unit second probably based on an optical frequency standard. Optical clocks in the visible part of the spectrum by no means need be the end of the evolution. Transitions in the UV or even in the XUV might give superior stability, as this important property scales as the transition frequency in use. It can be expected that the evolution in this field also in the future will keep its current pace since novel ideas like reading out of clock transitions by quantum information techniques [163, 164] or the use of nuclear transitions [165] will continuously shift the frontiers in optical frequency metrology.

REFERENCES

[1] Hall J. L., *IEEE Sel. Top. Quantum Electron.*, **6** (2000) 1136.

[2] Udem Th. *et al.*, in *Proceedings of the 1999 Joint Meeting of the European Frequency and Time Forum and the IEEE International Frequency Control Symposium, Besançon France*, IEEE catalog no. 99CH36313 (1999), pp. 602-625.

[3] Reichert J. *et al.*, *Opt. Commun.*, **172** (1999) 59.

[4] Reichert J. *et al.*, *Phys. Rev. Lett.*, **84** (2000) 3232.

[5] Diddams S. A. *et al.*, *Phys. Rev. Lett.*, **84** (2000) 5102.

[6] Jones D. J. *et al.*, *Science*, **288** (2000) 635.

[7] Holzwarth R. *et al.*, *Phys. Rev. Lett.*, **85** (2000) 2264.

[8] Udem Th., Holzwarth R. and Hänsch T. W., *Nature*, **416** (2002) 233.

[9] Oskay W. H. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 020801.

[10] Takamoto M. *et al.*, *Nature*, **435** (2005) 321.

[11] Gill P. and Margolis H., *Physics World*, **May** (2005) 35.

[12] *High Resolution Spectroscopy of Hydrogen*, in *The Hydrogen Atom*, edited by Bassani G. F., Inguscio M. and Hänsch T. W. (Springer Verlag, Berlin, Heidelberg, New York) 1989, pp. 93-102.

[13] Niering M. *et al.*, *Phys. Rev. Lett.*, **84** (2000) 5496.

[14] de Beauvoir B. *et al.*, *Eur. Phys. Lett. D*, **12** (2000) 61.

[15] Udem Th. *et al.*, *Phys. Rev. Lett.*, **82** (1999) 3568.

[16] Gerginov V. *et al.*, *Phys. Rev. A*, **73** (2006) 032504.

[17] Bize S. *et al.*, *Phys. Rev. Lett.*, **90** (2003) 150802.

[18] Peik E. *et al.*, *Phys. Rev. Lett.*, **93** (2004) 170801.

[19] Zimmermann M. *et al.*, *Laser Phys.*, **15** (2005) 997.
[20] Kourogi M. *et al.*, *IEEE J. Quantum Electron.*, **31** (1995) 2120.
[21] Brothers L. R. and Wong N. C., *Opt. Lett.*, **22** (1997) 1015.
[22] Imai K. *et al.*, *IEEE J. Quantum Electron.*, **34** (1998) 1998.
[23] Spence D. E., Kean P. N. and Sibbett W., *Opt. Lett.*, **16** (1991) 42.
[24] Diels J. C. and Rudolph W., *Ultrashort Laser Pulse Phenomena* (Elsevier, Amsterdam, Heidelberg, Tokyo) 2006.
[25] Eckstein J. N., *High Resolution Spectroscopy using Multiple Coherent Pulses* (Thesis, Stanford University, USA) 1978.
[26] *Frequency Standards in the Optical Spectrum*, in *The Hydrogen Atom*, edited by Bassani G. F., Inguscio M. and Hänsch T. W. (Springer Verlag, Berlin, Heidelberg, New York) 1989, pp. 123-133.
[27] Holzwarth R. *et al.*, *Appl. Phys. B*, **73** (2001) 269.
[28] Udem Th. *et al.*, *Opt. Lett.*, **24** (1999) 881.
[29] Diddams S. A. *et al.*, *Opt. Lett.*, **27** (2002) 58.
[30] Ma L. S. *et al.*, *Science*, **303** (2004) 1843.
[31] Stenger J. and Telle H. R., *Opt. Lett.*, **25** (2000) 1553.
[32] Krausz F. *et al.*, *IEEE J. Quantum Electron*, **28** (1992) 2097.
[33] Hasegawa A. and Tappert F., *Appl. Phys. Lett.*, **23** (1973) 142.
[34] Agrawal G. P., *Nonlinear Fiber Optics* (Academic Press, New York) 2001.
[35] Sutter D. H. *et al.*, *Opt. Lett.*, **24** (1999) 631.
[36] Morgner U. *et al.*, *Opt. Lett.*, **24** (1999) 411.
[37] Siegman A. E., *Lasers* (University Science Books, Mill Valley Ca USA) 1986.
[38] Bramwell S. R., Kane D. M. and Ferguson A. I., *Opt. Commun.*, **56** (1985) 112.
[39] Hollberg L. *et al.*, see fig. 9 in *IEEE J. Quantum Electron.*, **37** (2001) 1502.
[40] Rush D. W., Ho P. T. and Burdge G. L., *Opt. Commun.*, **52** (1984) 41.
[41] Bartels A. *et al.*, *Opt. Lett.*, **29** (2004) 1081.
[42] Swann W. C. *et al.*, *Opt. Lett.*, **31** (2006) 3046.
[43] Knight J. C. *et al.*, *Opt. Lett.*, **21** (1996) 1547.
[44] Ranka J. K. *et al.*, *Opt. Lett.*, **25** (2000) 25.
[45] Russell P. St. J., *Science*, **299** (2003) 358.
[46] Birks T. A., Wadsworth W. J. and Russell P. St. J., *Opt. Lett.*, **25** (2000) 1415.
[47] Holzwarth R. *et al.*, *Laser Phys.*, **11** (2001) 1100.
[48] Bartels A., Dekorsy T. and Kurz H., *Opt. Lett.*, **24** (1999) 996.
[49] Hoi M. *et al.*, *Phys. Rev. Lett.*, **96** (2006) 243401.
[50] Corwn K. L. *et al.*, *Phys. Rev. Lett.*, **90** (2003) 113904.
[51] Holman K. W. *et al.*, *Opt. Lett.*, **28** (2003) 851.
[52] Holzwarth R., PhD thesis Ludwig-Maximilians-Universität Munich (2001).
[53] Dudley J. M. *et al.*, *J. Opt. Soc. Am. B*, **19** (2002) 765.
[54] Ell R. *et al.*, *Opt. Lett.*, **26** (2001) 373.
[55] Fortier T. M. *et al.*, *Opt. Lett.*, **28** (2003) 2198.
[56] Matos L. *et al.*, *Opt. Lett.*, **29** (2004) 1683.
[57] Fortier T. M. *et al.*, *Opt. Lett.*, **31** (2006) 1011.
[58] Morgner U. *et al.*, *Phys. Rev. Lett.*, **86** (2001) 5462.
[59] Diddams S. A. *et al.*, *IEEE J. Quantum Electron*, **9** (2003) 1072.
[60] Fuji T. *et al.*, *Opt. Lett.*, **30** (2005) 332.
[61] Nelson L. E. *et al.*, *Appl. Phys. B*, **65** (1997) 277.
[62] Kubina P. *et al.*, *Opt. Express*, **13** (2005) 909.
[63] Adler F. *et al.*, *Opt. Express*, **12** (2004) 5872.
[64] McFerran J. J. *et al.*, *Opt. Lett.*, **31** (2006) 1997.

[65] See, for example, Gardener F. M., *Phaselock Techniques* (John Wiley & Sons, New York) 1979.

[66] Walls F. L. and DeMarchi A., *IEEE Trans. Instrum. Meas.*, **24** (1975) 210.

[67] Telle H. R., in *Frequency Control of Semiconductor Lasers*, edited by Ohtsu M. (Wiley, New York) 1996, pp. 137-167.

[68] Private communication T. Brabec.

[69] Haus H. A. and Ippen E. P., *Opt. Lett.*, **26** (2001) 1654.

[70] Xu L. *et al.*, *Opt. Lett.*, **21** (1996) 2008.

[71] Helbing F. W. *et al.*, *Appl. Phys. B*, **74** (2002) S35.

[72] Witte S. *et al.*, *Appl. Phys. B*, **78** (2004) 5.

[73] Prevedelli M., Freegarde T. and Hänsch T. W., *Appl. Phys. B*, **60** (1995) S241.

[74] Fischer M. *et al.*, *Phys. Rev. Lett.*, **92** (2004) 230802.

[75] Santarelli G. *et al.*, *Phys. Rev. Lett.*, **82** (1999) 4619.

[76] Uzan J. P., *Rev. Mod. Phys.*, **75** (2003) 403.

[77] Dirac P. A. M., *Nature (London)*, **139** (1937) 323.

[78] Webb J. K. *et al.*, *Phys. Rev. Lett.*, **87** (2001) 091301.

[79] Murphy M. T., Webb J. K. and Flambaum V. V., *Mon. Not. R. Astron. Soc.*, **345** (2003) 609.

[80] Quast R. *et al.*, *Astron. Astrophys.*, **417** (2004) 853.

[81] Srianand R. *et al.*, *Phys. Rev. Lett.*, **92** (2004) 121302.

[82] Shlyakhter A. I., *Nature*, **264** (1976) 340.

[83] Fujii Y. *et al.*, *Nucl. Phys. B*, **573** (2000) 377.

[84] Lamoreaux S. K. and Torgerson J. R., *Phys. Rev. D*, **69** (2004) 121701(R).

[85] Murphy M. T. *et al.*, arXiv:astro-ph/0612407v1.

[86] Murphy M. T., Webb J. K. and Flambaum V. V., arXiv:astro-ph/0611080 v3.

[87] Calmet X. and Fritzsch H., *Phys. Lett. B*, **540** (2002) 173.

[88] Schneider T., Peik E. and Tamm Chr., *Phys. Rev. Lett.*, **94** (2005) 230801.

[89] Diddams S. A. *et al.*, *Science*, **293** (2001) 825.

[90] Mohr P. J. and Taylor B. N., *Rev. Mod. Phys.*, **77** (2005) 1.

[91] Gabrielse G. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 030802.

[92] Wicht A. *et al.*, *Phys. Scr.*, **T02** (2002) 82.

[93] Cladé P. *et al.*, *Phys. Rev. Lett.*, **96** (2006) 03301.

[94] Bradley *et al.*, *Phys. Rev. Lett.*, **83** (1999) 4510.

[95] Pachucki K. and Jentschura U. D., *Phys. Rev. Lett.*, **91** (2003) 113005.

[96] Murphy M. T. *et al.*, sumitted to Mon. Not. R. Astron. Soc.

[97] Apolonski A. *et al.*, *Phys. Rev. Lett.*, **85** (2000) 740.

[98] Kienberger R. *et al.*, *Nature*, **427** (2004) 817.

[99] Baltuška A. *et al.*, *Nature*, **421** (2003) 611.

[100] Wahlström C. G. *et al.*, *Phys. Rev. A*, **48** (1993) 4709.

[101] Brabec T. and Krausz F., *Rev. Mod. Phys.*, **72** (2000) 545.

[102] Eden J. G., *Prog. Quantum Electron.*, **28** (2004) 197.

[103] Paulus G. G. *et al.*, *Phys. Rev. Lett.*, **85** (2000) 253004.

[104] Seres J. *et al.*, *Nature*, **433** (2005) 596.

[105] Goulielmakis E. *et al.*, *Science*, **305** (2004) 1267.

[106] Gerginov V. *et al.*, *Opt. Lett.*, **30** (2005) 1734.

[107] Fendel P. *et al.*, *Opt. Lett.*, **32** (2007) 701.

[108] Baklanov Ye. F. and Chebotayev V. P., *Appl. Phys. Lett.*, **12** (1977) 97.

[109] Meshulach D. and Silberberg Y., *Nature*, **396** (1998) 239.

[110] Eckstein J., Ferguson A. I. and Hänsch T. W., *Phys. Rev. Lett.*, **40** (1978) 847.

[111] Snadden M. J. *et al.*, *Opt. Commun.*, **125** (1996) 70.

[112] Marian A. *et al.*, *Science*, **306** (2004) 2063.
[113] Fortier T. M. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 163905.
[114] Witte S. *et al.*, *Science*, **307** (2005) 400.
[115] Cavalieri S. *et al.*, *Phys. Rev. Lett.*, **89** (2002) 133002.
[116] Zinkstok R. Th. *et al.*, *Phys. Rev. A*, **73** (2006) 061801(R).
[117] Gohle Ch. *et al.*, *Nature*, **436** (2005) 234.
[118] Jones R. J., Moll K. Thorpe M. and Ye J., *Phys. Rev. Lett.*, **94** (2005) 193201.
[119] Barnes J. A. *et al.*, *IEEE Trans. Instrum. Meas.*, **20** (1971) 105.
[120] Webster S. A. *et al.*, *Phys. Rev. A.*, **65** (2002) 052501.
[121] See, for example, contributions of Madej A. A., Bernard J. E., Riehle F. and Helmcke J., in *Frequency Measurement and Control*, edited by Luiten A. N. (Springer Verlag, Berlin, Heidelberg, New York) 2001.
[122] Evenson K. M. *et al.*, *Appl. Phys. Lett.*, **22** (1973) 192.
[123] Peik E., Schneider T. and Tamm Chr., *J. Phys. B*, **39** (2006) 145.
[124] Brusch A. *et al.*, *Phys. Rev. Lett.*, **96** (2006) 103003.
[125] Quinn T. J., *Metrologia*, **40** (2003) 103.
[126] Editor's note: *Documents concerning the new definition of the metre*, *Metrologia*, **19** (1984) 163.
[127] Quinn T. J., *Metrologia*, **30** (1993/1994) 523.
[128] Quinn T. J., *Metrologia*, **36** (1999) 211.
[129] Wineland D. J. *et al.*, *Phys. Rev. A*, **36** (1987) 2220.
[130] Blatt R., Gill P. and Thompson R. C., *J. Mod. Opt.*, **39** (1992) 193.
[131] Madej A. A. and Bernard J. E., *Frequency Measurement and Control: Topics in Applied Physics*, edited by A. N. Luiten, **79** (2001) 153.
[132] Riehle F., *Frequency Standards: Basics and Applications* (Wiley-VCH, Weinheim) 2004.
[133] Dehmelt H., *Bull. Am. Phys. Soc.*, **20** (1975) 60.
[134] Wineland D. J. and Dehmelt H., *Bull. Am. Phys. Soc.*, **20** (1975) 637.
[135] Wineland D. J. *et al.*, *Opt. Lett.*, **5** (1980) 245.
[136] Nagourney W., Sandberg J. and Dehmelt H., *Phys. Rev. Lett.*, **56** (1986) 2797.
[137] Itano W. M., *J. Res. Natl. Inst. Stand. Technol.*, **105** (2000) 829.
[138] Dubé P. *et al.*, *Phys. Rev. Lett.*, **95** (2005) 033001.
[139] Tamm Chr. *et al.*, to be published in *IEEE Trans. Instr. Meas.*
[140] Margolis H. S. *et al.*, *Science*, **306** (2004) 1355.
[141] Ito N. *et al.*, *Opt. Commun.*, **109** (1994) 414.
[142] Zibrov A. S. *et al.*, *Appl. Phys. B*, **59** (1994) 327.
[143] Kersten P. *et al.*, *Appl. Phys. B*, **68** (1999) 27.
[144] Beausoleil R. G. and Hänsch T. W., *Phys. Rev. A*, **33** (1986) 1661.
[145] Sterr U. *et al.*, *Appl. Phys. B*, **54** (1992) 341.
[146] Wilpers G. *et al.*, *Phys. Rev. Lett.*, **89** (2002) 230801.
[147] Sterr U. *et al.*, *C. R. Physique*, **5** (2005) 845.
[148] Sterr U. *et al.*, *Phys. Rev. A*, **72** (2005) 062111.
[149] Wilpers G., Oaetes C. W. and Hollberg L. W., *Appl. Phys. B*, **85** (2006) 31.
[150] Sorrentino F. *et al.*, arXiv:physics/0609133 v1 (2006).
[151] Itano W. M. *et al.*, *Phys. Rev. A*, **47** (1993) 3554.
[152] Bordé Ch. J. *et al.*, *Phys. Rev. A*, **30** (1984) 1836.
[153] Bordé Ch. J., *Phys. Lett. A*, **30** (1989) 10.
[154] Stoehr H. *et al.*, *Opt. Lett.*, **31** (2006) 736.
[155] Dick G. J., *Proceedings of 19th Annual Precise Time and Time Interval (PTTI) Applications and Planning Meeting, Redondo Beach, CA, 1987* (U. S. Naval Observatory) 1987, pp. 133-147.

[156] QUESSADA A. *et al.*, *J. Opt. B: Quantum Semiclassical Opt.*, **5** (2003) S150.

[157] KATORI H., IDO T. and KUWATA-GONOKAMI M., *J. Phys. Soc. Jpn.*, **68** (1999) 2479.

[158] KATORI H., in *Proc. 6th Symposium on Frequency Standards and Metrology*, edited by P. GILL (World Scientific, Singapore) 2002, p. 323.

[159] TAKAMOTO M. *et al.*, *J. Phys. Soc. Jpn.*, **75** (2006) 104302.

[160] BOYD M. M. *et al.*, *Phys. Rev. Lett.*, **98** (2007) 083002.

[161] LE TARGAT R. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 130801.

[162] BARBER Z. W. *et al.*, *Phys. Rev. Lett.*, **96** (2006) 083002.

[163] WINELAND D. J. *et al.*, in *Proc. 6th Symposium on Frequency Standards and Metrology*, edited by P. GILL (World Scientific, Singapore) 2002, p. 361.

[164] SCHMIDT P. O. *et al.*, *Science*, **309** (2005) 749.

[165] PEIK E. and TAMM CH., *Europhys. Lett.*, **61** (2003) 161.

This page intentionally left blank

Time scales and relativity

E. F. Arias

Bureau International des Poids et Mesures - 92312 Sèvres Cedex, France

1. – Introduction

The construction of the first caesium frequency standard at the National Physical Laboratory (UK) in 1955 is the origin of the atomic era in frequency referencing and timekeeping. It was quickly recognised that the caesium transition was the best choice and could serve as a reference for *frequencies*. The adoption of this transition for the definition of the second was more difficult, but did not raise fundamental objections. However, the acceptance of a time scale built by cumulating atomic seconds was not easily accepted; one argument was that atomic time differed from dynamical time (which is, in fact, true), but there was also the ancestral feeling that the motion of celestial bodies is time. A fundamental difference between dynamical timescales, based on planetary motions, and atomic time scales is that atomic time results from integration over frequency and that uncertainties are also integrated, leading to an unlimited departure from an ideally integrated time. In contrast, astronomical times are based on observations of positions of celestial bodies, with limited uncertainties, decreasing as a consequence of observational progress. At some time, the error on atomic time should exceed that on the reading of astronomical time. In spite of the opposition, international atomic time (TAI) was formally adopted in 1971, and a compromise between this continuous time scale and the apparent rotation of the celestial bodies was accepted with the definition of coordinated universal time (UTC). In those days, the precision of measurements did not require to appeal to relativistic theories. The progress of atomic time standards narrows the limits of what can be considered a local phenomenon or a local measurement, leading to a situation where the structure of an atomic clock cannot be seen as local. The modelling of a frequency standard itself in the framework of general relativity becomes a necessity.

The international timescales are calculated at the Bureau International des Poids et Mesures (BIPM). They are the result of the worldwide cooperation of national metrology laboratories and astronomical observatories that operate commercial caesium standards and that, in a smaller number, develop and maintain primary frequency standards. The algorithm used for the calculation of TAI has been designed to guarantee the reliability, the long term frequency stability, the frequency accuracy and the accessibility of the scale. It rests critically on the methods of clock comparison which are still the factor that can act to the detriment of a highly precise time scale.

In the near future we will be challenged by the comparisons of frequency standards accurate at a level at least one order of magnitude higher than the caesium fountains; these standards will be very probably called to provide a new definition of the second.

This text presents the evolution of timescales, from astronomy up to the atomic scale as realized today, on the framework of general relativity. A general description of the process of calculation of TAI and UTC is presented, together with the techniques used for current time and frequency comparisons.

The texts by Drs Andreas Bauch and Pierre Lemonde complement the notions provided in this lecture.

2. – Evolution of time scales and their associated interval units

In the 1950s, astronomical time scales were the standard for timekeeping. In those days, the precision of measurements did not require to appeal to relativistic theories. The goal of astronomers was to produce a time scale for worldwide use which was the best possible representation of absolute Newtonian time.

2˙1. *Universal Time (UT).* – Worldwide unification of time started in 1884, with the adoption of the Greenwich meridian as the origin of terrestrial longitudes, and a universal time associated to it [1]. The name "universal" time indicates any time scale based on the rotation of the Earth and referred to the prime meridian. The use of universal time was recommended by the International Astronomical Union (IAU) in 1948, even if it was of practical use since the end of the XIX century. In order to deal with the irregularities of UT, three kinds of "universal times" have been defined; UT0 is the time derived from astronomical observations, affected by the motion of the rotation axis on the Earth relative to the crust (known as polar motion), and by the irregular rate of the rotation of the Earth. These two effects are independent, and they can be clearly separated and studied.

The polar motion has two major components: a free oscillation with period of about 435 days, known as Chandler wobble, and an annual oscillation forced by the seasonal displacement of air and water masses. Figure 1 shows the trajectory of the polar axis, since 1900, as provided by the International Earth Rotation and Reference Systems Service (IERS). By eliminating from UT0 the effects of the polar motion, a form of universal time denominated UT1 is obtained. It is defined so that it is proportional to the rotation angle of the Earth in inertial space [2]; UT1 suffers from the irregularities of the rotation of the Earth; a secular deceleration and decade fluctuations.

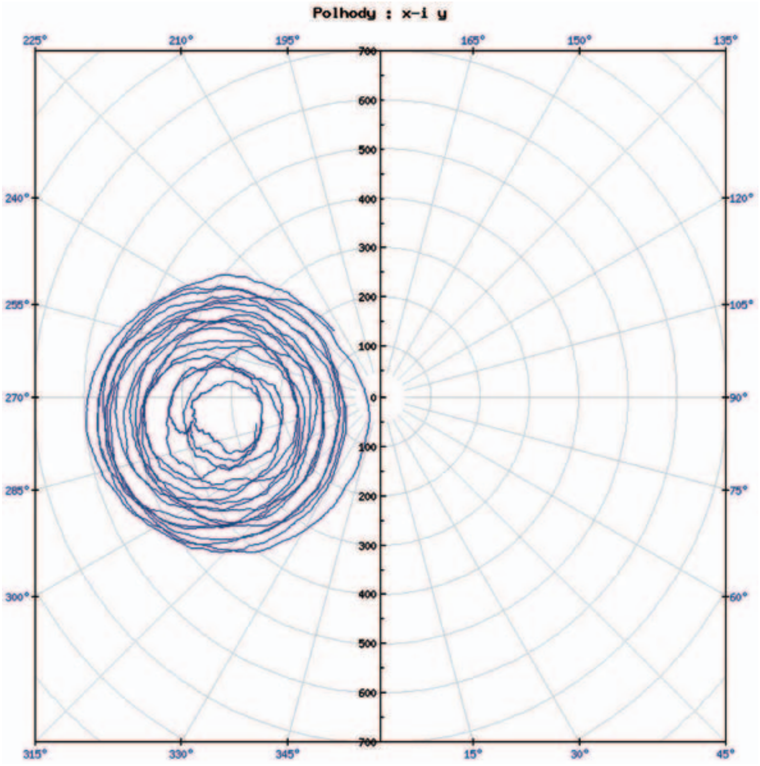


Fig. 1. – Path of the pole (plot provided by the IERS Earth Orientation Parameters Product Centre, Observatoire de Paris). The x-axis is oriented on the Greenwich meridian, the y-axis towards longitude 90° East. The unit of the central scale is the milliarsecond (mas).

The IERS has been charged since its creation, in 1988, of the monitoring of the irregularities of the Earth's rotation, such as has been done in the past by the Bureau International de l'Heure (BIH).

The variation of the length of the day, that is of the difference in duration of the rotational day from 86400 "standard" seconds, is at the millisecond order. These differences could be determined with precision after the arrival of atomic clocks.

The lack of uniformity of the realised time is characterized by an estimation of relative variations of rate, denoted by u, over specified intervals. Another important characteristic is the smallest uncertainty ε in assigning a date to an event. Table I gives the corresponding values of u and ε for the scale UT1 and for ephemeris time, which will be described in the next section [3]. Shorter-term fluctuations also exist, but they are not considered here because they could be smoothed out by use of good crystal clocks.

The unit of time (the second), was then defined as the duration of 1/86400 of a mean solar day. According to Leschiutta [4], this definition of the second was found for the first time in 1685! No metrological official definition existed for the second derived from the rotation of the Earth.

Time metrology was an activity developed in astronomical observatories. A time determination consisted in observing star transits to have the data necessary to calculate corrections to the clocks that had a rate different to the sidereal rotation. These corrections were added to the clock readings to obtain local time, and after adding the longitude, universal time. By extrapolation, the clock provided universal time on real time. Clearly, local realizations of universal time were different, and the need of an organization centralizing these activities was evident. After the end of World War I, the BIH was established at the Paris Observatory. One of the responsibilities of the BIH was, until 1988, to unify the time. Based on the contributions of the observatories, the BIH calculated the readings of an average clock, and referred the time signals emissions from the observatories to it. These offsets could reach, in some cases, 0.2 s.

2·2. *Ephemeris Time (ET).* – The confirmation of the irregularities of the Earth's rotation came from the study of the motion of planets and of the Moon. When comparing an ephemeris calculated with a time argument which is the uniform time of the Newtonian theory to the observed positions dated with a universal time scale, discrepancies were detected. It was concluded that these discrepancies arose from the non-uniformity of the Earth's rotation. In the early 1930s it seemed clear that the time embedded in the dynamical equations of the Solar System was a representation of uniform time and, after some fifty years of difficult research, Ephemeris Time ET, was defined in 1950, on the basis of the orbital motion of the Earth. The poor precision in positioning the Sun with respect to the stars made it impossible to exploit the good uniformity of ET in acceptable delays. A better precision of reading was obtained by defining a secondary ephemeris time by the motion of the Moon. But this motion, perturbed by ocean tides and geophysical phenomena, requires an empirical calibration against the fundamental ET which, although it extended over centuries, strongly limits the uniformity.

A new definition of the second was consequently associated to the orbital motion of the Earth, and it became a fraction of a particular tropical year. Its duration has been chosen so that it corresponded to the average duration of the second of mean solar time over the century. When this definition was adopted, in 1960 [5], the second of ET was shorter by 1.4×10^{-8} than the second of mean solar time.

The determination of Ephemeris Time was, in the practice, rather complex. It was determined, in post-real time, in the form of a correction to be applied to the universal time; it was based on the principle of the uniformity of the time argument of the Newtonian theory, and the difference between a calculated ephemeris and the corresponding observation at the date measured in universal time, served to represent the correction that should be applied to UT to derive ET. The use of Ephemeris Time remained limited to astronomical dynamics, and it never became a world wide used time scale.

During the short period of life of ET there was an inconsistency between the time scale used in practice and the unit of time. While the unit was associated to the dynamical time, UT remained the practical time scale. This situation persisted until the adoption of atomic time.

Table I. – *Relative lack of uniformity u and uncertainty of reading ε of astronomical time scales (1955).*

Time scale	u	ε/s
UT1 secular	5×10^{-11}/year	0.001
decade	4×10^{-8}	
ET (Earth orbit)	$\sim 10^{-11}$	~ 10
ETn (Moon orbit)	$\sim 5 \times 10^{-9}$	0.1

Table I gives an order of magnitude of u and ε for two realizations of ET: that trough the orbital motion of the Earth, and that defined via the orbital motion of the Moon.

2˙3. *The downing of atomic time.* – The essential quality of atomic frequency standards to be used in the development of atomic time scales is their accuracy. Defined in this context, accuracy is the ability of the standard to provide the natural frequency, or a known sub-multiple of the adopted atomic transition for the unperturbed atom. By postulate, this frequency is constant so that the time obtained by integration is uniform. Strictly speaking this applies at the location of the standard; in relativity, it is *proper time*, as we shall discuss later. A real standard has a defect in accuracy characterized by a relative uncertainty. On the long term, this uncertainty represents the departure from uniformity of the time scale generated by the standard.

The work leading to the construction of atomic clocks started in the 1940s with the experiments on magnetic resonance and the use of molecular beam techniques. In 1945 there was enough knowledge to build an atomic clock, but it could not be better than the crystal standards used at that time. The experiments of Ramsey [6] and his studies on the ways of minimizing errors in atomic clocks had been a fundamental to the construction in 1955 of the first caesium frequency standard by Essen and Parry at the National Physical Laboratory (NPL) in the United Kingdom. This standard was not a clock, it served to calibrate the frequency of an external quartz clock at intervals of a few days. In 1957 Essen and Parry published their measurements of the quartz ring oscillators and of astronomical time against the caesium second, and estimated that the frequency of the clock was known with a relative standard deviation of 2×10^{-10} in terms of the caesium resonance [7].

It was already admitted that the caesium frequency should be expressed in terms of the second of ET. It must be noted that any change of a unit of measurement must keep continuity with the previous definition; measurements made in the past would still be valid within the accuracy of the old unit. Markowitz and Hall at the United States Naval Observatory (USNO) undertook a precise determination of ET by a worldwide program of observation with the Markowitz Moon Camera. The value of $\nu_{\mathrm{Cs}} = 9\,192\,631\,770\,\mathrm{Hz} \pm 30\,\mathrm{Hz}$ was found [8], the uncertainty on the frequency being almost entirely due to ET. This value of ν_{Cs} was finally adopted for the definition of the SI second in 1967. The *Conférence Générale des Poids et Mesures* (CGPM) decided [9] that "the second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium-133 atom". This definition was completed, in 1997 by declaring that it refers to a caesium atom at rest at a thermodynamic temperature of 0 K.

Some fifty commercial caesium clocks operated in a few national laboratories in 1957-1958. They were less accurate than laboratory standards, but their very long term stability in frequency was excellent and they were apt to continuous operation as clocks. Several local independent atomic times TA(k) (this is an official designation, k being the name of the laboratory producing the scale) were established by integration over frequency, or by use of commercial caesium clocks, or by a combination of the data of both types of instruments. This raised two new problems: a) how to compare these time scales with accuracy compatible with that of the standards and b) how to average them to form a mean atomic time scale ensuring a better uniformity and reliability than individual scales.

The uncertainty of time comparisons based on radio time signals, of order of 1 ms, corresponds to a frequency uncertainty of a few units of 10^{-11} in relative value over one year; this was already much too large to exploit the accuracy of the standards at the end of the 1950s. Better frequency comparisons could be performed by referring the frequency of the standards to a common broadcast frequency at very low frequency (VLF), by measurement of phase variation. It was thus possible to construct a mean frequency standard and, by integration, a mean atomic time. But the uncertainty of time comparison between this mean time scale and local scales remained at the level of ±1 ms. A remarkable advance occurred in 1967 when the firm Hewlett-Packard demonstrated the possibility of using commercial flights to transport its caesium clocks allowing time transfer with an uncertainty of 1 μs in operational mode. Many calibrations of the VLF links were then performed by clock transportation, mostly by the USNO.

Mean atomic time scales based on international clock data were established at the USNO and at the BIH. The BIH scale became the International Atomic Time (TAI) in 1972. This scale is continuous since July 1955, although the methods of computation were adapted to the changing techniques of frequency and time comparisons.

The unification of time on the basis of the atomic time scale of the BIH was recommended by the International Astronomical Union (IAU, 1967), the International Union of Radio Sciences (URSI, 1969) and the International Radio Consultative Committee of the International Telecommunication Union (CCIR, 1970). Ultimate consecration came from the official recognition by the 14th CGPM in 1971 [10], which introduced the designation *International Atomic Time* and the universal acronym TAI.

2'4. *The birth of Coordinated Universal Time UTC.* – A clear problem at that time to accept universally TAI was to provide time to those users that required astronomical time; this was the case of sea navigation and other domestic applications. There was still a need to keep two different time scales, and to avoid the enormous confusion that this would create, Universal Coordinated Time (UTC) was defined by the International Radio Consultative Committee of the International Telecommunication Union (CCIR, presently

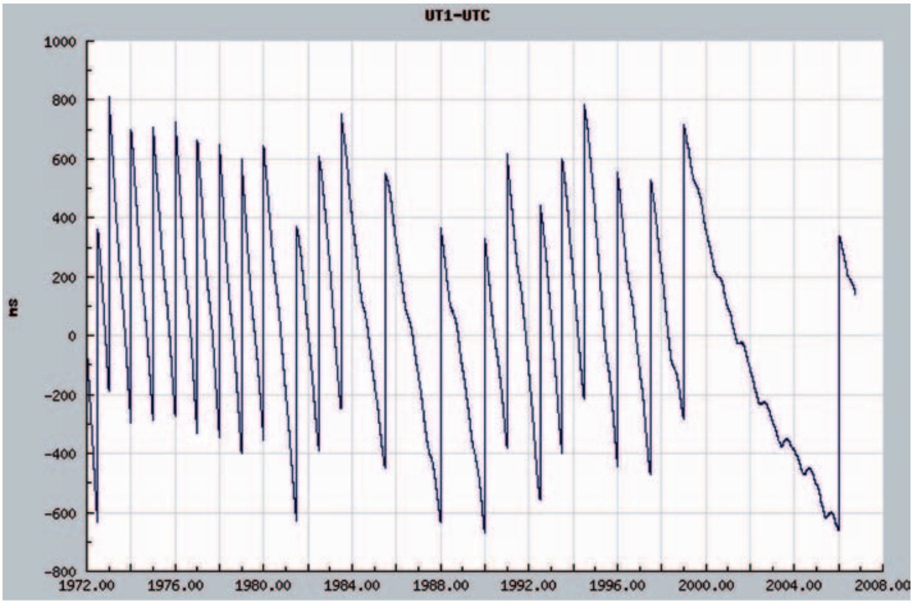


Fig. 2. – Values of $[UT1 - UTC]$ since 1972 (plot provided by the IERS Earth Orientation Parameters Product Centre, Observatoire de Paris). Unit is the millisecond (ms).

International Telecommunication Union, Radiocommunication Sector, ITU-R), and its use endorsed by the 15th CGPM in 1975 [11].

TAI was never disseminated directly and UTC, approximating UT1, continues to rule the world. The definition of UTC is provided by the tolerance for the time offset $[UT1 - UTC]$. Since 1972, UTC differs from TAI by an integer number of seconds, changed when necessary by insertion of a *leap second* to maintain $|UT1 - UTC| < 0.9\,\mathrm{s}$. Figure 2 shows the evolution of the difference $[UT1 - UTC]$, evaluated by the IERS.

TAI and UTC have numerous applications in time synchronization at all levels of precision; from the minute needed by the general public, to the nanoseconds required in the most demanding applications.

TAI is the basis of realisation of time scales used in dynamics, for modelling the motions of artificial and natural celestial bodies, with applications in the exploration of the Solar System, tests of theories, geodesy, geophysics, studies of the environment. In all these applications, relativistic effects are important.

UTC is the practical time scale, represented in real time by approximations given by clocks in national laboratories, and broadcasted by time signals. UTC serves as the basis of legal time in many countries.

3. – Metrologic qualities of a time scale

The phenomenon giving origin to a time scale should be reproducible with a frequency that is, ideally, constant. But this is never exactly the case, so we must be able to identify the causes of its variation, and to eliminate or at least minimize them. The realizations of the second of the International System of Units (SI) [12] differ from the ideal duration settled in its definition; we should be capable of reducing these differences in the process of construction a time scale. One solution is to average; we can average in time, but taking into account that since the measurements are not simultaneous, a process that keeps the frequency memory is essential. Or, we can average over the realizations of the second in different laboratories; in this case there are two sources of uncertainty: the individual experiments that converge to a realization, and the algorithm used to fabricate the scale.

TAI is an integrated time scale built by accumulation of seconds, which means that uncertainties are also accumulated. This was one of the criticisms at the time of discussing on the adoption of atomic time in replacement of ephemeris time. The main point was on the increasing discrepancy in the long term, which must be taken into account in some applications. No doubt was on the higher accuracy of TAI compared to ET.

The following elements are necessary for the construction of an integrated time scale, such as TAI:

a) a periodic phenomenon;

b) a definition of the unit, derived from the frequency of the phenomenon;

c) realizations of the unit, that is frequency standards, called clocks if they operate continuously; and

d) an algorithm of calculation adapted to the required characteristics of the scale.

A time scale is characterized by its reliability, frequency stability and accuracy, and accessibility.

The *reliability* of a time scale is closely linked with the reliability of the clocks whose measurements are used for its construction; at the same time, redundancy is also requested. In the case of the international reference time scale, a large number of clocks are required; this number is today about 300, most of them highly performing commercial caesium atomic standards and active auto-tuned hydrogen masers.

The *frequency stability* of a time scale is the capacity of maintaining a fixed ratio between its unitary scale interval and its theoretical counterpart. A means of estimating the frequency stability of a time scale is by calculating the Allan variance [13], which is the two-sample variance designed for the statistical analysis of time series, and depends on the sampling interval.

The *frequency accuracy* of a time scale is the aptitude of its unitary scale interval to reproduce its theoretical counterpart. After the calculation of a time scale on the basis of an algorithm conferring the required frequency stability, frequency accuracy is improved by comparing the frequency (rate) of the time scale with that of primary frequency standards, and by applying, if necessary, frequency (rate) corrections.

The *accessibility* to a worldwide time scale is its aptitude to provide a way of dating events for everyone. It depends on the precision that is required. We consider here only the ultimate precision which requires a delay of a few tens of days in order to reach the long term frequency stability required for a reference time scale. Besides, the process needs to be designed in such a way that the measurement noise is eliminated or at least minimized, this requires a minimum of data sampling intervals.

4. – Relativity and time scales

4˙1. *Proper and coordinate quantities.* – In the 1960s, general relativity was still a rather esoteric theory, except for a few specialists and cosmologists. Before general relativity, the time derived from the reproducibility of local phenomena and the time derived from the Newtonian dynamics were considered as representations of the same absolute time. In fact, Newtonian time is today adequate for most technical and scientific applications. A consequence of the accuracy of frequency standards is that the relativistic theory is now in current use in many applications. This has also led to a general reflection on metrology in the framework of general relativity [14].

In relativity we distinguish locally measurable quantities from coordinates that depend on a choice of conventions. The first is the case of *proper quantities* that result from observation or measurement without any particular choice of a space-time reference frame; they are valid in local experiments and observations. Proper quantities are directly measurable with a standard. *Coordinate quantities* depend on conventional choices, such as space-time reference frames or conventions of synchronization. In this case, the graduation unit varies with respect to proper units, since it depends on place and time. We will thus use the denomination *scale unit* to refer to the unit on the coordinate graduation.

Following these definitions, *proper time* is the (local) time measured by a clock. The time of an observer placed at the geocentre is also a form of proper time. Times based on the Newtonian dynamics are manifestations of *coordinate times.* Time is a coordinate in the space-time coordinate system arbitrarily chosen. The coordinate time difference between two events is dependent on the chosen reference system.

The second realized in a laboratory k by a standard fixed relative to it is the *proper second*, associated to the proper time of the laboratory k.

The time scale adopted as a reference worldwide, is a representation of *coordinate time.* International Atomic Time is this reference, whose scale unit is not the proper second (locally realized), but the proper second as realized on the rotating geoid.

The insertion of TAI in this general scheme was slow. Already in 1967 G. Becker from the Physikalisch-Technische Bundesanstalt (PTB) clearly defined the relation between the *proper unit of time* for local use and the time coordinate (*coordinate time*) in extended domains [15]. Although the need to correct the frequency of standards to refer them at sea level was recognized (the correction is about 1×10^{-13} per kilometre of altitude in relative value, for clocks fixed on the surface of the Earth), it was generally, but wrongly considered that TAI had the form of a proper time [16]. In 1980, the Consultative Committee for the Definition of the Second (CCDS) declared that TAI is coordinate time defined in a geocentric reference frame with the SI second as realised on the rotating geoid as the scale unit. The IAU did not accept this definition at its General Assembly of 1982 and preferred to wait for a global treatment of space-time reference systems. This was accomplished in the framework of general relativity in 1991, with the adoption of a specified metric [17]. In 2000, a metric extended to higher-order terms was adopted by the IAU [18]. In these developments, several theoretical coordinate times were defined, for use in the vicinity of the Earth and for the dynamics of the solar system. All these coordinate times are realised on the basis of TAI after relativistic transformations.

The necessity of a relativistic treatment of time comparisons was demonstrated in 1972 by an around-the-world transportation of caesium clocks [19]. In the context of general relativity there is no *a priori* definition of simultaneity of distant events, so that synchronization, or more specifically, the comparison of remote clocks is arbitrary. The formulae are based on a choice of a convention for synchronization, known as *coordinate synchronization.* This is the natural choice when clock comparison is to be used in the generation of coordinate time scales. According to Klioner [20], two events characterized by space-time coordinates (t_1, x_1, y_1, z_1) and (t_2, x_2, y_2, z_2) in some reference system are considered simultaneous if the values of the time coordinate are equal, that is, $t_1 = t_2$. If we follow this definition, synchronization depends on the choice of the space-time coordinate system, defined in a relativistic framework.

In general relativity, local physics keeps its familiar form, provided that the effects of special relativity are taken into account. It is however important to note that the progress of atomic time standards narrows the limits of "local". We arrive at a situation where the structure of an atomic clock cannot be seen as local. The modelling of a frequency standard itself in the framework of general relativity is becoming a necessity [21]. In applications, the need of a relativistic treatment appears as a consequence of precise measurement techniques based on atomic time and frequency standards: telemetry by lasers and radars, angular measurements by very long baseline interferometry. The fields which are concerned include celestial and terrestrial reference systems, geodesy, geophysics, positioning by satellite systems, dynamics in the Solar System (motion of planets, satellites, space probes), etc. Originally, the relativistic treatment often took the form of relativistic corrections to the Newtonian model. Subsequently, the work of experts in general relativity who have undertaken an educational effort [22, 23] has led to a much more satisfactory comprehensive treatment.

We can now say that general relativity is accepted as a "new classical" framework for metrology, geodesy and fundamental astronomy, which is a consequence of atomic timekeeping.

4'2. *Coordinate systems.* – The choice of the most suitable coordinate system is fundamental for time measurements and time and frequency comparisons. To cover the regions of the universe where time metrology applies, it will be necessary to make use of two non-rotating systems in space, one centred at the barycentre of the Solar System, the other with origin at the centre of mass of the Earth. Due to the dynamics of the Earth, a third system, rotating with the Earth will be used. The International Astronomical Union (IAU) gave in 1991 the basis for the definitions of the non-rotating systems [24].

In general relativity, there are no privileged systems; the choice of a system depends only on the convenience of having the simplest expression for the equations that represent phenomena. The space-time metric establishes a relationship between coordinate differences of two infinitely close events to the infinitesimal interval ds. The space-time coordinates $x^0 = ct$, x^1, x^2, x^3 are recommended to be chosen so that in a coordinate system with origin at the barycentre of any system of masses, the interval ds is expressed by

$$\text{(1)}\qquad \mathrm{d}s^2 = -c^2\mathrm{d}\tau^2 = -\left(1-\frac{2U}{c^2}\right)(\mathrm{d}x^0)^2 + \left(1+\frac{2U}{c^2}\right)\left[(\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2 + (\mathrm{d}x^3)^2\right],$$

where c is the speed of light, τ is the proper time and U the sum of the gravitational Newtonian potentials of the system of masses considered, of value zero at infinity, and the Newtonian tidal potential generated by the bodies external to the system, of value zero at the barycentre.

The consistent units to be used are the SI second for the unit of *proper time* and the SI metre for the unit of *proper length.* The metric (1) can be used without any loss of generality for time measurement on the Earth or in its vicinity; when representing a geocentric system, it generates relativistic errors of maximum order 10^{-18} at 3×10^5 km from the Earth, much smaller than the uncertainties in frequency standards (10^{-15} at present).

In the geocentric rotating system, the metric defined by eq. (1) can be written as

$$\text{(2)}\qquad \begin{aligned}\mathrm{d}s^2 = -\left(1-\frac{2U}{c^2}\right)(\mathrm{d}x^0)^2 + \left(1+\frac{2U}{c^2}\right)\times \\ \times\left[\mathrm{d}r^2 + r^2\mathrm{d}\phi^2 + r^2\cos^2\phi(\omega^2\mathrm{d}t^2 + 2\omega\mathrm{d}t\mathrm{d}L + \mathrm{d}L^2)\right],\end{aligned}$$

where r is the geocentric coordinate distance; ϕ is the geocentric latitude, measured from the equator; L is the longitude, reckoned positive towards the east; and ω is the angular velocity of rotation of the Earth.

4˙3. *Practical realizations of the coordinate systems.* – The barycentric non-rotating coordinate system is centred at the barycentre of the Solar System. Its axes are defined through a non-rotation relative to the directions to a set of extragalactic compact radio sources; it is realized through the International Celestial Reference Frame (ICRF) [25]. The coordinates of the some 700 radio sources in the ICRF are determined from observations of very long-baseline interferometry (VLBI) with an uncertainty at the submilliarcsecond level.

The geocentric non-rotating coordinate system axes are oriented parallel to those of the barycentric system, but it is centred at the centre of mass of the Earth. In this case the potential U in eq. (1) represents the gravitational potential of the Earth, plus the Sun-Moon tidal potential.

The third system, synchronous to the rotating Earth, has the x^3-axis oriented close to the rotation axis, while the axis x^1 points towards the origin of longitudes. It is represented by the International Terrestrial Reference Frame (ITRF) [26]. The coordinates and velocities of some 200 sites on the surface of the Earth, determined from space geodetic observations, represent the ITRF. The site coordinates are known with centimetric uncertainty.

The rotation of the ITRF relative to the ICRF is represented by the Earth orientation parameters; they are calculated from space techniques observations by the IERS [27].

4˙4. *Coordinate times.* – The coordinate $t = x^0/c$ represents, depending on the chosen system, barycentric or geocentric coordinate time.

The IAU defined in 1991 three coordinate times for use at the coordinate systems centred at the barycentre of the Solar System and at the centre of mass of the Earth. These definitions were extended in 2000 to be adapted to the improvement in accuracy of atomic clocks.

Barycentric Coordinate Time (TCB) is the time coordinate in the barycentric coordinate reference system (BCRS); Geocentric Coordinate Time (TCG) is the time coordinate in the geocentric coordinate reference system. The scale unit of TCB and TCG is consistent with the second of the SI. Physical realizations of TCB and TCG can be obtained from TAI.

The origins of TCG or TCB are related to TAI by the equation:

$$TCG\ (\text{or}\ TCB) - TAI = 32.184\,\text{s on 1 January 1977, 0 h TAI, in the geocentre.} \tag{3}$$

The value 32.184 has been conventionally adopted to assure the continuity with ephemeris time, and represents the number of seconds that TAI differed from ET on 1 January 1977.

Another coordinate time, denominated terrestrial time TT has been defined as having a scale unit with duration very close to the duration of the proper second on the rotating geoid. TT differs from TCG by a constant rate,

$$\frac{\mathrm{d}(TT)}{\mathrm{d}(TCG)} = 1 - L_G, \tag{4}$$

where L_G is a defining constant of value $6.969\ 290\ 134 \times 10^{-10}$.

TAI is a realization of TT, with an offset of 32.184 s, that is

$$TT = TAI + 32.184\,\text{s}. \tag{5}$$

Practically, TAI is computed on the basis of monthly packets of data, including measurements of primary frequency standards over the year preceding the month of calculation, whereas TT is the result of a computation using all available data from primary frequency standards. Each realization of TT is represented by TT(BIPMyy), where yy indicates the year of computation [28, 29]. The last realization TT(BIPM05) has been established using data from all available primary frequency standards until December 2005. TAI itself appears as a realisation of an ideal Terrestrial Time TT. Terrestrial Time is obtained from a Geocentric Coordinate Time TCG by a linear transformation chosen so that the mean rate of TT is close to the mean rate of the proper time of an observer located on the rotating geoid.

The barycentric coordinate system BCRS is associated to the coordinate time TCB. Its scale unit is based on the proper second. As stated in eq. (3), TCB has the same reading than TCG on 1 January 1977, 0 h TAI at the geocentre. TCB and TCG are related by the following expression, to order c^{-2}:

$$TCB - TCG = c^{-2}\left\{\int_0^t \left[\frac{v_E^2}{2} + U_P(\bar{x}_E)\right] \mathrm{d}t + \bar{v}_E \cdot (\bar{x} - \bar{x}_E)\right\}, \tag{6}$$

where $\bar{x}_E$ and $\bar{v}_E$ are, respectively, the barycentric position vector and velocity of the geocentre, $\bar{x}$ is the barycentric position vector of the observer, U_P is the gravitational potential provoked by the bodies of the Solar System except for the Earth, and t represents TCB.

5. – Realization of TAI and UTC

TAI is the reference time scale, defined in the context of general relativity as seen previously. It is calculated on the basis of differences of clock readings and of frequencies, by using an algorithm that follows the principles of Algos [30-32], the algorithm developed at the BIH in 1973 that fixed the principles of the construction of TAI. Since 1988 it has been calculated at the BIPM as a result of an international cooperation.

UTC is calculated as derived from TAI by the application of leap seconds. The dates of leap seconds of UTC are decided and announced by the International Earth Rotation and Reference Systems Service (IERS). The difference between TAI and UTC amounts to 33 s at the date of this publication, and this will remain the case at least until July 2007.

TAI is the uniform time scale that provides a precise reference for scientific applications, whereas UTC is the time scale of practical use that serves for international coordination in timekeeping and for the definition of legal national times.

5˙1. *An essential tool: clock comparisons.* – The calculation of a time scale on the basis of the readings of clocks located in different laboratories requires the use of methods of comparison of distant clocks. A prime requisite is that the methods of time transfer do not contaminate the frequency stability of the clocks, and in fact they often were in the past a major limitation in the construction of a time scale (see section on clocks).

The uncertainty of clock comparison varies today between ten nanoseconds and a nanosecond or even less for the best links, *a priori* sufficient to compare the best atomic standards over integration times of a few days. This assertion is strictly valid for frequency comparisons, where only the denominated type A uncertainty (evaluated by sta-

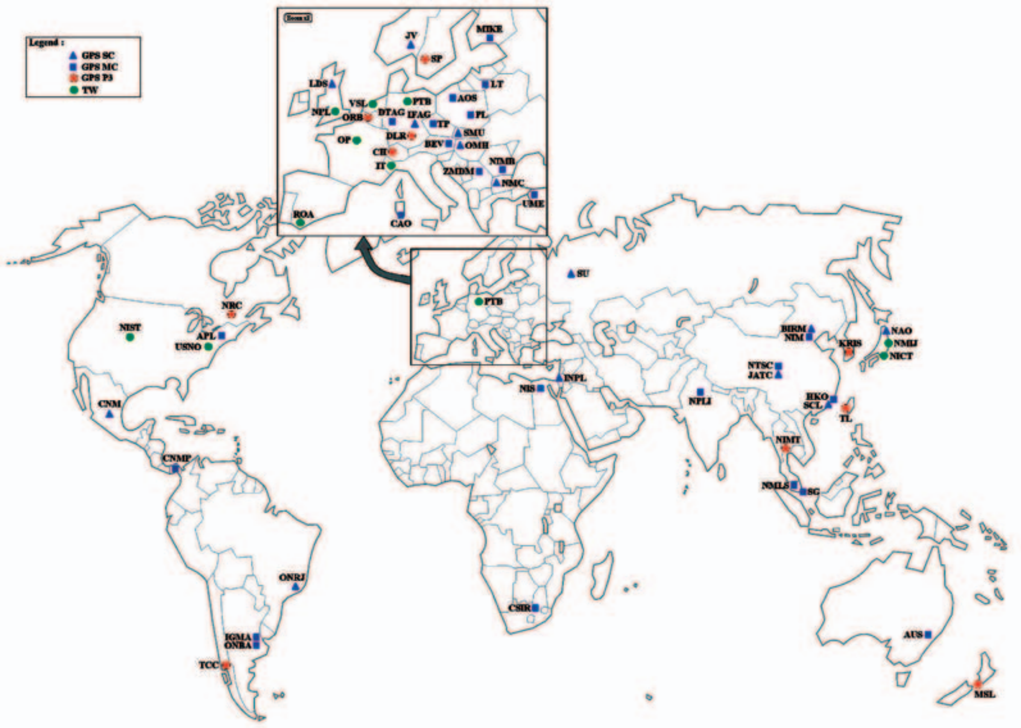


Fig. 3. – Laboratories participating to the calculation of timescales at the BIPM as in October 2006. Laboratories operating GPS equipment are compared to the PTB; laboratories equipped with TWSTFT stations in Europe and North America are compared to the PTB; those in the Asia-Pacific region are compared to NICT.

tistical methods) affects the process. In the case of time comparisons, the systematic uncertainty (type B), coming mainly from the calibration, should be considered in addition. For a complete description of the type A and type B uncertainties, refer to the *Guide to the expression of uncertainty in measurement* [33]. In the present situation, calibration contributes with an uncertainty that surpasses the statistical component, and that can reach 20 ns for non calibrated equipment or equipment with very old calibration (see table II). Repeated equipment calibrations are indispensable for clock comparison, and this is a task in which the BIPM puts many efforts.

A network of international time links has been established by the BIPM to organise these comparisons (see fig. 3). The structure of time links has recently changed as a consequence of the progress in clocks, in time transfer by GPS and the availability of clock products coming from the International GNSS Service (IGS) (refer to section "Use of GPS for time transfer"). Until October 2006 it was a star-like scheme with links from laboratories to a pivot laboratory in each continent, and long baselines providing the links between the pivot points. Since October 2006, all laboratories equipped with GPS receivers are compared to a unique pivot — the Physikalisch-Technische Bundesanstalt (PTB) in Germany.

The participating laboratories provide time transfer data in the form of a comparison of their local realizations UTC(k) with respect to another time scale (for example the time of the GPS) or to another local realisation of UTC.

5·2. *Use of GPS for time transfer.* – The use of the satellites of the Global Positioning System (GPS) since the early 1990s in time comparisons introduced a major improvement in the construction and dissemination of time scales. It consists in using the signal broadcast by GPS satellites, which contains timing and positioning information. It is a one-way method, the signal being emitted by a satellite with an atomic clock on board and received by specific equipment installed in a laboratory. For this purpose, GPS receivers have been developed and commercialized to be used specifically for time transfer. The common-view method proposed in the 1980s by Allan and Weiss [34] has been in use for the comparison of distant clocks in the last decades. It consists in the simultaneous reception of the same emitted signal at two Earth laboratories. This method has been developed and used to remove the common errors in the GPS signals coming from errors in the satellite position, instabilities of the satellite clocks, errors in the transmission of the signal between the satellite and the receiver, plus the effects of the intentional degradation (known as "selective availability") that was applied to the GPS signal until May 2000.

The most important error sources of clock comparison with GPS have been substantially reduced, making the method of common view less advantageous face to improvements that can be reached with a different strategy. The number of satellites used in a comparison is limited to those in common view, imposing also a sky distribution of satellites that biases the solution due to the non-homogeneous distribution of effectively used satellites. To compare clocks between laboratories at long distances, where common views are not possible, one or two intermediate laboratories are necessary, with a degradation of the quality of the final result.

The first generation of receivers used for time comparison were single-channel, single-frequency C/A code (Coarse Acquisition) receivers. In this case the information used is the code that modulates the L1 frequency of the carrier phase. The receiver's manufacturers developed later multi-channel receivers, operating also in one frequency, but allowing simultaneous observations of satellites over the horizon. The increasing number of this kind of receivers in laboratories improved the quality of time comparisons. The propagation of the signal, in the microwave region, is affected by the atmospheric effects, producing a supplementary delay. The ionosphere can provoke delays that will introduce significant errors, much bigger during periods of high solar activity. Dual-frequency reception eliminates the ionospheric delays increasing the accuracy of time transfer. Multi-channel, dual-frequency receivers are increasing in number in laboratories participating to the calculation of the timescales. These are geodetic-type receivers that provide the P-code (precise) observations on two frequencies denominated L1 and L2, and with a noise that is smaller than the noise on the C/A code.

As a consequence of these improvements, clock comparison by GPS time transfer can be achieved today directly between any two laboratories on the surface of the Earth with improved stability due to the increment in the number of measurements that can be used to compute the link.

Since September 2006, the method of common view has been replaced by a new one, named "GPS all-in-view", where each laboratory can be compared with a realization of the GPS time (the system time of the GPS constellation, calculated from an ensemble of clocks in the satellites and on Earth), or to the IGS time (a time scale calculated at the IGS, which is used as the reference for the IGS products) [35]. The current generation of TAI at the BIPM makes use of the GPS all-in-view method using the GPS time as the reference for comparison of all laboratories with the PTB.

Thanks to new hardware and to improvements in data treatment and modelling, the uncertainty of clock synchronization via GPS fell from a few hundreds of nanoseconds at the beginning of the 1980s to 1 ns or better today [36]. The effects of ionospheric delay introduce one of the most significant errors in GPS time comparison and in particular, in the case of clocks compared over long baselines. Dual-frequency receivers installed in some of the participating laboratories permit the removal of the delay introduced by the ionosphere, thus increasing the accuracy of time transfer. GPS observations with single-frequency receivers used in regular TAI calculations are corrected for ionospheric delays by making use of ionospheric maps produced by the IGS [37]. All GPS links are corrected for satellite positions using IGS post-processed precise satellite ephemerides.

The Russian satellite system of global navigation GLONASS is not yet used for time comparison in TAI on a routine basis, since the satellite constellation is in process of completion and it will become rapidly more stable. Studies conducted at the BIPM and in other laboratories prove [38, 39] that the system is potentially useful for accurate time transfer.

To facilitate the data exchange for time transfer and dissemination, directives on a common format and, standard formulae and parameters are given by the Consultative Committee for Time and Frequency (CCTF). Modern GPS and GPS/GLONASS receivers installed in national time laboratories that contribute to the calculation of TAI provide in an automated way time transfer data accordingly to these directives [40]. GPS transfer data is provided in the form of the time series of differences between a local realization UTC(k) and the GPS time.

5˙3. *Two-way satellite time and frequency transfer.* – After about 25 years of experimentation, the method of two-way satellite time and frequency transfer (TWSTFT) started to be extensively used in TAI at the beginning of the XXI century [41]. The TWSTFT technique utilizes a telecommunication geostationary satellite to compare clocks located in two receiving-emitting stations. Two-way observations are scheduled between pairs of laboratories so that their clocks are simultaneously compared at both ends of the baseline. The clocks are directly compared, using the transponder of the satellite. It has the advantage of a two-way method over the one-way method of eliminating or reducing some sources of systematic error such as ionospheric and tropospheric delays, uncertainty on the positions of the satellite and the ground stations. The differences between two clocks placed in the two stations are directly computed. The first TWSTFT link was introduced in TAI in 1999 [42]. Since then, the number of laboratories operating two-way equipment has increased, allowing links within and between North America, Europe and the Asia-Pacific region. Two-way observations are scheduled today on two-minute intervals every two hours, setting the type A uncertainty of time transfer below the nanosecond.

The time links by the TWSTFT technique have a different configuration than that of GPS links. Time comparisons within Europe and between Europe and the USA have the PTB as the pivot laboratory, while laboratories equipped with TW stations in the Asia-Pacific region are linked to the National Institute of Information and Communications Technology (NICT, Japan). UTC(NICT) and UTC(PTB) are compared to provide the link Europe-Asia.

5˙4. *Calibration of time transfer equipment and link uncertainties.* – Calibration of the laboratory's equipment for time transfer is fundamental for the stability of TAI and for its dissemination. Campaigns of GPS time equipment differential calibration are organised by the BIPM to compensate for internal delays in laboratories by comparing their equipment with travelling GPS equipment. Successive campaigns with BIPM travelling receivers are conducted on a permanent basis with the result that more than 70% of the GPS equipment used in TAI has been calibrated since 2001 [43-45]. The situation for the TWSTFT links is rather different; the laboratories organise with the support of the BIPM calibrations of the TWSTFT equipment [46-49] by using a portable TW station loaned by the Technical University of Graz (Austria). While waiting to have all stations thus calibrated, two-way links in TAI are calibrated at the BIPM by using the corresponding GPS link.

The BIPM estimates type A (u_{A}) and type B (u_{B}) uncertainties of all time links in TAI [50]. u_{A} is the uncertainty calculated from statistical methods by taking into account the level of phase noise in the raw data used for clock comparison; u_{B} is the uncertainty on the calibration. Some links have been selected to show examples of their values in table II, they are indicated with u_{A} and u_{B}, respectively.

For two decades, GPS C/A-code observations have provided a unique tool for clock comparison in TAI, rendering any test of its performance with respect to other methods impossible. The present situation is quite different; the introduction of the TWSTFT technique has allowed the opportunity of comparing the results of clock comparisons traditionally obtained with the GPS technique to those coming from an independent technique, and made the system more reliable. For the links where the two techniques are available, both GPS and TWSTFT links are computed; the best being used in the calculation of TAI, the other kept as a backup. The links with geodetic-type, multi-channel, dual-frequency receivers (indicated with the acronym GPS P3) have further increased the reliability of the system of time links, providing a method of assessing the performance of the TWSTFT technique. Comparison of results obtained on the same baselines with the different techniques shows equivalent performances for GPS geodetic-

Table II. – *Characteristics of some of the time links in TAI. The technique is indicated as follows: GPS MC for GPS multi-channel C/A data; GPS SC for GPS single-channel C/A data; GPS P3 for GPS multi-channel dual-frequency P code data; TWSTFT for two-way satellite time and frequency transfer data. Uncertainties u_A, u_B are described in the text. In the calibration type EC indicates equipment calibration, LC (technique) is for a link calibrated using the mentioned "technique", BC (technique) is used for a link calibrated using "technique" to transfer a past equipment calibration through a discontinuity in the operation of the link, NA stands for "not available". Conventional acronyms for the laboratories are developed in the BIPM reports.*

Link Lab 1 / Lab 2	Technique	u_A/ (ns)	u_B/n (s)	Calibration Type Lab 1 / Lab 2	Distance/km (approx.)
NIST/PTB	GPS SC	1.5	5.0	GPS EC/GPS EC	10000
NIST/PTB	TWSTFT	0.5	5.0	BC(GPS P3)	10000
OP/PTB	GPS SC	2.5	5.0	GPS EC/GPS EC	600
OP/PTB	GPS P3	0.7	5.0	GPS EC/GPS EC	600
OP/PTB	TWSTFT	0.5	1.0	LC(TWSTFT)	600
USNO/PTB	GPS P3	0.7	5.0	GPS EC/GPS EC	7000
USNO/PTB	TWSTFT	0.5	1.0	LC(TWSTFT)	7000
ONRJ/PTB	GPS SC	7.0	20.0	NA/GPS EC	10000

type dual-frequency receivers and TWSTFT equipment, when two-way sessions have a daily regularity (1 ns or less).

At the moment, 58 laboratories participate to the calculation of TAI at the BIPM, providing 57 time links. 84% of the links in TAI are obtained by using GPS equipment (63% with GPS single-frequency receivers; 21% with GPS geodetic-type dual-frequency receivers) and about 12% of the links are provided by TWSTFT observations. There still remain a small number of laboratories equipped with old-type receivers not adapted to provide data in the standard format.

6. – The algorithm Algos

6·1. *The general scheme.* – In the establishment and dissemination of time scales the quantities which intervene are only time and frequency differences at dates which can be loosely specified because these quantities vary slowly. In each participating laboratory k the approximation to UTC denoted UTC(k) is used as the reference for local clock differences and frequencies. The algorithm treats data over a 30-day period, with clock measurements available every five days, the so-called standard dates (Modified Julian Dates ended by 4 and 9). Every month, laboratory k provides to the BIPM with a clock file giving the differences between the readings of the clocks in the laboratory and UTC(k). Time transfer files provide the differences between UTC(k) and an external reference. Comparisons between laboratories j, k have the form of $[UTC(j) - UTC(k)]$. The dissemination of a global time scale T takes the form of time series of $[T - UTC(k)]$ at selected dates.

Making use of the clock and time transfer data, the algorithm Algos calculates an averaged time scale called *Free Atomic Time Scale* denoted EAL. This scale has an optimized frequency stability for a selected averaging time, but its frequency is not constrained to be accurate. Then, TAI is obtained by application of frequency corrections to EAL based on the data of primary frequency standards (PFS). The next step is to produce UTC by addition to TAI of an integer number of seconds (negative for the time being). The output of the process is $[UTC - UTC(k)]$.

6˙2. *Clocks in TAI.* – Fifty-eight time laboratories from 42 states members or associates to the Metre Convention participate in the calculation of TAI at the BIPM in November 2006. They contribute data each month from about 300 clocks. About 85% of clocks are either commercial caesium clocks or active, auto-tuned hydrogen masers. Commercial caesium clocks with high performance tubes realise the atomic second with a relative frequency accuracy of 1×10^{-13}, almost one order of magnitude better than the standard model, and they have an excellent long term frequency stability. Active hydrogen masers also benefit from high-frequency stabilities of the order of 10^{-15} over 1 day, but they do not serve to the realization of the second of the SI.

6˙3. *Free Atomic Time Scale (EAL), clock weighting and frequency prediction.* – To improve the stability of EAL, a weighting procedure is applied to the clocks. The weight of a clock is considered as constant during the 30-day period of computation and continuity with the previous period is assured by clock frequency prediction, procedure that renders the scale insensitive to changes in the set of participating clocks. The algorithm is able to detect abnormal behaviour of clocks and disregard them, if necessary; this is done in an iterative process that starts by the weights obtained in the previous month, and serves as an indicator of the behaviour of the clock in the month of computation. In the case of commercial caesium clocks, for averaging times around 30 days the predominant noise is random walk frequency modulation. All clocks in TAI are treated with this same frequency prediction model, but a revision appears to be necessary to take into account the increasing number of participating hydrogen masers, for which the predominant frequency noise is a linear drift (18% of the total number).

To avoid the possibility that very stable clocks indefinitely increase their weights and come to dominate the scale, a maximum relative weight is fixed for every period of calculation. Since January 2001, the maximum relative weight is fixed as a function of the number of participating clocks (N) as $\omega_{\max} = A/N$ (A is a constant equal to 2.5 at present), allowing a clock to reach the maximum weight when its variance computed from 12 consecutive 30-day samples is, at most, 5.8×10^{-15} [51].

The medium-term stability of EAL, expressed in terms of an Allan deviation [13], is estimated to be 0.6×10^{-15} for averaging times of 20 to 40 days. The frequency fluctuations of clocks that serve to characterize their weights are evaluated with respect to EAL.

6˙4. *Primary frequency standards (PFS).* – The accuracy of TAI is assured by the primary frequency standards developed in some laboratories reporting their frequency measurements to the BIPM. According to the directives of the CCTF, a report of a

Table III. – *Primary frequency standards having contributed to TAI since January* 2002. u_{B} *is the type B uncertainty as stated in the last report to the BIPM used for Circular T* [54], *expressed in* 10^{-15}.

PFS	Type	u_{B}
IT-CSF1	Cs fountain	0.8
LNE-SYRTE-FO2	Cs/Rb double fountain	0.61
LNE-SYRTE-FOM	Cs fountain	1.1
LNE-SYRTE-JPO	Optically pumped Cs beam	6.3
NICT-O1	Optically pumped Cs beam	5.5
NIST-F1	Cs fountain	0.36
NPL-CSF1	Cs Fountain	1.0
PTB-CS1	Magnetically defl. Cs beam	8.0
PTB-CS2	Magnetically defl. Cs beam	12.0
PTB-CSF1	Cs fountain	2.6

PFS should include the measurement of the frequency of the standard relative to that of a clock participating in TAI, and a complete characterization of its uncertainty as published in a peer-reviewed journal. Six caesium fountains and two optically pumped caesium beam standards have contributed, more or less regularly, in the last two years to TAI with measurements over 10 to 30-day intervals. Two magnetically deflected caesium beam standards of the PTB (CS1 and CS2) are operated in a continuous manner and contribute permanently to both the accuracy of TAI and the stability of EAL as a clock. Table III gives the main characteristics of these primary frequency standards. In 2006, the definition of the second of the SI is realised, at best, by the primary frequency standards with an accuracy of order 10^{-15}.

Based on the frequency measurements of the PFS reported to the BIPM during a 12-month period, the fractional deviation d of the unitary scale interval of TAI from its theoretical value (the unitary scale interval of TT) is evaluated, together with its uncertainty [52]. A filter is applied to the individual measurements; this filtering process takes into account the correlation terms of successive measurements reported for the same standard and a model of the frequency instability of TAI [53].

In order to keep the unitary scale interval of TAI as close as possible to its definition, a process called *frequency steering* has been implemented. It consists in applying a correction to the frequency of EAL when d exceeds a tolerance value, generally fixed to 2.5 times its uncertainty. These frequency corrections should be smaller than the frequency fluctuations of the time scale in order to preserve its long-term stability. Over the period 1998-2004, frequency steering corrections of $\pm 1 \times 10^{-15}$ have been applied, when necessary, for intervals of two months at least. The values of d demonstrated that the unitary scale interval of TAI had significantly deviated from its definition and that the steering procedure was in need of revision. A different strategy for the frequency steering was adopted in July 2004. A frequency correction of variable magnitude, up to 0.7×10^{-15} is applied for intervals of one month at least, if the value of d reaches 2.5 times its uncertainty.

7. – Secondary representations of the second

The second of the SI is defined as derived from the frequency of a caesium transition, and realized at best at national time laboratories maintaining caesium fountains. In the last years there has been a rapid improvement in the development of optical frequency standards that realize frequencies with uncertainties comparable to those of caesium standards. These devices could in the future provide realizations of the second with an accuracy at least one order higher than that of the present realization.

The Consultative Committee for Time and Frequency (CCTF), considering that new frequency standards based in other microwave transitions and on optical transitions could eventually be considered as the basis for a new definition of the second, recommended in 2001 to examine and approve accurate frequency measurements of atom and ion transition frequencies made relative to the caesium frequency standard as secondary representations of the second. To accomplish this task, a working group that puts together experts working in the domain of time and frequency, meets regularly to study reports of frequency measurements to be included in a list of secondary representations of the second. For the value of the unperturbed frequency of a quantum transition to be accepted as a secondary representation of the second, its uncertainty must be evaluated and documented to meet the requirements for the primary frequency standards for use in the formation of TAI; in any case, this uncertainty should be no larger than about a factor of ten of the primary frequency standards that serve as the best realization of the second. Following these criteria, only frequency standards with accuracies of around one part on 10^{14} should be considered [55].

As in November 2006, five transition frequencies have been recommended by the CCTF and endorsed by the International Committee for Weights and Measures (CIPM). Although these systems can prove to have reproducibility better than the caesium, they are limited in uncertainty by the caesium, since they are secondary representations. The list includes the Rb microwave standard as evaluated by the LNE-SYRTE [56] with an estimated relative uncertainty of 3×10^{-15}, the $^{88}\mathrm{Sr}^+$ ion transition as reported by the NPL [57] and the NRC [58] with an estimated relative uncertainty of $7{\times}10^{-15}$, the $^{199}\mathrm{Hg}^+$ as evaluated at the NIST [59] with an estimated relative uncertainty of 3×10^{-15}, the $^{171}\mathrm{Yb}^+$ ion transition as reported by the PTB [60] with an estimated relative uncertainty of 9×10^{-15}, and the $^{87}\mathrm{Sr}$ neutral atom transition reported by the NICT [61] and the JILA [62, 63] with an estimated relative uncertainty of 1.5×10^{-14}.

8. – Dissemination and access to the time scales

The BIPM disseminates in post-real time the reference time scales TAI and UTC through the BIPM *Circular T* [54].

Access to UTC is provided in the form of differences $[UTC - UTC(k)]$ making at the same time the local approximations UTC(k) traceable to UTC; starting in January 2005, their uncertainties are also published [45]. The use of the integer number of seconds of $[TAI - UTC]$ leads to TAI.

The values of the frequency corrections on TAI and their intervals of validity are regularly reported. This information is needed for the laboratories to steer the frequency of their UTC(k) to UTC.

Circular T provides wide access to the best realisation of the second through the estimation of the fractional deviation d of the scale interval of TAI with respect to its theoretical value based on the SI second, calculated as explained above. The values of d for the individual contributions of PFS are also published, giving access to the second as realised by each of the primary standards.

Access to GPS time with an uncertainty of a few nanoseconds and to GLONASS time with an uncertainty of a few tens of nanoseconds is provided via their differences with respect to TAI and UTC.

Within *Circular T*, the time links used for the calculation of one month, with their respective type A and type B uncertainties are detailed, accompanied by information about the technique used in the calibration of the time transfer equipment or link.

The ftp server of the BIPM time section gives access to clock data and time transfer files provided by the participating laboratories, as well as the rates and weights for clocks in TAI in each month of calculation. This information is particularly useful for laboratories in the study of their clocks behaviour.

Results for a complete year are published in the *Annual Report of the BIPM Time Section* [46], together with information about the laboratories' equipment, time signals and time dissemination services, as reported by the laboratories to the BIPM.

Data used for the calculation of TAI, *Circular T*, some tables of the *Annual Report* and all relevant results and information are available on the ftp server of the BIPM time section (`www.bipm.org`).

GPS satellites disseminate a common time scale designated as *GPS time*. It is the system time for GPS. The GPS time was set to UTC on 6 January 1980, and since then it has not been adjusted to UTC by leap seconds. Therefore $[TAI - GPS\ time] = 19\,\mathrm{s} + C$, were C is a small quantity which is at the most $1\,\mu\mathrm{s}$, and in practice, of order of $10\,\mathrm{ns}$. GPS time is the result of clock combination steered to the realisation of UTC(USNO) (modulo 1 second), from which it cannot differ in more than one microsecond, the exact difference is contained in the GPS navigation message. UTC(USNO) represents UTC at the level of a few nanoseconds, and its dissemination via GPS gives the widest access to a real-time approximation of TAI and UTC.

9. – List of acronyms used in the text (in English and French when applicable)

BCRS	Barycentric coordinate reference system
BIH	*Bureau international de l'heure*
BIPM	*Bureau International des Poids et Mesures* International Office for Weights and Measures
C/A	coarse acquisition
CCDS	*Comité Consultatif pour la définition de la seconde* Consultative committee for the definition of the second
CCIR	International Radio Consultative Committee
CCTF	Consultative Committee for Time and Frequency
CGPM	*Conférence Internationale des Poids et Mesures* International Conference for Weights and Measures
CIPM	*Comité International des Poids et Mesures* International Committee for Weights and Measures
EAL	*Echelle atomique libre* Free atomic scale
ET	*Temps des éphémérides* Ephemeris Time
GCRS	Geocentric coordinate reference system
GLONASS	GLObal Navigation Satellite System, Russia
GNSS	Global navigation satellite system
GPS	Global positioning system
IAU	International Astronomical Union
ICRF	International celestial reference frame
IERS	International Earth rotation and reference systems service
IGS	International GNSS Service
IT	acronym used in *BIPM Circular T* for *Istituto Nazionale di Ricerca Metrologica* (INRiM), *Torino*, Italy
ITRF	International terrestrial reference frame
ITU-R	International Telecommunication Union, Radiocommunication Sector
LNE-SYRTE	Laboratoire National d'Essais, Systèmes de reference temps-espace Paris
NICT	National Institute of Information and Communications Technology Tokyo, Japan
NIST	National Institute for Standards and Technology, Boulder, USA France
NPL	National Physical Laboratory, Teddington, United Kingdom
NRC	National Research Council, Canada
ONRJ	Observatorio Nacional Rio de Janeiro, Brazil
OP	Observatoire de Paris

PFS	Primary frequency standard
PTB	Physikalisch-Technische Bundesanstalt, Germany
SI	*Systéme nternational d'unités*
	International system of units
TAI	*Temps atomique international*
	International atomic time
TA(k)	atomic time maintained at laboratory k
TCB	*Temps coordonnée barycentrique*
	Barycentric coordinate time
TCG	*Temps coordonné géocentrique*
	Geocentric coordinate time
TT	*Temps terrestre*
	Terrestrial time
TWSTFT	two-way satellite time and frequency transfer
URSI	*Union Radio-Scientifique Internationale*
	International Union of Radio Science
USNO	United States Naval Observatory, Washington DC, USA
UT	*Temps universel*
	Universal time
UT0	*Temps universel* "0"
	Universal time "0"
UT1	*Temps universel* "1"
	Universal time "1"
UTC	*Temps universel coordonné*
	Coordinated universal time
UTC(k)	coordinated universal time realized at laboratory k
VLBI	Very Long Baseline Interferometry
VLF	very low frequency

REFERENCES

[1] *Proceedings of the International conference for the adoption of a single prime meridian and a universal time* (Gibson Bros., Washington) 1884.

[2] Guinot B., *Time and the Earth's rotation (IAU Symp. 82)* (Reidel) 1979, p. 7.

[3] Guinot B. and Arias E. F., *Metrologia*, **42** (2005) S20.

[4] Leschiutta S., *Metrologia*, **42** (2005) S10.

[5] *Comptes Rendus a la 11 ème Conférence Générale Des Poids et Mesures*, **86** (Bureau International Des Poids et Mesures) 1961.

[6] Ramsey N. F., *J. Res. NBS*, **88** (1983) 301.

[7] Essen L. and Parry J. V. L., *Philos. Trans. R. Soc. A*, **250** (1957) 45.

[8] Markowitz W., Hall R. G., Essen L. and Parry J. J. L., *Phys. Rev. Lett.*, **1** (1958) 105.

[9] *13th Conference Generale Des Poids et Mesures*, *Metrologia*, **4** (BIPM, 1968) 43.

[10] *14th Conference Generale Des Poids et Mesures*, *Metrologia*, **8** (BIPM, 1972) 35.

[11] *15th Conference Generale Des Poids et Mesures*, *Metrologia*, **11** (BIPM, 1975) 180.

[12] *Bureau International Des Poids et Mesures*, *Le Système International d'Unités SI*, 8th edition (2006).

[13] Allan D. W., Hellwig H., Kartaschoff P., Vanier J., Vig J., Winkler G. M. R. and Yannoni N.F., *Proceedings of the 42th Annual Frequency Control Symposium* (1988) p. 419.

[14] Guinot B., *Metrologia*, **34** (1997) 261.

[15] Becker G., *Comité Consultatif pour la Définition de la Seconde*, 4$^{\rm e}$ session (1969) S26.

[16] Guinot B., *Celestial Mechanics*, **38** (1986) 155.

[17] Iau, *Proceedings of the 21st General Assembly* (Kluwer, Dordrecht) 1991 (Resolution A4).

[18] Iau, *Proceedings of the 24th General Assembly* (Astronomical Society of the Pacific, Provo, USA) 2000. (Resolutions B1.3 and B1.9)

[19] Hafele J. C. and Keating R. E., *Science*, **177** (1972) 166.

[20] Klioner S. A., *Cel. Mech. Dynam. Astron.*, **53** (1992) 81.

[21] Borde Ch. J., *Metrologia*, **39** (2002) 435.

[22] Brumberg V. A., *Essential Relativistic Celestial Mechanics* (Adam Hilger) 1991.

[23] Soffel M. H., *Relativity in Astrometry, Celestial Mechanics and Geodesy* (Springer-Verlag) 1989.

[24] Bergeron J. (Editor), *International Astronomical Union, IAU Trans.* Vol. **21** B (Kluwer, Dordrecht, Boston, London) 1991.

[25] Ma C., Arias E. F., Eubanks T. M., Fey A., Gontier A.-M., Jacobs Ch., Sovers O. J., Archinal B. and Charlot P., *Astron. J.*, **116** (1998) 516.

[26] Altamimi Z., Sillard P. and Boucher C., *J. Geophys. Res.*, **107** (2001) B10.

[27] International Earth Rotation and Reference Systems Service, Bulletin B, IERS EOP Product Centre, Observatoire de Paris (monthly).

[28] Guinot B., *Astron. Astrophys.*, **192** (1988) 370.

[29] Petit G., *Proceedings of the 35th PTTI Systems and Applications Meeting* (United State Naval Observatory) 2003, p. 307.

[30] Bureau International de l'Heure, *BIH Annual Report for 1973* (1974) Observatoire de Paris.

[31] Audoin C. and Guinot B., *The Measurement of Time* (Cambridge University Press) 2001.

[32] Guinot B. and Thomas C., Establishment of International Atomic Time, Annual Report of the BIPM Time Section, **1**, D1-D22 (1988).

[33] ISO, *Guide to the expression of uncertainty in measurement* (1995).

[34] Allan D. W. and Weiss A. M., *Proceedings of the 34th Annual Symposium on Frequency Control* (United State Naval Observatory) 1980, pp. 334-346.

[35] Jiang Z. and Petit G., *Time Transfer with GPS satellite All-in-View, Proceedings ATF 2004* (United State Naval Observatory) 2004, p. 236.

[36] Defraigne P. and Petit G., *Metrologia*, **38** (2003) 184.

[37] Wolf P. and Petit G., *Proceedings of the 31st PTTI* (United State Naval Observatory) 1999, pp. 419-428.

[38] Azoubib J. and Lewandowski W., *Metrologia*, **37** (2000) 55.

[39] Lewandowski W., Foks A., Jiang Z., Nawrocki J. and Nogas P., *"Recent progress in GLONASS time transfer"*, *Proceedings of the Joint IEEE FCS and PTTI* (Vancouver, Canada) 2005, pp. 728-734.

[40] Lewandowski W., Azoubib J., Gevorkyan A. G., Bogdanov P. P., Klepczynsky W. J., Miranian M., Danaher J., Koshelyaevsky N. B. and Allan D. W., *Proceedings of the 28th PTTI* (United State Naval Observatory) 1997, pp. 367-386.

[41] Lewandowski W. and Azoubib J., *Proceedings of IEEE/EIA International Frequency Control Symposium* (2000) pp. 586-597.

[42] Azoubib J. and Lewandowski W., *4th BIPM TWSTFT Monthly Report* (BIPM) 1999 pp. 1-11.

[43] Lewandowski W. and Moussay P., *Rapports BIPM-2002/02*, 2002; *BIPM-2003/04*, 2003; *BIPM-2003/05*, 2003.

[44] Lewandowski W., Tisserand L., *Rapports BIPM-2004/05*, 2004; *BIPM-2004/06*, 2004; *BIPM-2004/08*, 2004.

[45] Petit G., Jiang Z. Moussay P., Powars E., Dudle G. and Uhrich P., *Proceedings of the 15th EFTF* (2001) pp. 164-166.

[46] Matsakis D., *Proceedings of the 34nd PTTI* (United State Naval Observatory) 2003, pp. 437-456.

[47] Cordara F., Lorini L., Pettiti V., Tavella P., Piester D., Becker J., Polewka T., de Jong G., Koudelka O., Ressler H., Blanzano B. and Karel C., *Calibration of the IEN-PTB TWSTFT link with a portable reference station, Proceedings of the 18th EFTF, 2004* (Guildford, UK) on CD-ROM..

[48] Lewandowski W., Cordara F., Lorini L., Pettiti V., Bauch A., Piester D. and Koudelka O., *"A Simultaneous calibration of the IEN/PTB time link by GPS and TWSTFT portable equipment", Proceedings of the 18th EFTF, 2004* (Guildford, UK) on CD-ROM.

[49] Piester D., Hlavac R., Achkar J., de Jong G., Blanzano B., Ressler H., Becker J., Merck P. and Koudelka O., *"Calibration of Four European TWSTFT Earth Stations with a Portable Station Through Intelsat 903", Proceedings of the 19th EFTF, 2005*, on CD-ROM.

[50] Lewandowski W. and Azoubib J., *Proceedings of the 34th PTTI* (United State Naval Observatory) 2003, pp. 413-424.

[51] Azoubib J., *Proceedings of the 32th PTTI* (United State Naval Observatory) 2001, pp. 195-209.

[52] Arias E. F. and Petit G., *Proceedings of the Joint IEEE FCS and PTTI* (Vancouver, Canada) 2005, pp. 244-246.

[53] Azoubib J., Granveaud M. and Guinot B., *Metrologia*, **13(4)** (1977) 87.

[54] *Circular T*, a monthly publication of the BIPM.

[55] Gill P. and Riehle F., *"On secondary representations of the second", Proceedings of the 20th EFTF, Braunschweig, Germany* (PTB) 2006, pp. 282-288.

[56] LNE-SYRTE, Report to the CCL-CCTF Joint Working Group on secondary representations of the second (2005).

[57] Zsymaniec K. *et al.*, *Metrologia*, **43** (2006) L18.

[58] Madej A. A., Bernard J. E., Dubé P. and Marmet L., *Phys. Rev. A*, **70** (2004) 012507.

[59] Oskay W. H. *et al.*, *Phys. Rev. Lett.*, **97** (2006) 020801.

[60] Schneider T., Peik E. and Tamm Chr., *Phys. Rev. Lett.*, **94** (2005) 020801.

[61] Takamoto M., Hong F.-L., Higashi R. and Katori H., *Nature*, **435** (2005) 321.

[62] Hong F.-L, Takamoto M, Higashi R., Fukuyamay., Jiang J. and Katori H., *Opt. Express*, **13** (2005) 5253.

[63] Ludlow A. D., Boyd M. M., Zelevinsky T., Foreman S. M., Blatt S., Notcutt M., Ido T. and Ye J., *Phys. Rev. Lett.*, **96** (2006) 033003.

Temperature metrology: Low and ultra-low temperatures

R. Rusby

Division of Engineering and Process Control, National Physical Laboratory
Teddington TW11 0LW, UK

1. – Introduction

Temperature is a key parameter in thermodynamics, chemical kinetics and statistical mechanics. It enters thermodynamics fundamentally as the potential for heat transfer, and as such it is the driver of almost all manufacturing industries, from steel, petrochemicals, plastics, glass, ceramics, etc, to foods and pharmaceuticals. Heat provides the basis for the generation of most electric and automotive power, in which the efficiency, as the 2nd law tells us, increases with the temperature of the hot reservoir. The need to operate with ever greater efficiency in aero engines in particular has led to the development of materials which can withstand highly hostile environments, and it also demands closer measurement and control of the temperature to ensure that safe limits are not exceeded.

Conversely, lowering the temperature reduces physical, chemical and biological activity. Chilling or freezing are key to the preservation, distribution and storage of foodstuffs and biological samples. Liquefaction of gases is important for the transportation of natural gas, in the use of oxygen in hospitals and in steel-making, hydrogen and oxygen for rocket propulsion, and helium for cooling large superconductive magnets. Liquid nitrogen and helium are used in a multitude of large and small activities as convenient portable sources of cooling. Cryogenic engineering is therefore big business.

In research, metrology and technology, cooling is sometimes needed just to reduce thermal noise as an unwanted background to a detection system, but it is often fundamental to the study or application of the phenomenon of interest: *i.e.* to observe or make use of low-energy physical effects which are otherwise disrupted by thermal activity. Thus if the characteristic energy of a physical interaction is less than the thermal energy of the particles of the system, $Ei < kT$, where k is the Boltzmann constant and T is the (thermodynamic) temperature, then the phenomenon will be wiped out. Hence, all ordering phenomena have an associated temperature above which they are not observed.

The temperature spectrum is therefore an energy spectrum, in principle extending indefinitely upwards, and also indefinitely down towards, but never reaching, the absolute zero.

2. – Refrigeration and thermometry

It is almost 100 years since the last of the "permanent" gases (those that could not be liquefied by pressure alone) was finally condensed. In a multi-stage process, culminating in a Joule-Thomson expansion through an orifice, Kamerlingh Onnes collected a small sample of liquid helium in a glass vacuum flask, in Leiden in 1908.

Cooling cycles involving the compression and expansion of gases continue to be mainstay of refrigeration, being capable of large heat extraction rates. Moreover, the cold gas or liquid produced can be conveniently piped or transported from the source to the point of use. Alternatives based on solid (magnetic or electric) refrigeration are confined to relatively small-scale and specialised local applications. Some processes are rather complex, such as those which are routinely used at ultra-low temperatures in "dilution" refrigerators and nuclear cooling, as we will see later. The temperature reached is limited by loss of cooling power, which eventually falls to the point where it is balanced by the heat load placed on the process.

Since temperature must be a parameter of state in any cooling process, the working substance itself has potential for use as a thermometer. Thus the thermometer associated with gas-based (p-v-T) refrigeration is the classical constant-volume gas thermometer (CVGT), evaporation is linked with vapour pressure thermometry, magnetic refrigeration with thermometry using paramagnetic susceptibility or magnetisation, Peltier cooling with thermoelectric thermometry, etc. These thermometers are all, in principle at least, primary thermodynamic thermometers, although in most cases the properties are not calculable accurately enough and some empirical calibration input is needed.

There is another class of thermometer, where the relevant parameters can be measured and related to T through the laws of statistical mechanics. These include radiation thermometry, which is governed by the Planck law, and electronic noise thermometry, which is based on the Nyquist law for Johnson noise in a resistor. In recent years other devices have been developed, such as tunnel junctions, where the I-V characteristics can be used (Coulomb blockade, shot noise), with some possibility for application in "on-chip" refrigeration. Table I provides a summary of thermometers, but further discussion is held back until a later section.

Table I. – *Thermometers and refrigerators.*

Property/system	Thermometer	Key equation	Refrigerator
Gas	CVGT	$pV = nRT[1 + \ldots]$	expansion
Liquid-vapour equilibrium	vapour pressure	$\mathrm{d}p/\mathrm{d}T = (S_g - S_l)/(V_g - V_l)$	evaporation
^{3}He-^{4}He mixtures	osmotic pressure	$\Pi v_4 = xRT$	^{3}He dilution
Paramagnetic susceptibility	magnetic	$\chi = C/(T + \Delta)$	demagnetisation
Thermoelectricity	thermocouple	$E = \int S_{\mathrm{A}} \mathrm{d}T$	Peltier cooling
Electron gas	noise thermometer	$\overline{V_T^2} = 4kTR(T)\Delta f$	-
Tunnel junctions	Shot noise	$S_{\mathrm{I}}(V) = (2eV/R)\coth(eV/2kT)$	?
	Coulomb blockade	$eV_{1/2} = 5.439NkT$	?
Electrical conduction	resistance thermometer	$R = f(T)$	-
Crystal resonance	quartz, NQR	$\nu = f(T)$	-
Thermal radiation	radiation thermometers	$E_\lambda = (2\pi hc^2/\lambda^5)[e^{(hc/k\lambda T)} - 1]^{-1}$	-

3. – Secondary thermometers

While an ideal thermometer would be based on fundamental principles and lead directly to T through a rigorous relationship, such methods are in practice extremely difficult and are rarely applied directly. Instead practical thermometry is based almost exclusively on sensors which have good functional properties, but where the link with theory is at best incomplete: they are "secondary", and they must be calibrated in some way before they can be used.

Almost every material property has some temperature dependence, and therefore almost every material is a potential thermometer. Most have limited or no practical applicability, but a few have the necessary sensitivity, stability and convenience to be produced commercially, and the best are used for defining or maintaining the "practical" scales.

There are several features which thermometers must have to be useful, such as adequate sensitivity, reasonable ease of measurement and calibration, low dissipation and the ability to make good contact with the object under measurement. In varying degrees, depending on the application, they should also have wide range of use, fast response, good stability, compatibility with their environment including EMC, small (sometimes microscopic) size, convenience of installation and replacement, interchangeability of the sensor, insensitivity to magnetic fields, etc., with fully automated and multiplexed instrumentation and communications, and all for a few euros. Not surprisingly, the ideal thermometer has not yet been made.

Solid and liquid expansion has been exploited for years in mechanical thermometers or actuators, but some, *e.g.* quartz crystal oscillators and nuclear quadrupole thermometers, have been engineered as temperature-to-frequency converters. This has obvious instrumental advantages but, for one reason or another, they have not made big commercial inroads. Likewise, many opto-thermal devices (*e.g.*, using fluorescence, reflectometry or refractometry) have been produced, with niche applications but with limited general popularity. The more successful thermometers use electrical sensors and instrumentation, especially the various types of resistance thermometer and thermocouple.

In fact, the pre-eminent thermometer for almost 100 years now has been the platinum resistance thermometer, PRT, both as the Standard Platinum Resistance Thermometer, SPRT, and its ruggedised versions, various forms of Industrial Platinum Resistance Thermometer, IPRT. Any newcomer must offer something which the PRT cannot provide.

The resistivity of platinum has good temperature sensitivity over a wide range, with a broadly linear, though gently curved, characteristic. It has good resistance to chemical attack and can be used, with increasing difficulty, almost to 1000 °C. SPRTs are normally of the "long-stem" type in which the sensing resistor is a coil of fine annealed platinum wire supported on a former in a stain-free manner, connected by long ($\sim 500\,\mathrm{mm}$) leads to a termination at room temperature, the whole being enclosed in a protective tube. At low temperatures the resistor is held inside a small (approximately 5 mm diameter by 50 mm long) helium-filled capsule which is totally included within the cryostat, the connections being made using fine wires. The capsule-type SPRT is illustrated in fig. 1.

The SPRT sensing element traditionally has a resistance, R, of about $25.5\,\Omega$ at 273 K, which gives a sensitivity of $0.1\,\Omega\,\mathrm{K}^{-1}$. This is large enough that modern commercial resistance bridges can be used with a resolution of a few μK, but at low temperatures the resistance and sensitivity become inconveniently small. The range of use is in practice limited to 13.8 K, where $R \sim 0.03\,\Omega$. Below this other thermometers must be used, and the rhodium-iron resistance thermometer has become the preferred option for metrology. This is modelled on the capsule-type SPRT but uses a wire of rhodium alloyed with 0.5% (by atom) of iron. The resistance anomaly due to the dissolved iron is such that good resistance and sensitivity is maintained down to 0.5 K or lower.

While low resistance thermometers make good standards over wide ranges, they are not generally advantageous from the measurement point of view, as the voltage, V, tends to be small for an acceptable self-heating, V^2/R. High-resistance semiconductor sensors (germanium, ruthenium oxide, "Cernox", or just carbon) are often preferred, because of the high sensitivity, greater ease of measurement and also their small size. However, individual sensors do not have such good stability or range of use. Other devices in use are silicon or GaAs diodes: these have very high voltage sensitivity ($> 1\,\mathrm{mV/K}$), and good interchangeability, but a high dissipation and susceptibility to noise. Capacitance sensors, using the permittivity of strontium titanate or aluminium silicate samples, have also been used, particularly where measurements are needed in high magnetic fields.

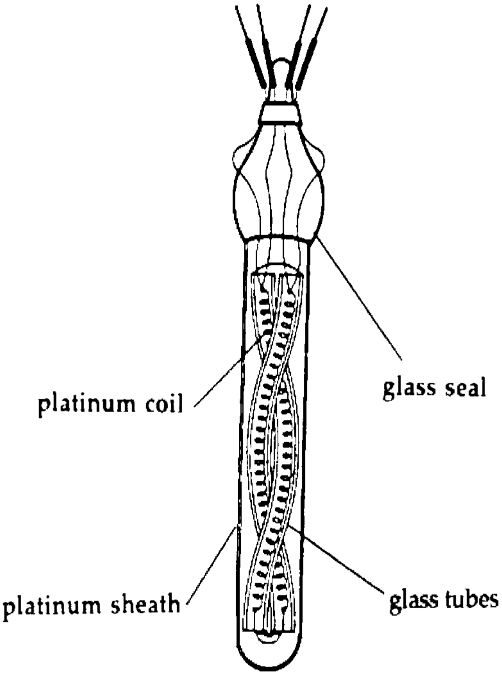


Fig. 1. – Capsule-type Standard Platinum Resistance Thermometer, H.Tinsley Type 5187L.

They are prone to drift and are susceptible to voltage spikes. Figure 2 shows how voltage sensitivities compare for some low-temperature sensors, with the caution that this is just one of the many criteria to address in selecting a sensor.

We now need to see how these thermometers can be used with proper traceability back to thermodynamics and the SI.

4. – The unit of thermodynamic temperature

As with other physical quantities, temperature metrology begins with the definition of a unit. Although historically this was based on the so-called "fundamental interval" of 100 ° between the melting point of ice and the boiling point of water, since 1954 the unit has been the kelvin. This is defined by assigning the value 273.16 K to the triple point of water, the unique temperature at which the liquid, solid and vapour phases of water coexist in equilibrium. The water must of course be pure, but what is pure water? For realisations of the unit at the current state of the art, where uncertainties may be less than 0.03 mK ($\sim 10^{-7}$ of T), the isotopic composition of the water becomes important. To clarify the situation the CIPM affirmed in 2005 [1] that the isotopic compositions of hydrogen and oxygen should be those specified by the International Atomic Energy Agency for "Vienna Standard Mean Ocean Water".

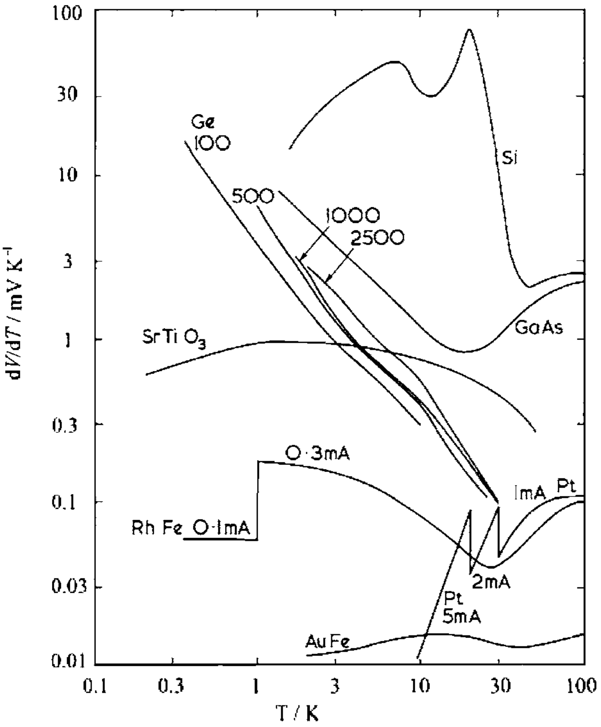


Fig. 2. – Voltage sensitivity, dV/dT for some low-temperature thermometers.

The triple point can be realised in practice very precisely and reproducibly, using clean glass cells containing the water under vacuum, *i.e.* at its vapour pressure, as illustrated in fig. 3. To prepare the cell for use, a mantle of ice is formed around the central thermometer well and, after the cell has rested for some days, a thin layer of the ice is melted to create a liquid film immediately around the well. When inserted into the well, the thermometer will accurately sense the temperature of the liquid-solid interface. The cell is long enough that a long-stem SPRT can reach the required temperature without being significantly influenced by conduction of heat along the stem. It must also be protected from radiative heat transfer.

The depth of water in the cell gives rise to the main correction which must be applied to the temperature, that due to the hydrostatic pressure head. The thermometer senses the temperature in the region near the bottom of the well, and at this depth the pressure at the liquid-solid interface is increased above the vapour pressure alone. The effect is to reduce the temperature by $0.73\,\mathrm{mK\,m^{-1}}$, or about $0.2\,\mathrm{mK}$ in a typical cell. The measured resistance must therefore be increased by the corresponding amount, to obtain $R(273.16\,\mathrm{K})$.

Other influences which may need to be taken into account are corrections for differences from the specified isotopic composition, and for impurity effects, where necessary.

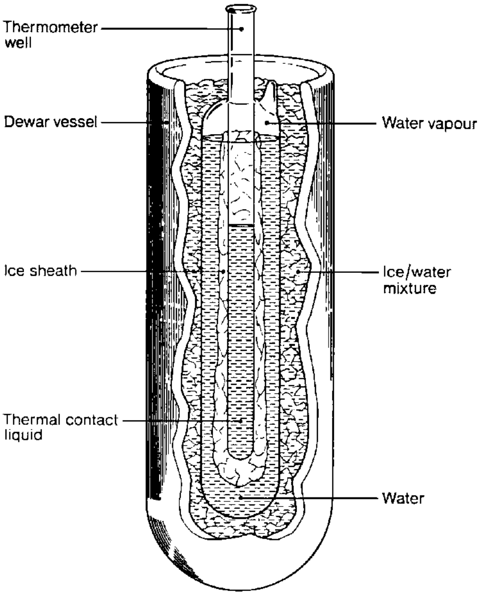


Fig. 3. – A water triple-point cell in an ice bath.

Guidance on the realisation of the triple point of water and other fixed points, and their use in calibrating SPRTs, is provided by the *Comité Consultatif de Thémometrie* in the *Supplementary Information for the ITS-90* [2] and by most national measurement institutes.

5. – The measurement of thermodynamic temperature

Having defined and realised the unit (or, rather, 273.16 units), one would like to multiply and subdivide it to allow standards to be set up for dissemination at any required value, but temperature is not amenable to this process: it is an intensive and indivisible quantity rather than an extensive and divisible one. Its value in other states of matter is not easily related to the triple point of water or any other reference point. To proceed, it is necessary to establish states of equilibrium at each new temperature and operate a thermometer whose parameters of state are well understood theoretically, so that measurements can be used to deduce the thermodynamic temperature according to the definition of the unit. There are few thermometers for which the theory and practice have been developed with the necessary accuracy, and the experiments are all extremely difficult. They divide into three main systems, those based on the properties of gases, photons and electrons, and these are now briefly reviewed.

5·1. *Gas thermometry.* – Gases have been used for thermodynamic thermometry since its inception, and we shall see that they are still pre-eminent for much of the range, though the techniques have changed. In classical gas thermometry, the virial equation of state for a gas is used to relate pressure, p, and molar density, ρ $(= n/V)$, of the gas to T:

$$pV = nRT\bigl[1 + B(T)(n/V) + C(T)(n/V)^2 + \dots\bigr], \tag{1}$$

where R is the molar gas constant, and $B(T)$ and $C(T)$ are the second and third virial coefficients.

In the constant-volume gas thermometer, CVGT, the density is kept approximately constant by maintaining a fixed amount of gas, n, in a rigid bulb of volume, V, and the pressure is measured at a succession of temperatures to generate a scale. To avoid having to determine the density absolutely, the temperatures are measured relative to a reference temperature, T_{r} (directly or indirectly, this must be 273.16 K), so that ratios of pressure can be taken and the density cancels out, to first order of accuracy. The gas constant also need not be known. Equation (1) then becomes

$$p/p_{\mathrm{r}} = (T/T_{\mathrm{r}})\bigl[1+B(T)(n/V)+C(T)(n/V)^2+\dots\bigr]/\bigl[1+B(T_{\mathrm{r}})(n/V)+C(T_{\mathrm{r}})(n/V)^2+\dots\bigr].$$

Several difficulties are apparent:

- Unless the bulb is closed and includes a pressure transducer, it must communicate with the external measurement system along a (capillary) tube. This has additional volume, the "dead-space", and temperature-dependent corrections must be applied for the amount of gas in this space.

- Variable (pressure and temperature dependent) amounts of the gas will be adsorbed on the surface of the bulb and wider gas system, or in cracks and crevices, and so are removed from the sample, leading to measured pressures which are too low.

- The volume will vary with the temperature, and to a lesser extent with the pressure, due to expansion and dilatation of the bulb.

- The pressure (ratio) measurement is required to be at first order in accuracy, and this will introduce many experimental complications.

- The virial coefficients must be known.

Many other practical difficulties will arise, due to establishing stable states of equilibrium, recording the results using (usually) resistance thermometers, relating them to known states (fixed points) or comparing them with other thermodynamic determinations.

In order to cope with the non-idealities of the gas (represented by the virial expansion in eq. (1)) it is in principle necessary to plot an "isotherm": that is, to carry out measurements at a number of densities and extrapolate the measurements to zero density, to determine the limiting value of pV/n, which is RT. Again, it is helpful to plot the isotherms relative to an isotherm at 273.16 K, to avoid the absolute measurement of pressure and density, and then to fit the family of isotherms all together. In this way

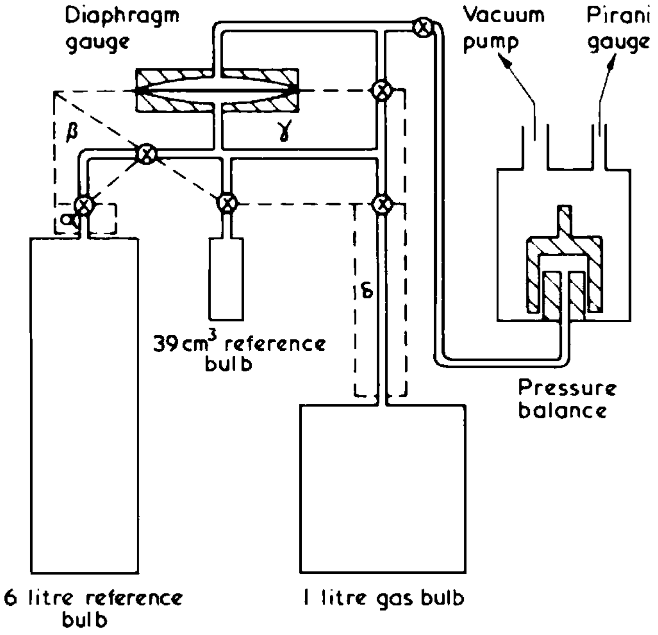


Fig. 4. – The CVGT apparatus of Berry.

the ratios of the temperatures, T/T_r, and the values of the virial coefficients, $B(T)$ and $C(T)$, can be determined.

Figure 4 shows a schematic diagram of the CVGT apparatus of Berry [3], which was used to establish the NPL-75 scale in the range 2.6 K to 27.1 K, relative to the triple point of water. The bulb, made of oxygen-free high-conductivity copper with a gold-plated interior surface, was 1 litre in volume. A stainless-steel reference volume of 6 litres maintained near 273.16 K (actually, in melting ice) was used to admit known amounts of gas into the bulb at low temperatures. The figure indicates four sections of connecting pipe work, whose volumes contributed to the dead-space volume, and a 39 cm^3 volume which was used to confirm the measurements of the dead-spaces by an expansion method. The gas sample is terminated at a diaphragm gauge, which measures the (small) pressure differences between the gas in the bulb and that generated by a pressure balance. The two main innovations in this work were to use a large bulb, so as to reduce the effect of the dead volume, and to use a pressure balance rather than a mercury barometer, for convenience and because it gave a constant relative uncertainty of a few parts in 10^6 in ratios of the pressures generated.

Other experiments in gas thermometry have been undertaken, (*e.g.* [4-7]), but a notable advance in recent years has been the accurate *ab initio* calculation of the second virial coefficient, $B(T)$, for helium [8], which simplifies the procedure for this gas (only). Even so, primary constant-volume gas thermometry is no longer practised, and two alternative gas-based techniques are now preferred.

The first is dielectric-constant gas thermometry, DCGT, in which the need to track amounts of gas and the bulb dimensions is replaced by a direct *in situ* measurement of the gas density. A capacitor is included in the bulb, and the change of capacitance when the gas is admitted is measured. Thus

$$\Delta C/C(0) = (\varepsilon - 1) + \kappa\,\varepsilon\,p\,,$$

where the relative permittivity of the gas $(\varepsilon - 1)$ can be related to the density, and the second term is a correction due to the compressibility of the capacitor. The temperature is then obtained by application of the equation of state, including the virial corrections. This technique is discussed in some detail in the paper by Fischer.

5˙2. *Acoustic gas thermometry.* – The second technique, which avoids many of the difficulties of constant-volume gas thermometry, is acoustic gas thermometry. The speed of sound, u, in a gas is given by

$$u^2 = \gamma RT/M\big[1 + A_1(T)p + A_2(T)p^2 + \dots\big], \tag{2}$$

where γ is the (known) ratio of specific heats, M is the molar mass of the gas, and the non-idealities are now represented by $A_1(T)$ and $A_2(T)$, the second and third acoustic virial coefficients, which are accurately related to $B(T)$ and $C(T)$. Thus if u is measured at a number of pressures, an isotherm can be plotted whose intercept leads to a value of RT.

The utility of measuring the speed of sound in gases as a primary means of measuring temperature has long been appreciated. The parameters are all intensive, so in principle there is no need to measure quantities such as volumes and the amounts of gas contained in them. Moreover, pressure is not required to be measured to first order of accuracy, being only needed for evaluating the non-idealities.

The main challenge of the method is the measurement of u with the necessary uncertainty, and in spite of significant advances this is still the limitation. Firstly, velocity decomposes into two extensive quantities, length and time. Time (or frequency) is in principle not difficult, but the length has always been problematical. This can be easily understood in the simple time-of-flight method in which a pulse is launched into the gas at one point and detected at another. Better results can be achieved interferometrically, for example in a cylindrical cavity in which the distance traversed by a piston between successive resonances is measured, each separation corresponding to half of the acoustic wavelength.

Fixed-frequency variable-path interferometry was undertaken in the 1960s and 1970s at NBS (NIST) and NPL for temperature measurement in the range 2 K to 20 K [9, 10], to provide a much-needed primary scale in the liquid-helium-to-liquid-hydrogen temperature range. While these experiments achieved significant advances and reached uncertainties in the region of 1 in 10^4, some second-order acoustic effects had to be overcome, by design or by correction.

One particular effect is that at high frequencies (*e.g.* $\sim 1\,\text{MHz}$, as used at NBS), which can readily be excited by quartz crystal oscillators, many different modes of propagation are possible and are not easily resolved. The velocity measured is likely to be a group velocity constituted of the phase velocities of the different modes. It is very difficult, if not impossible, to analyse the situation well enough to make a valid correction.

Consequently a low-frequency (few kHz) instrument was preferred at NPL, in which the higher modes were excluded and the propagation was truly plane wave. It had the compensating disadvantage of larger boundary layer losses due to viscous and thermal effects at the cavity walls, but in this case the effects could be modelled and corrections calculated [11].

Although the NBS and NPL acoustic experiments were successful up to a point, parallel advances in constant-volume gas thermometry [3], relative to the triple point of water, pushed the target uncertainty to a few parts in 10^5, and the acoustic scales were superseded. Later, Moldover and others [12-16], have used spherical acoustic resonators, which can have well-defined acoustic properties and very high Q-values. The resonators are made by carefully machining two hemispheres to the required tolerances and fitting them together at the equator.

In this situation one must vary the frequency, rather than the path, and in practice it is swept through a series of resonances corresponding to the different modes and harmonics propagated. Spherical resonators have been rigorously modelled and applied to the measurement of temperature and the determination of R. Here one of the main difficulties is the need to know the cavity radius, so as to convert the measured frequency to the required velocity.

These methods are discussed in detail by Gavioso in this volume.

5·3. *Other thermodynamic techniques.* – The use of thermal (blackbody, Planckian) radiation in temperature measurement is largely restricted to high temperatures, and here we only note in passing the absolute radiometric measurements of Quinn and Martin [17], based on the Stefan-Boltzmann T^4 law for total radiant exitance, which extended down to 144 K. New developments in this and other radiometric methods are discussed in the paper by Fox.

This leaves electronic noise thermometry as the final thermodynamic technique of importance to be considered.

Johnson noise arises from the instantaneous fluctuations in the local density of electrons within a resistor, which produces a fluctuating voltage across the resistor according to the Nyquist formula [18]

$$\overline{V_T^2} = 4kTR(T)\Delta f,$$

where $\overline{V_T^2}$ is the mean-square Johnson noise voltage measured in a bandwidth Δf across the resistor $R(T)$ at a temperature T.

Noise thermometry is attractive in that it gives access to thermodynamic temperature through the measurement of an electrical noise signal which is independent of all the

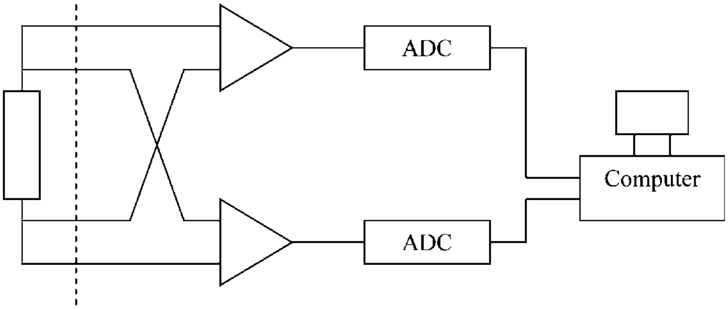


Fig. 5. – Schematic drawing of a cross-correlation noise thermometer, originated by Brixy [20].

material properties of the sensor except its resistance. It can in principle be used to measure T over a very wide range and, if the difficulties can be overcome, it would be a truly practical thermodynamic thermometer.

In addition, the availability of reference voltages based on the Josephson effect ($V_\mathrm{J} = nhf/2e$) and reference resistances based on the quantised Hall effect ($R_\mathrm{H} = h/ie^2$) leads to the possibility of measuring the Johnson noise from a resistor in terms of the fundamental constants h and e and thus providing a route to the determination of T (or k) through the direct comparison of the two energy units hf and kT.

The main difficulty is that the signal levels generated in a typical noise thermometer are small. For example, the noise generated by a $100\,\Omega$ sensor at a temperature of 300 K observed in a bandwidth of 100 kHz is about $0.4\,\mu$V rms. Effective measurement of such signal levels requires low-noise amplifiers with gains typically in the region of 10^6. However, amplifiers with carefully designed input stages based on field effect transistors have been developed with input noise contributions which are negligible for temperature accuracies at the mK level [19].

The mean-square voltage, since it is a noise signal, fluctuates about its expected value and the resulting uncertainty σ in a single measurement is given by

$$\frac{\sigma^2}{\overline{V_T^2}^2} \geq \frac{1}{\tau_\mathrm{m} \Delta f}\,,$$

where τ_m is the measurement time. To achieve an uncertainty of 3 mK at 300 K with a 100 kHz bandwidth requires a measurement time of at least 27 hours. (In contrast, an industrial measurement with an uncertainty of 1 K at 1000 K would only require 10 seconds.) In order to shorten this measurement time, either a higher measurement bandwidth is required, if the parallel capacitance permits, or a number of measurement systems could be set up, whose results are combined.

The noise generated by the wires which connect the sensor to the amplifiers has to be eliminated from the measurement. This is conveniently done by making a four-wire connection to the sensor and using two amplifier chains as shown in fig. 5. If the product of the output signals from the amplifiers is taken, then only noise signals shared by both channels give a non-zero expectation in the output. Noise voltages in the leads and voltage noise in the amplifiers are eliminated. Current noise from the amplifiers is not eliminated and has to be made sufficiently small by design. Recent developments in analogue-to-digital converters and computer processing mean that the multiplication is most effectively carried out by digitising the signal and calculating the product in a computer. Note that the four-wire sensor configuration also gives an accurate definition of the resistance value of the sensor.

The frequency response of the signal path has a direct effect on the accuracy of the measurement and this effect is most easily reduced by switching between two sensors (at, for example, two temperatures whose ratio is to be measured) at the reference plane indicated by the dotted line in fig. 5. The noise powers need to be matched: thus the resistances need to be such that the product $TR(T)$ is the same for both. With such a technique, it is important that the sensors and their connecting wires also have the same frequency characteristic.

To date a number of noise thermometers have been set up, notably as described in [20] and relative uncertainties of $< 10^{-5}$ seem to be in prospect. These will provide invaluable confirmation (or otherwise) of gas-based results using an entirely different system.

Carrying out ratio measurements with two sensors, with one at the triple point of water, establishes temperatures relative to the present definition. To make a noise measurement traceable back to the maintained quantum electrical units, and so allow direct independent measurement of T, a known electrical signal has to be substituted at the reference plane. This has been done very effectively at NIST using an array of Josephson junctions biased with a high-speed pulse train having a pseudo-random pattern chosen to emulate the noise produced by a resistive sensor [21]. The Josephson-based noise source was substituted for the resistive sensor at 100 second intervals and the authors reported a type-A uncertainty of 10 mK on a measurement at the gallium triple point (~ 30 parts in 10^6 at 302 K) using data collected over 16 hours and a 100 kHz signal bandwidth.

However, the measured temperature differed by 44 mK from the ITS-90 value. A critical factor with this method is the exactness of the substitution of the noise source for the sensor, and the measurement data showed some non-ideal frequency response behaviour. Further work to understand this is in progress.

6. – International Temperature Scale of 1990, ITS-90

Because the measurement of temperature from first principles is so difficult and time-consuming, in practice temperature metrology proceeds in two stages. In the first, the thermodynamic data that have been produced are compared and assessed, with the aim of achieving a consensus of values of T over the whole range of interest. Then an "International Temperature Scale" is constructed, to serve as the basis for everyday measurement standards.

The purpose of the ITS is to set out the measurements which are needed to calibrate certain specified practical thermometers in such a way that temperatures obtained us-

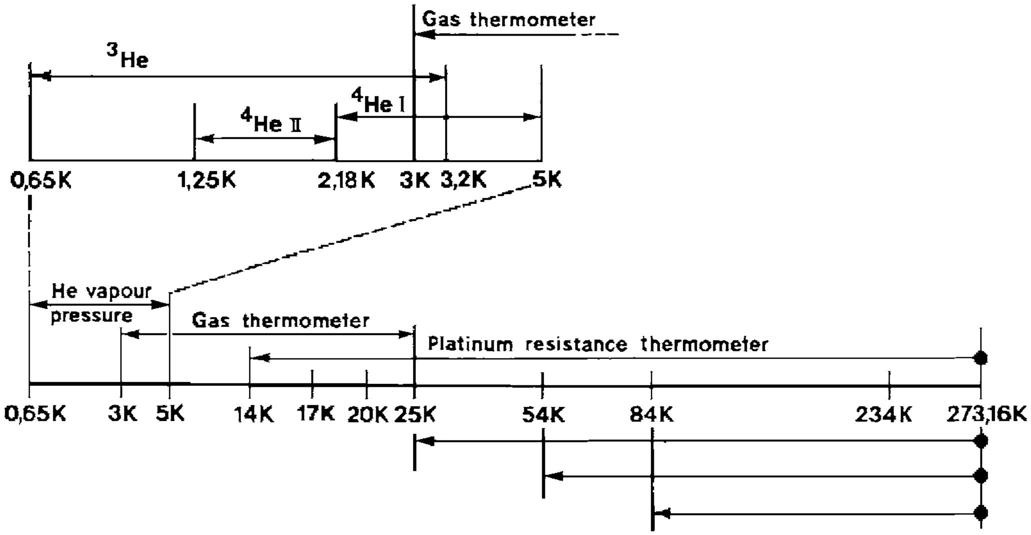


Fig. 6. – The sub-ranges of the ITS-90 below 273.16 K [2].

ing them are precise and reproducible, while at the same time closely approximate the corresponding thermodynamic values.

The key points are that the practical thermometers are more precise and reproducible than a thermodynamic thermometer can hope to be, but that they are also accurate enough to be used as a convenient substitute for thermodynamic measurements. The practical thermometers which have been the mainstay of temperature metrology since the first ITS in 1927, the standard platinum resistance thermometers (SPRTs) in their various forms, are well adapted as transfer standards for the next tier of dissemination, the calibration services run by national measurement institutes and accredited laboratories.

In its text the ITS-90 [22] specifies that the SPRT resistance should be measured at the triple point of water and at a series of other fixed points, all being highly reproducible states of matter such as the triple points of gases and freezing points of pure metals, depending on the range. For convenience and flexibility, a number of sub-ranges are allowed, so that one need only go as far up or down the scale as is desired. The sub-ranges below 273.16 K are schematically shown in fig. 6, taken from the *Supplementary Information for the ITS-90* [2], and shows the special arrangements for temperatures below the triple point of neon, 24.5561 K, discussed later.

Long-stem SPRTs are commonly used from the triple point of argon (83.8058 K) to the freezing point of zinc (692.677 K) or aluminium (933.473 K), and helium-filled capsule-type SPRTs are typically used from the melting point of gallium (320.9146 K) to the triple point of hydrogen (13.8033 K).

The apparatus and techniques for realising the fixed points have been described in the literature and in the *Supplementary Information*, to which reference should be made.

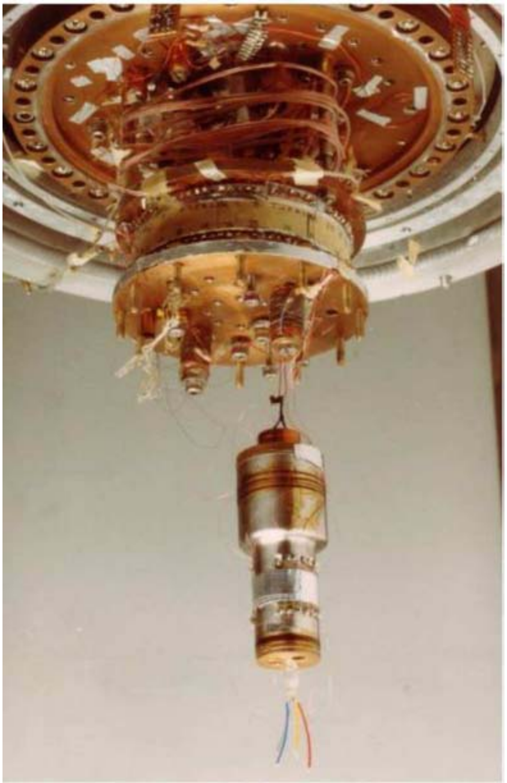


Fig. 7. – NPL stainless-steel triple-point cell suspended in the cryostat.

Here, in fig. 7, we show a cryogenic triple-point cell as used at NPL for hydrogen, neon, oxygen and argon, and some other gases. The cell is made of clean stainless steel, with a copper block welded into the base for receiving three cSPRTs and bringing them into good contact with the interior of the cell. The gas is mainly contained in the top chamber (the "expansion space") but on cooling in an adiabatic calorimeter it condenses and freezes in the narrow gap around the copper block.

When the calorimeter has settled just below the triple-point temperature with minimal heat flow to or from the cell, a series of heat pulses are applied, and measurements of the thermometer resistance are made at several distinct points on the melting plateau. From these the resistance value at the liquidus point (melted faction = 100%) is deduced, see fig. 8, [23]. The general principles are described in the *Supplementary Information*, and detailed considerations concerning isotopic and other effects, with reference to hydrogen, are discussed by Fellmuth *et al.* [24].

The ITS-90 specifies formulae for interpolating the SPRT calibration throughout the range. The guiding principle is that the resistance ratios,

$$W(T_{90}) = R(T_{90})/R(273.16\,\mathrm{K}),$$

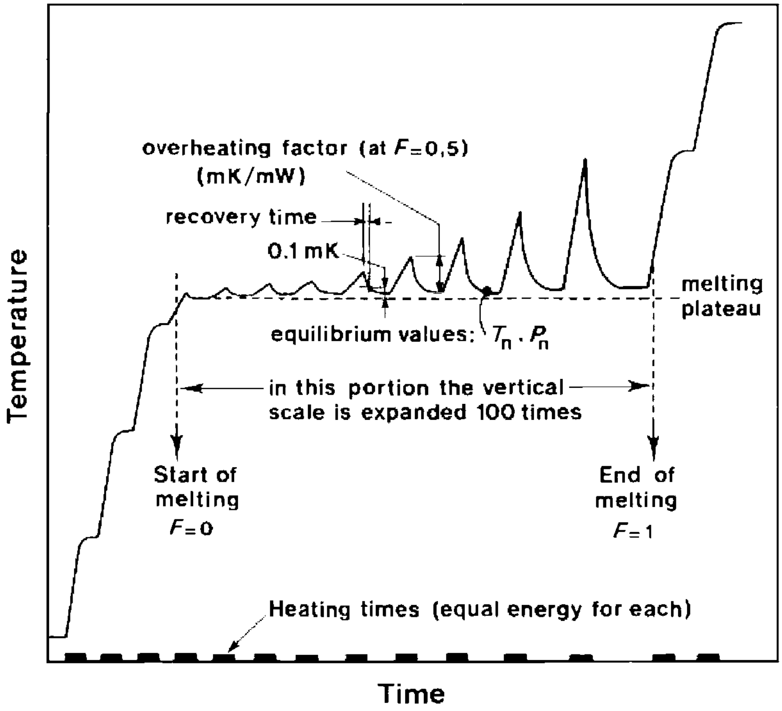


Fig. 8. – Melting plateau for a cryogenic triple point, from Pavese [23].

are very similar for all high-quality SPRTs. Therefore a reference function specifies $W_r(T_{90})$ as a polynomial function of T_{90}, and the calibration need then only determine the small deviations from the reference function: $\Delta W = W(T_{90}) - W_r(T_{90})$. The use of resistance ratios also avoids the need for absolute standards of resistance.

Figure 9 shows on the left how the resistance ratio (or resistivity) depends on temperature, with the loss of sensitivity at low temperatures. On the right, the deviations between thermometers are seen to be typically less than 0.0003 in W, which is less than 0.1 K in temperature equivalence. These deviations can be fitted with relatively few parameters, and correspondingly few fixed points are needed to calibrate SPRTs over wide ranges (see fig. 6): deviation functions with up to 7 terms can achieve satisfactory ($\sim$ 0.5 mK) interpolation of ΔW, compared with the 12th-degree equation which is needed to specify $W_r(T_{90})$.

The variability between SPRTs is due mainly to the effects of impurities and crystal defects in the platinum, which introduce unwanted resistance through temperature-independent scattering of electrons. (However, the decrease in some deviations at low temperatures, seen in fig. 9, is likely to be due to residual temperature-dependent scattering from magnetic impurities such as iron and manganese.) Because of this variability, different SPRTs interpolate slightly differently, which is the so-called "non-uniqueness"

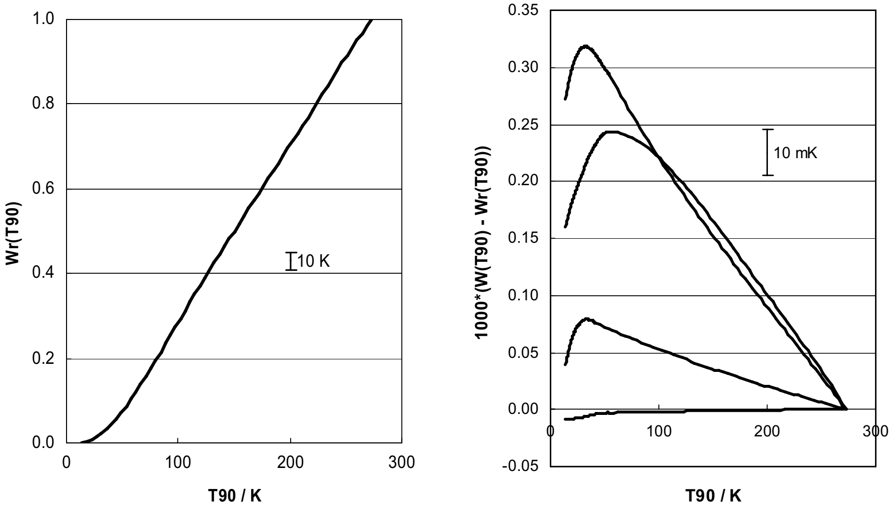


Fig. 9. – Resistance ratio $W_\mathrm{r}(T_{90})$ *vs.* T_{90}, left, and deviations for four SPRTs, right. Note the 3000-fold difference in the scale of the ordinates.

between SPRT calibrations, and to keep this within reasonable limits, the ITS-90 specifies criteria for the minimum acceptable temperature coefficient of resistance, which means that well-annealed wires of high purity must be used.

6·1. *The ITS-90 below the triple point of neon, 24.5561 K.* – As we have seen, the SPRT is not suitable for defining the ITS-90 below 13.8 K. Among other resistance thermometers, the rhodium-iron resistance thermometer is sufficiently sensitive and very stable, and hence suitable for maintaining a calibration, but the characteristics of the alloy are more variable and not so amenable to interpolation using a reference function [25].

To bridge the gap between the vapour-pressure range of liquid hydrogen ($T > 13.8\,\mathrm{K}$) and of liquid helium ($T < 5\,\mathrm{K}$), the ITS-90 resorts to a quasi-thermodynamic instrument, the interpolating constant-volume gas thermometer, iCVGT. The range of use is specified to be 3 K to 24.5561 K, overlapping the SPRT and helium vapour-pressure ranges, see fig. 6. Many of the difficulties of thermodynamic determinations are removed by interpolating between the iCVGT pressures at triple point of neon, the triple point of hydrogen and a helium vapour-pressure point in the range between 3 K and 5 K. A simple quadratic equation is adopted for this. In spite of the simplifications, the instrument is still slow and difficult, and not frequently used, and the arrangement is only acceptable because the calibrations of rhodium-iron thermometers, which are the result of the experiment, can be maintained more-or-less indefinitely, with only occasional checks at the fixed points. A realisation of the ITS-90 in this range has been described by Meyer and Reilly [26].

For temperatures below 5 K, the ITS-90 specifies the measurement of the vapour pressures of ^{4}He (1.25 K to 5 K) and ^{3}He (0.65 K to 3.2 K), and gives equations for the

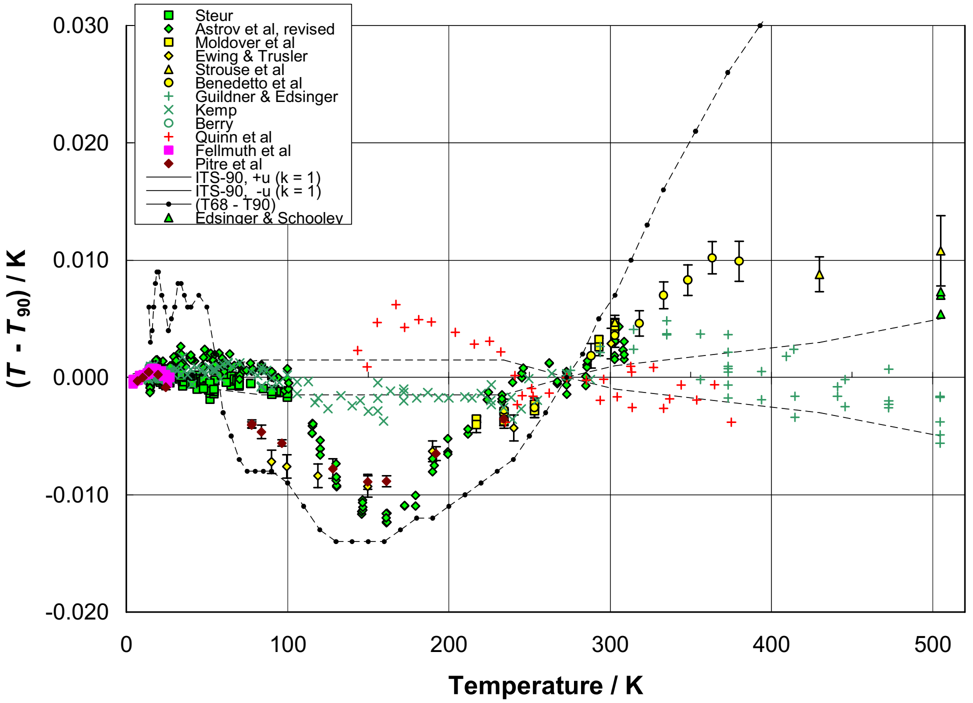


Fig. 10. – Differences between published determinations of thermodynamic temperature and the ITS-90, from various authors. Note the uncertainty bars ($k = 1$) on the recent acoustic thermometry data.

dependence of the pressure on T_{90}. A realisation therefore requires that a small volume of pure liquid (^{4}He or ^{3}He) is condensed in a copper block, with an access tube for filling and measurement of the pressure of the liquid-vapour equilibrium. Several descriptions of apparatus have been published, such as Rusby and Swenson [27], Meyer and Reilly [28] (who used the same apparatus as for their gas thermometry).

6˙2. *Thermodynamic accuracy of the ITS-90.* – To complete this brief discussion of the ITS-90 at low-temperatures, it remains to ask how closely T_{90} is now thought to represent T. Figure 10, produced by CCT Working Group 4 in 2005 [29], is a collation of data which shows the differences between published determinations of thermodynamic temperature and the ITS-90. A notable feature is the good consistency between the various recent experiments using acoustic thermometry mentioned earlier. These suggest that significant errors exist in the ITS-90, with a difference in slope of about 10^{-4} (1 mK in 10 K) over the range from 150 K and 350 K. Also notable is the erratic behaviour of the differences with the former scale, the IPTS-68, when compared with the ITS-90, especially at low temperatures.

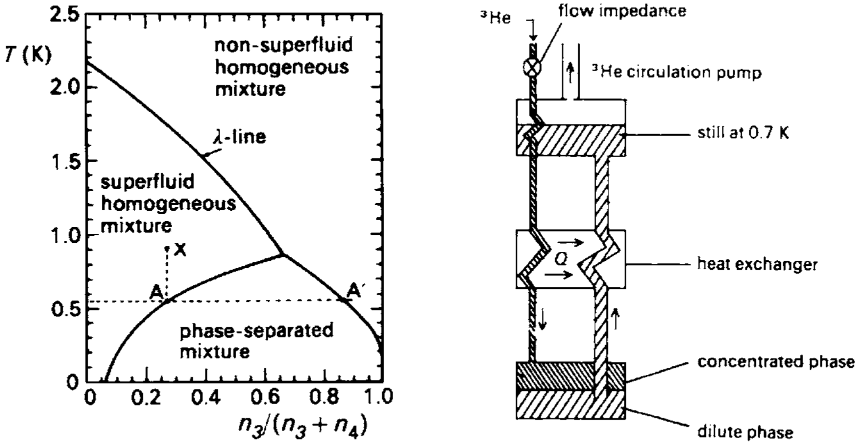


Fig. 11. – Left: Phase diagram for ^{3}He-^{4}He mixtures, temperature *vs.* ^{3}He concentration, and Right: schematic diagram of a dilution unit. Reprinted by permission from McClintock, Meredith and Wigmore [30].

7. – Ultra-low temperatures

We now descend into the mysterious world of ultra-low temperatures (ULT), which is conveniently interpreted as "sub-kelvin". We are interested in the range of the dilution refrigerator (down to about 5 mK) to satisfy commercial needs, but we also want to include the "^{3}He features"; that is, the superfluid transitions at 2.444 mK and 1.896 mK, and the Néel transition in the solid phase, at 0.902 mK, because they are used as reference points in sub-millikelvin work.

As before, we will begin by seeing how one can reach these temperatures, and then look at some of the thermometers which are available to measure them. The dilution refrigerator, the workhorse of ULT research, is a truly remarkable machine. It uses the heat of dilution, or mixing, of ^{3}He in ^{4}He which, as we will see, is analogous to evaporative cooling. The dilution takes place in the mixing chamber, where liquid ^{3}He is encouraged to diffuse into liquid ^{4}He from which it can be led off, distilled and recirculated, thereby achieving a continuous process. To see how it works, we must look at the phase diagram for mixtures of ^{3}He and ^{4}He.

In fig. 11, left, we see first the "lambda-line", which marks the boundary between superfluid and normal mixtures. In pure ^{4}He it is at 2.1768 K, but as ^{3}He is added the transition takes place at lower and lower temperatures, until at about 0.85 K the lambda-line splits into two branches. These mark the boundaries of the phase-separation region in which the mixture spontaneously separates into the normal ^{3}He-rich "concentrated" phase, and the superfluid ^{4}He-rich "dilute" phase.

Thus if one cools a mixture of, say, 30% ^{3}He, superfluidity is encountered at about 1.7 K, and phase separation at about 0.6 K. A concentrated phase appears at point A' and, being lighter, this sits on top of the dilute phase (at point A), the position of the interface being such as to accommodate the total mixture. On further cooling, the concentrated phase becomes more concentrated, and the dilute phase more dilute. Eventually the concentrated phase becomes almost 100% ^{3}He, but the dilute phase can always hold at least 6% ^{3}He. This is crucial, because ^{3}He atoms can easily migrate through the superfluid dilute phase, which behaves as a quasi-vacuum.

Referring to fig. 11, right, in the dilution cycle ^{3}He crosses the boundary between the two phases, in the mixing chamber, absorbing the heat of dilution in a kind of upside-down evaporation. It then diffuses to the still at about 0.7 K, where it is preferentially evaporated because its vapour pressure is much higher than that of ^{4}He. It is then fed back in and recondensed through a flow impedance in the concentrated side. Thus there are several extraordinary features, all of which are crucial to the process:

- the existence of the phase separation;
- the existence of a substantial heat of dilution;
- the superfluidity of the dilute phase;
- the significant solubility of ^{3}He in ^{4}He, even at $T = 0$;
- the factor of ~ 30 between the vapour pressures of ^{3}He and ^{4}He at ~ 0.7 K, so that the distillate is almost all ^{3}He.

From the point of view of building a dilution refrigerator, the mixing chamber and still present few problems: they are just chambers with inlets and outlets, as necessary. The key to a successful design is the parts in between: the heat exchangers. The cooling power, $\dot{Q}$, though substantial, falls off with the square of the temperature:

$$\dot{Q}/W = 84\dot{n}T^2,$$

where $\dot{n}$ is the flow rate typically $< 10^{-4}\,\mathrm{mol\,s^{-1}}$, and T is the mixing chamber temperature. The base temperature of the refrigerator is reached when the dilution process can absorb only the heat load on the mixing chamber due to the incoming ^{3}He plus any other heat leaks. The returning ^{3}He must therefore be efficiently cooled by the flow of ^{3}He in the dilute phase, and so it is the design of the heat exchangers which is the limiting factor. Continuous counterflow heat exchangers can initially be used but, to overcome the Kapitza thermal boundary resistance below 0.1 K, which increases with T^{-3}, discrete heat exchangers are needed with blocks of sintered silver powder to provide large surface areas in both the dilute and concentrated phases. Base temperatures are generally about 5 mK, though 2 mK has been achieved [31].

7·1. *Nuclear cooling.* – Dilution refrigerators are available commercially and, while they are complex and require skill to run, in modern versions they are at least semi-automated. However, they do not reach the ^{3}He feature temperatures, and for sub-mK experiments they are just the platform for further cooling. This is done using demagnetisation of a nuclear paramagnet, such as copper, as illustrated in the S-T diagram in

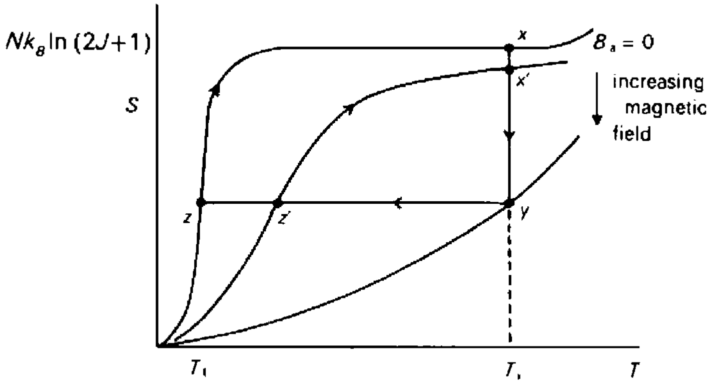


Fig. 12. – Entropy-temperature diagram for a system of magnetic dipoles at three fields, illustrating the nuclear cooling process, reprinted by permission from McClintock, Meredith and Wigmore [30].

fig. 12. The nuclear moments, of spin J, are magnetised by a powerful magnet, the heat of magnetisation being extracted by the mixing chamber of the dilution refrigerator at about 10 mK. The sample moves from Point x to Point y on the diagram. With the sample magnetised at the maximum available field, a heat switch is opened to isolate it. The field is then ramped slowly down, adiabatically (*i.e.* at constant entropy), and the sample therefore cools from Point y to Point z. Very large temperature changes can be achieved, from 10 mK to 1 μK or less for copper – but for the nuclear spins only: the temperature of the lattice and electrons lags behind, being cooled only through the weak spin-lattice coupling. Cooling an experiment, for example a cell of ^{3}He, in addition requires measures to increase the surface area and so overcome the Kapitza boundary resistance.

We shall not discuss nuclear cooling further, but only mention the use of PrNi$_5$, in which the internal field at the nucleus is enhanced by a factor of about 12, and good alignment of the spins occurs in a modest (6 T) applied field at $\sim$ 10 mK. Experiments can then be cooled to $\sim$ 0.3 mK, and this can be the first stage of a two-stage nuclear cooling process, the second stage, with copper, reaching $< 30\,\mu$K.

Several books are available which discuss ULT techniques and thermometry (*e.g.* [30, 32-34]).

8. – ULT Thermometry

Having achieved the cooling, how can one measure the temperature reached? As we have seen, a wide variety of techniques and devices have been used to measure low temperatures, and ultra-low temperatures are no different. Figure 13 summarises the more important thermometers over five decades of T, down to 1 mK, though we will not be able to refer to all of them in detail. They range from those which function independently as primary thermometers (*e.g.*, nuclear orientation and noise), which are needed for generating scales, to those which are "semi-primary" (NMR and magnetic

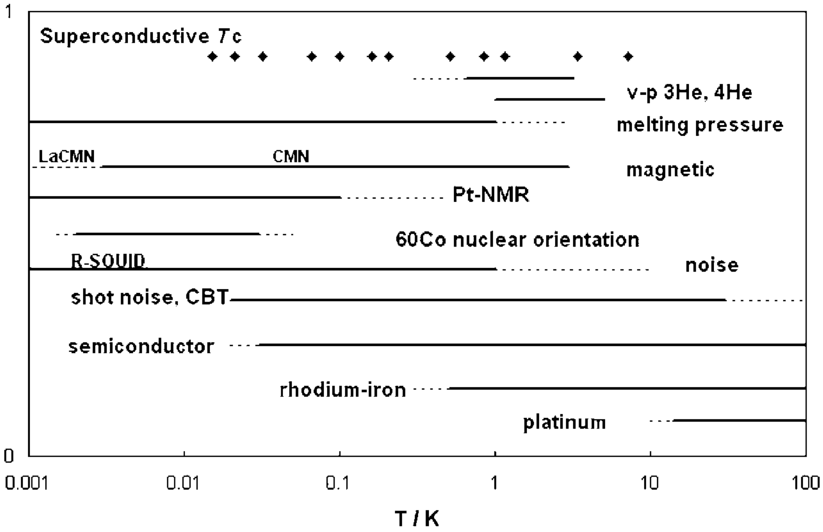


Fig. 13. – Ranges of use of low-temperature sensors and thermometers.

thermometers), requiring only limited calibration input, and those which are secondary practical everyday devices such as resistance thermometers.

It is apparent from fig. 13 that the practical thermometers do not extend all the way to 1 mK. The main reason for this is, again, the increase in the Kapitza boundary resistance as the temperature falls, which means that dissipations of even nW become unacceptable. Metallic resistance thermometers are not practical much below 1 K, because of the low resistances and sensitivities, but semiconductive thermometers are widely used to about 0.03 K. Germanium, ruthenium oxide, zirconium oxynitride (Cernox), and carbon-based resistors give good sensitivity until the dissipation leads to excessive overheating. Then the semiconductor resistances "saturate" at a temperature below which the heat dissipated cannot be conducted away.

8˙1. *Superconductive reference points.* – Figure 13 includes superconductive reference points which are very useful for the *in situ* calibration of secondary devices and magnetic thermometers, and for comparisons with other thermometry (LT and ULT practitioners are wise not to rely on a single thermometer). The transition to superconductivity, at temperature T_c, is detected by observing the exclusion of magnetic flux (the Meissner effect) which changes the mutual inductance of a pair of coils. Formerly two devices, each containing five samples and ranging in total from tungsten ($T_c \sim 15\,\text{mK}$) to lead ($T_c \sim 7.2\,\text{K}$), complete with detection coils, were produced, calibrated and sold by NBS (NIST) as SRM 767 and 768 [35]. Following an EU-funded collaboration [36] in ULT Dissemination, a new device, the SRD1000 is now available from HDL Hightech Developments, Leiden, [37] with ten samples in the range 15 mK to 1.2 K. The package includes detection coils, screens against external fields, electronics, and a calibration by

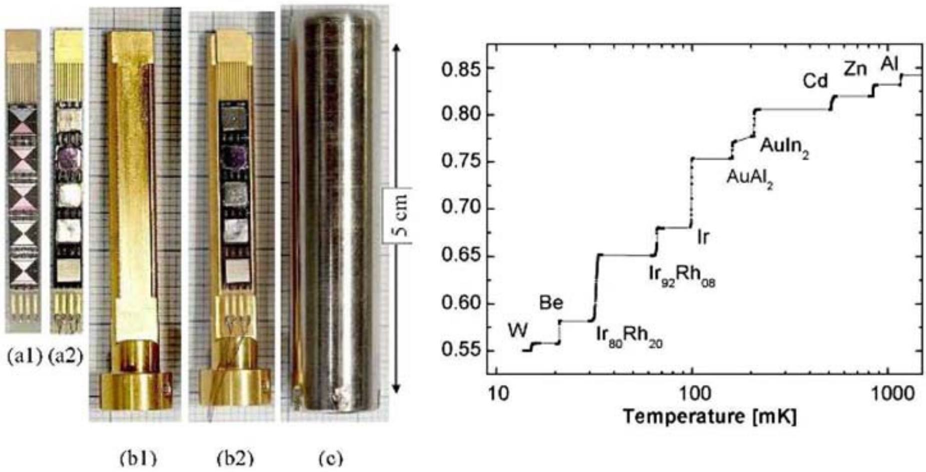


Fig. 14. – SRD1000 assembly, and device output in V as a function of temperature [37].

PTB. The calibration is needed because the transitions of individual samples vary by significant amounts. Figure 14 shows the output of the SRD1000, with the characteristic staircase of ten transitions.

8˙2. *Magnetic thermometry.* – Cerium magnesium nitrate (CMN) magnetic-susceptibility thermometers are often used because they can cover wide ranges with only two or three calibration points, and in their most refined form they have been used to generate laboratory scales such as the NIST CTS-2 [38] and the PTB-96 [39], see sect. **9**. The preferred method is to measure the susceptibility of a single-crystal sphere oriented with the c-axis perpendicular to the measuring field. The governing equation is the Curie-Weiss law:

$$\chi = C/(T - \Delta),$$

where Δ, the Weiss constant, is about 0.27 mK. However, poor thermal diffusion in the crystal limits the range to perhaps 0.05 K, below which powdered samples must be used [40]. Deviations from the Curie-Weiss law occur at a few mK, as the Curie temperature is approached, due to interactions between the cerium ions. To operate to lower temperatures it is necessary to reduce the interactions by substituting most of the cerium with non-magnetic lanthanum, though obviously this reduces the sensitivity. However, $(\mathrm{d}\chi/\mathrm{d}T)$ varies as T^{-2}, so this is acceptable at low enough temperatures.

CMN is the classic paramagnetic salt, which was also widely used for cooling by adiabatic demagnetisation before dilution refrigerators became available. To avoid some of the problems of thermal contact, a metallic paramagnet has recently been used for thermometry [41]. This uses palladium with a trace of dissolved iron, and good results have been obtained down to ~ 1 mK.

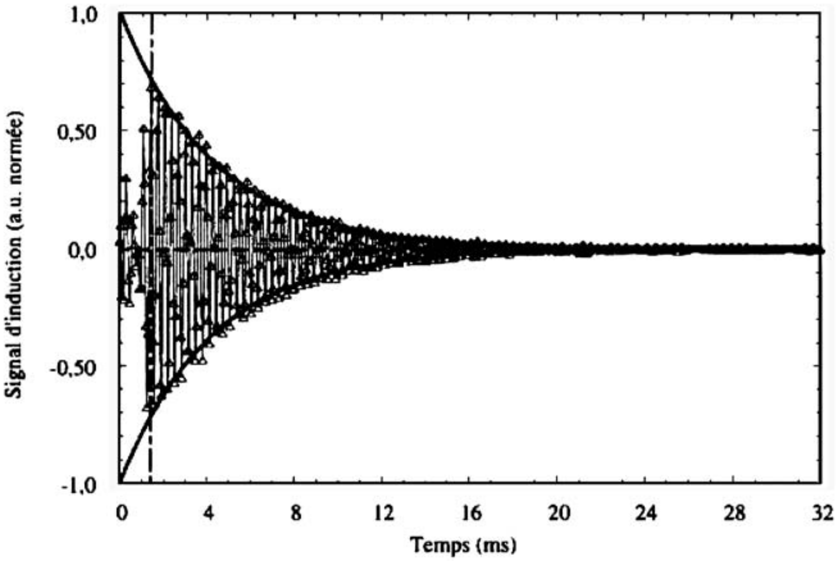


Fig. 15. – Free induction decay for pulsed Pt-NMR (H. Godfrin, private communication).

We now consider two other forms of magnetic thermometer which are extremely useful at millikelvin and sub-mK temperatures. The first is the nuclear orientation (γ-ray anisotropy) thermometer, see [32-34], in which the emission of γ-rays following a β-decay from radioactive nuclei in a metallic lattice is a measure of the polarisation of the nuclear moments in a high magnetic field. At "high" temperatures (say, $> 100\,\text{mK}$) the distribution is isotropic, but as kT falls below the nuclear hyperfine splitting energy, the nuclei are increasingly polarised and the emission of the γ-rays is increasingly anisotropic. The sensitivity falls off again as the polarisation saturates, but for ^{60}Co in a ^{59}Co lattice, good results can be obtained in the range from 3 to 30 mK. In this case, and for ^{54}Mn in Ni, the polarisation is produced by the internal magnetic field of the host lattice.

The temperature is calculated from the known hyperfine splittings and the count rates relative to the isotropic high-temperature limit. The method calls for a small low-activity radioactive sample mounted in contact with the experiment, with counting equipment outside the cryostat. It is therefore practically non-invasive, though the energy absorbed by the sample following the initial β-decay does cause a modest self-heating.

Nuclear orientation is useful as a convenient means of checking other thermometry in its range of use, and also for qualifying the performance of dilution refrigerators against specifications, because it is independent of any other thermometer or calibration. However, the counting is quite time-consuming (~ 1 hr for a measurement to $< 1\%$ of T) and the method does not extend low enough for sub-mK research into ^{3}He and other systems. For this Pt-NMR is the preferred thermometer.

Nuclear magnetisation or susceptibility is most commonly measured by resonance methods, and we will now briefly look into NMR thermometry using a platinum sample. The susceptibility of the nuclear spins follows the Curie law down to the lowest achievable temperatures (microkelvins), and if the spin-lattice relaxation time is short, the nuclei reach equilibrium with the electrons quickly. The sample, which is usually platinum, is in the form of a brush of fine wires because the skin effect would prevent penetration of the oscillating field into bulk material.

The sample, which is located in a steady uniform applied field, is subjected to an orthogonal RF pulse which tips the nuclear spins out of alignment. The magnetisation then precesses around the applied field at the Larmor frequency, and the precession can be detected in a pick-up coil (actually the same as the tipping coil) while the spins decay exponentially back into alignment. This "free induction decay" signal, see fig. 15, occurs with the characteristic spin-spin relaxation time which is proportional to the magnetisation and hence $1/T$. The thermometer can be calibrated at a single point, usually around 10 mK, and extrapolated down to the nuclear ordering temperature ($< 1\,\mu$K). The main concern is that magnetic impurities may cause a temperature dependence in the relaxation time. See refs. [32-34] and [42] for further details.

8˙3. *Noise thermometry with SQUIDS.* – Noise thermometry was mentioned earlier as a potentially interesting primary method at higher temperatures and, given that the noise power is always small and reduces with T, it may be surprising that it could even be considered at ultra-low temperatures. Two factors make it not only possible, but attractive. The first is that whereas uncertainties of $< 10^{-5}$ would normally be needed, at ULT an absolute measurement precision of 10^{-2} is already good, and 10^{-3} is broadly state of the art. This considerably relaxes the statistical demands on noise thermometry, where the measurement time scales inversely with the square of the uncertainty. Secondly, superconductive quantum interference devices (SQUIDs) are available, which are extremely sensitive detectors of voltage, as in a resistive R-SQUID, or of magnetic flux, as in an inductively-coupled current-sensing noise thermometer.

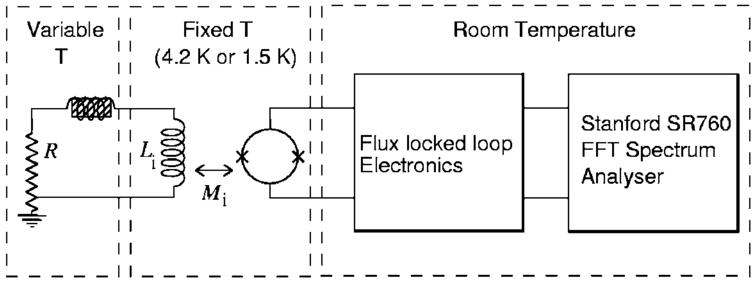

Variable T
Fixed T (4.2 K or 1.5 K)
Room Temperature
R
L_i
M_i
Flux locked loop Electronics
Stanford SR760 FFT Spectrum Analyser

Fig. 16. – Schematic diagram of the current-sensing noise thermometer. One end of the resistor is electrically grounded to ensure adequate cooling of electrons. A fixed-point device is incorporated into the SQUID input circuit.

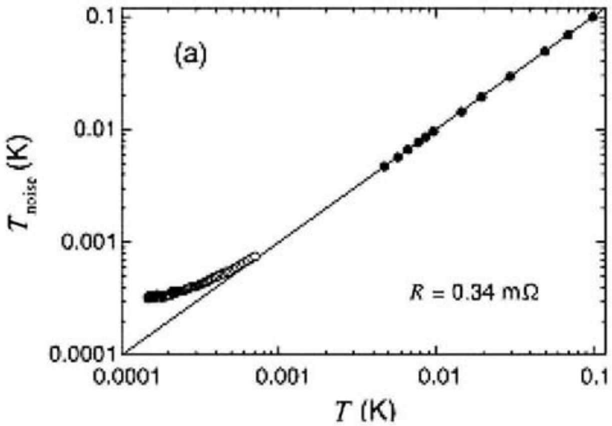


Fig. 17. – Measurements with the current-sensing noise thermometer at RHUL, compared with the PLTS-2000 and a Pt-NMR thermometer below 1 mK [43].

R-SQUIDs were used by both NIST and PTB in the development of their ULT scales [38,39]. In short, the noise resistor carrying a small current is connected in parallel with the Josephson junction, which converts the bias voltage to a frequency. This voltage fluctuates because of the Johnson noise in the resistor. The Josephson frequency is detected by coupling with an RF tank circuit which is amplified and demodulated at room temperature. The frequency is then repeatedly counted and subjected to statistical analysis. Count times were typically several hours, to achieve uncertainties of 10^{-3} or better.

More recently a DC SQUID has been used in a current-sensing noise thermometer at the Royal Holloway University of London (RHUL) [43]. A schematic diagram of the experimental set-up is shown in fig. 16. The resistor is connected to the input coil of a DC SQUID using a shielded superconducting twisted pair. The SQUID is held at fixed temperature (*e.g.* 4.2 K) and the resistive sensor is connected to the experiment whose temperature is to be measured. The mean square noise current flowing in the SQUID input coil per unit bandwidth, due to the thermal noise in the resistor, is given by

$$\langle I_{\mathrm{N}}^2 \rangle = \frac{4kT}{R}\left(\frac{1}{1+\omega^2\tau^2}\right),$$

where $\omega = 2\pi f$, and the resistance R can be taken to be independent of frequency f for the relevant geometries and frequencies. The time constant $\tau = L_{\mathrm{T}}/R$, where L_{T} is the total inductance of the input circuit, which includes in this case a superconductive sample to provide a single *in situ* calibration point (alternatively the device could operate absolutely if the gain of the amplifiers is measured). The output of the flux-locked loop electronics is fed to a spectrum analyser, and the data are then fitted to the above equation, after subtracting a background white-noise power, in order to extract the temperature.

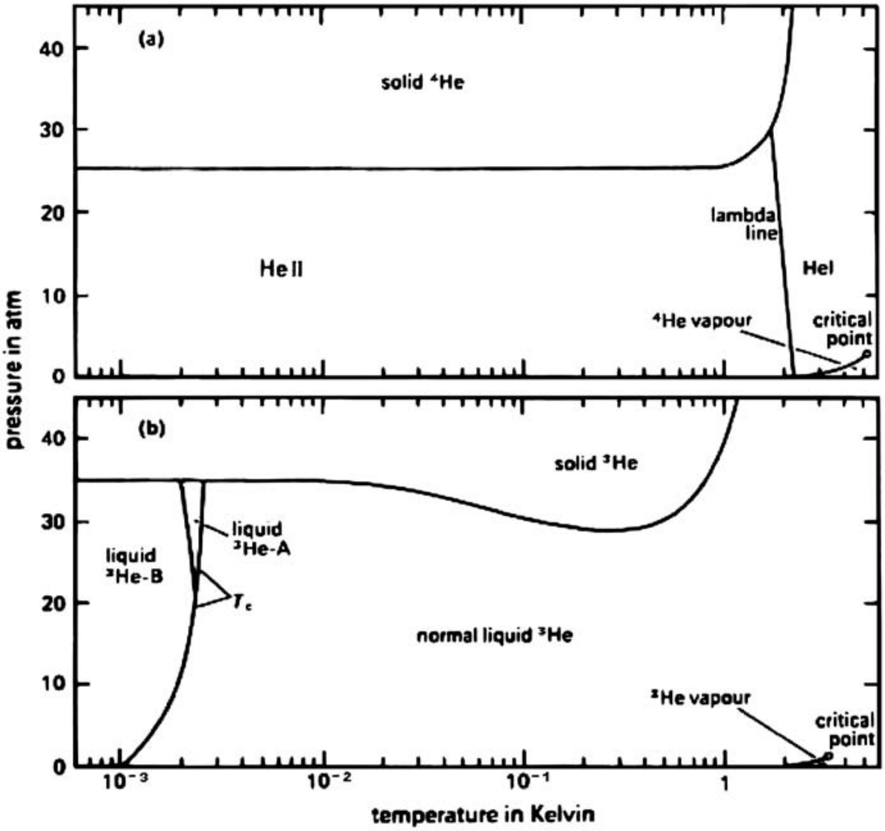


Fig. 18. – Phase diagrams for ^{4}He and ^{3}He under zero magnetic field, reprinted by permission from McClintock, Meredith and Wigmore [30].

Commercially available DC SQUIDs have an energy sensitivity of around 500 h, compared with around 10^5 h for the RF SQUIDs, and correspondingly faster measurements are possible. However, there is a trade-off between the speed and the amplifier noise temperature, and hence the resistance must be chosen to suit the desired operating temperature range. The RHUL thermometer has been tested from 4.2 K to below 1 mK, see fig. 17, but there is evidence of heat leaks at the lowest temperatures.

9. – ^{3}He melting pressure thermometry and the PLTS-2000

The last thermometer to be considered is based on equilibrium between solid and liquid ^{3}He, *i.e.* on the relation between the melting pressure, p_3, of ^{3}He and temperature. The isotopes of helium are unique in that the interatomic forces are so weak compared with the zero-point motion that solidification does not occur at the vapour pressure: there is no conventional triple point, and it is necessary to apply pressures in excess of 2.5 MPa to produce solid. The phase diagrams of fig. 18 illustrate this, and also show that in the case of ^{3}He there is a significant minimum in the melting curve at $T_{\min} = 315\,\mathrm{mK}$. According to the Clausius-Clapeyron equation, this means that the solid has greater entropy than the liquid, and in fact this is due to the nuclear magnetic entropy of $k \ln 2$.

Although it does not look like it in this representation, either side of the minimum the sensitivity $\mathrm{d}p_3/\mathrm{d}T$ is easily good enough for useful thermometry and therefore, given an equation relating p_3 with T, the melting-pressure is a good practical system for thermometry. This was first proposed by Scribner and Adams [44], and apart from the narrow region around the minimum, $T_{\min}$, the thermometer can cover more than three decades of temperature. As ^{3}He is a thermodynamic property of a well-defined substance, it was chosen to be the basis for the Provisional Low Temperature Scale, PLTS-2000 [45] between 0.9 mK and 1 K. We conclude with a short discussion of this development.

During the 1980s and 1990s there was a considerable effort to establish temperature scales for the range below 1 K, notably at NIST and PTB. The NIST work led to a Cryogenic Temperature Scale, CTS-2 [38], which was based on CMN linked to the ITS-90 around 1.2 K and a resistive-SQUID noise thermometer for temperatures down to 7 mK, the lower limit of the dilution refrigerator used. Additional data were obtained from a nuclear orientation thermometer. One of the main objectives was to carry out a definitive determination of the ^{3}He melting pressure relation, using a cell provided by Greywall [46]. The work resulted in a polynomial representation of the melting pressure $p_3(T)$ from 6.3 mK to 700 mK [47].

Meanwhile, Ni and co-workers at the University of Florida, had used Pt-NMR and nuclear orientation thermometry to link the NIST work with the ^{3}He features, and had published (p_3, T) values for them [48], with a polynomial equation from the Néel point to 250 mK.

In a parallel project undertaken at PTB (Berlin), a nuclear cooling stage was included and measurements were made over the whole range down to the ^{3}He features. The PTB-96 scale [39] was based on CMN and Pt-NMR, with a resistive-SQUID noise thermometer to provide a thermodynamic foundation. As at NIST, a melting-pressure cell was included and the scale is recorded as a polynomial equation for $p_3(T)$, in this case all the way from the Néel point to 1 K.

These projects were major advances over what had been done previously. The history of melting-pressure thermometry, and the values of the feature temperatures in particular, was reviewed by Soulen and Fogle [49], showing that discrepancies between various values for the A-transition, for example, had been tens of %. Unfortunately, the UF/NIST and PTB-96 values still differed by about 6% from each other, and no new data seemed likely to appear which might resolve the discrepancy.

Therefore, in 2000, a temperature scale was derived, using thermodynamic calculations to ensure a smooth behaviour as far as possible. This was accepted by the CIPM as the PLTS-2000, between 0.9 mK and 1 K [45], and recommended for use until such time as new information becomes available. The standard uncertainty is estimated to be no more than ±0.5 mK down to 0.5 K, ±0.2 mK at 0.1 K, ±0.3 mK at 25 mK, and about ±0.02 mK (2%) at 0.9 mK. It has been implemented in many ULT laboratories, but it remains the situation today that it needs verification at its lowest extremity.

Table II. – *Values of* p_m/MPa *and* T_{2000}/K *for the* ^{3}He *melting-pressure features, with estimated standard uncertainties with respect to thermodynamic temperature,* $\Delta T/\mu$K*, and the standard uncertainties of the current best realizations,* $\delta T_\mathrm{r}/\mu$K.

Point	p_m/MPa	T_{2000}/mK	$\Delta T/\mu$K	$\delta T_\mathrm{r}/\mu$K
Minimum	2.93113	315.24	360	10
A	3.43407	2.444	48	0.7
A-B	3.43609	1.896	38	2.8
Néel	3.43934	0.902	18	1.1

9·1. *Realisation of the PLTS-2000.* – Fundamentally the operation of a melting pressure thermometer (MPT) requires that a sample of pure liquid and solid ^{3}He be brought to equilibrium at the desired temperature, and the equilibrium pressure be measured absolutely. Several designs of MPT have been described, see for example [39, 46].

A consequence of the inversion in pressure is that for temperatures below T_min the cell cannot communicate with an external measuring system: a line drawn at constant pressure in fig. 18 passes through the solid domain, so the sensing capillary is blocked with a plug of solid ^{3}He, and direct contact with the pressure in the cell is lost. A melting-pressure cell must therefore include a pressure transducer, and this leads to the particular usefulness of the intrinsic fixed points discussed below.

In the region around the minimum the thermometer cannot be used directly for lack of sensitivity, and some interpolation is required. The method for doing this is not specified in the PLTS-2000, but it can use any secondary sensor (CMN, resistive, etc) which can be calibrated in this range. Although it may be surprising to base a scale on a thermometer where there is a region of low or zero sensitivity, the minimum has the compensating advantage of providing an in-built pressure fixed point which can be used for the calibration of the pressure transducer, as will be seen. Schuster *et al.* [50] show that by stepping carefully through the minimum, the temperature can also be located, within about 10 μK.

The three other features may be detected and used as fixed points of pressure and temperature for the *in situ* calibration of the pressure transducer. They are the transition to the superfluid "A" phase, the "A to B" transition in the superfluid and the Néel transition in the solid. The pressure and temperature values of the four fixed points on the PLTS-2000 are shown in table II, with the estimated thermodynamic uncertainty (at $k = 1$) in the temperatures, $\Delta T/\mu$K, and the uncertainty of the best practical realization of each point, $\delta T_\mathrm{r}/\mu$K. For the three low-temperature features, δT_r comes from the pressure resolution with which they can be observed (about ±3 Pa for p_A and $p_\mathrm{Néel}$, ±10 Pa for $p_\text{A-B}$). For the minimum, the pressure resolution is also about ±3 Pa; δT_r comes from locating the point of zero derivative in [50]. The uncertainty in the assigned absolute pressure values was estimated in [45] to be ±60 Pa.

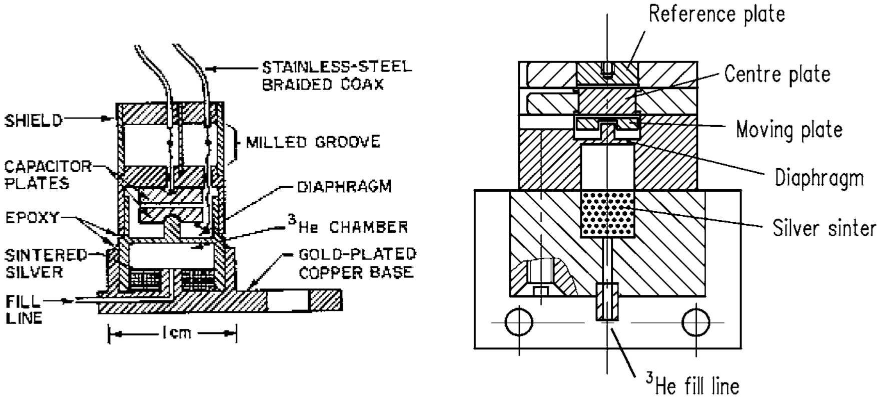


Fig. 19. – ^{3}He melting-pressure cells of Greywall and Busch [46] and Schuster *et al.* [50].

In practical realisations the pressure transducer relies on the capacitive sensing of the displacement of a diaphragm in the cell. The interior, which is typically only a few $\times 100\,\text{mm}^3$ in volume, contains a sinter, usually of silver powder, to promote thermal contact with the liquid ^{3}He and reduce the time constant for equilibrium. Two examples are shown in fig. 19, in which a parallel-plate capacitor senses the displacement of the diaphragm. The most critical design parameters of the transducer are the diameter and thickness of the diaphragm, which is usually made of coin silver or BeCu, and the parallelism of the capacitance plates in order to achieve the desired sensitivity, linearity and reproducibility of the device.

The design, construction, and operation of melting pressure thermometers is reviewed in [50-52], and *Supplementary Information for the realisation of the PLTS-2000* is in preparation by CCT WG4.

10. – Future prospects

There have been major advances in refrigeration and thermometry at low temperatures in recent decades, but there is still some way to go before low and ultra-low temperatures are available in hands-off turn-key systems. Refrigeration is primarily based on the use of fluids, and the resulting complicated vacuum pipework inherently carries risks of leaks —it has been said that the probability goes with the square of the number of joints [53]. In thermometry there are errors and inconveniences in the ITS-90 and PLTS-2000: the latter especially needs confirmation before it can be fully integrated into international thermometry and its provisional status removed.

In the last few years mechanical coolers —"cryocoolers"— have become widely used in place of liquid refrigerants, and magnetic refrigerators have been developed for fast turn-round cooling to $\sim 30\,\text{mK}$. These developments will continue and bring with them greater possibilities for automatic table-top operation, but they will also require small robust electrical temperature sensors, with established traceability to the ITS-90 or PLTS-2000.

Meanwhile, solid-state mesoscopic electronic systems have been produced both for studying quantum phenomena and for incorporating refrigeration and thermometry "on-chip". The possibilities have been reviewed by Pekola *et al.* [54] and by Giazotto *et al.* [55], in which devices based on the width of the Fermi function in tunnel-junctions, or electronic Brownian or other gated refrigerators, are considered to have potential.

To date Coulomb blockade thermometers have reached the market-place [56], and a shot-noise thermometer has been described [57]. They have been investigated mostly below 4.2 K, where the modest relative uncertainties achieved so far (~ 0.1 to 1% of T) are nevertheless interesting, but they can in principle be extended to much higher temperatures. Not relying on superconductivity, they could even be stretched to room temperature, though the greater demands and increased competition will be formidable obstacles to overcome.

* * *

The author is grateful to past and present colleagues at NPL, and partners in the EU project "ULT Dissemination", in the CCT and in the wider scientific community, for many fruitful collaborations over the years.

REFERENCES

[1] CIPM Recommendation 2, CI-2005, BIPM, 2005, at `www.bipm.org`. See also Document CCT/05-07, and the *International System of Units*, 8th edition (BIPM) 2006.

[2] *Supplementary Information for the ITS-90*, 1990, publication of the CCT (BIPM) 1990, available at `www.bipm.org`.

[3] Berry K. H., *Metrologia*, **15** (1979) 89.

[4] Steur P. P. M. and Durieux M., *Metrologia*, **23** (1986) 1.

[5] Kemp R. C., Kemp W. R. G. and Besley L. M., *Metrologia*, **23** (1986) 61.

[6] Astrov D. N., Belyansky L. B., Dedikov Y. A., Polunin S. P. and Zakharov A. A., *Metrologia*, **26** (1989) 151, and Astrov D. N., Belyansky L. B. and Dedikov Y. A., *Metrologia*, **32** (1995/96) 393.

[7] Edsinger R. E. and Schooley J. F., *Metrologia*, **26** (1989) 95.

[8] Hurly J. J. and Moldover M. R., *J. Res. Natl. Inst. Stand. Technol.*, **105** (2000) 667.

[9] Plumb H. H. and Cataland G. A., *Metrologia*, **2** (1966) 129.

[10] Colclough A. R., *Proc. R. Soc. London, Ser. A*, **365** (1979) 349.

[11] Colclough A. R., *Metrologia*, **9** (1973) 75.

[12] Moldover M. R., Trusler J. P. M., Edwards T. J., Mehl J. B. and Davis R. S., *J. Res. NBS*, **93** (1988) 85.

[13] Strouse G. F., Defibaugh D. R., Moldover M. R. and Ripple D. C., *Temperature, its Measurement and Control in Science and Industry*, edited by Ripple D. C., Vol. **7** (American Institute of Physics, New York) 2003, pp. 31-36.

[14] Ewing M. B. and Trusler J. P. M., *J. Chem. Thermodynamics*, **32** (2000) 1229.

[15] Benedetto G., Gavioso R., Spagnolo R., Marcarino P. and Merlone A., *Metrologia*, **41** (2004) 74.

[16] Pitre L., Moldover M. R. and Tew W. L., *Metrologia*, **43** (2006) 142.

[17] Quinn T. J. and Martin J. E., *Philos. Trans. R. Soc. London, Ser. A*, **316** (1985) 85.

[18] Nyquist H., *Phys. Rev.*, **32** (1928) 110.

[19] White D. R., Galleano R., Actis A., Brixy H., de Groot M., Dubbeldam J., Reesink A. L., Edler F., Sakurai H., Shepard R. L. and Gallop J. C., *Metrologia*, **33** (1996) 325.

[20] Brixy H., Hecker R., Oehman J., Rittinghaus K. F., Setaiwan W. and Zimmermann E., *Temperature, its Measurement and Control in Science and Industry*, edited by Schooley J. F., Vol. **6** (American Institute of Physics, New York) 1992, pp. 993-996.

[21] Nam S. W., Benz S. P., Dresselhaus P. D., Burroughs C. J., Tew W. L., White D. R. and Martinis J. M., *IEEE Trans. Instrum. Meas.*, **54** (2005) 653.

[22] Preston-Thomas H., *Metrologia*, **27** (1990) 3-10 and 107, available as a publication of the CCT through `www.bipm.org`.

[23] Pavese F., *BIPM Monographie*, 84/4, 1984.

[24] Fellmuth B., Wolber L., Hermier Y., Pavese F., Peroni I., Szmyrka-Grzebyk A., Lipinski L., Tew W. L., Nakano T., Sakurai H., Tamura O., Head D., Hill K. D. and Steele A. G., *Metrologia*, **42** (2005) 171.

[25] Rusby R. L., *Temperature, its Measurement and Control in Science and Industry*, edited by Schooley J. F., Vol. **5** (American Institute of Physics, New York) 1982, pp. 829-833.

[26] Meyer C. W. and Reilly M. L., *Proceedings of Tempmeko '96*, edited by Marcarino P. (Levrotto and Bella, Torino) 1997, pp. 39-44.

[27] Rusby R. L. and Swenson C. A., *Metrologia*, **16** (1980) 73.

[28] Meyer C. and Reilly M., *Metrologia*, **33** (1996) 383.

[29] Rusby R. L., Moldover M. R., Fischer J. F., White D. R., Steur P. P. M., Hudson R. P., Durieux M. and Hill K. D., *Document CCT/05-19*, BIPM.

[30] McClintock P. V. E., Meredith D. J. and Wigmore J. K., *Matter at Low Temperatures* (Blackie & Son Ltd, Glasgow) 1984.

[31] Vermeulen G. A. and Frossati G., *Cryogenics*, **27** (1987) 139.

[32] Pobell F., *Matter and Methods at Low Temperatures*, 2nd edition (Springer-Verlag, Berlin and Heidelberg) 1996.

[33] Betts D. S., *Refrigeration and Thermometry below One Kelvin* (Sussex University Press) 1976.

[34] Lounasmaa O. V., *Experimental Principles and Methods below 1 K* (Academic Press, London and New York) 1974.

[35] Schooley J. F. and Soulen R. J. jr., *Temperature, its Measurement and Control in Science and Industry*, edited by Schooley J. F., Vol. **5** (American Institute of Physics, New York) 1982, pp. 251-260.

[36] Rusby R. L., Head D., Cousins D., Godfrin H., Bunkov YU. M., Rapp R. E., Gay F., Meschke M., Lusher C., Li J., Casey A., Shvarts Dm., Cowan B., Saunders J., Mikheev V., Pekola J., Gloos K., Hernandez P., Triqueneaux S., de Groot M., Peruzzi A., Jochemsen R., Chincure A., van Heumen E., de Groot G. E., Bosch W., Mathu F., Flokstra J., Veldhuis D., Hermier Y., Pitre L., Vergé A., Fellmuth B. and Engert J., *Temperature, its Measurement and Control in Science and Industry*, edited by Ripple D. C., Vol. **7** (American Institute of Physics, New York) 2003, pp. 89-94.

[37] Bosch W. A., Engert J., van der Hark J. J. M., Liu X. Z. and Jochemsen R., in *Proceedings of the 24th International Conference on Low Temperature Physics, Orlando, 2005*, *AIP Conf. Proc.*, **850** (2005) 1589, See also `www.xs4all.nl/~hdleiden/srd1000`.

[38] Fogle W. E., Soulen R. J. jr. and Colwell J. H., *Temperature, its Measurement and Control in Science and Industry*, edited by Schooley J. F., Vol. **6** (American Institute of Physics, New York) 1992, pp. 91-96.

[39] Schuster G., Hoffmann A. and Hechtfischer D., PTB-Th-2, 2005. See also *Czech. J. Phys.*, **46** (1996) 481, and Document CCT/96-25, BIPM.

[40] Greywall D. S. and Busch P. A., *Rev. Sci. Instrum.*, **60** (1989) 471.

[41] Gloos K., Smeibidl P., Kennedy C., Singsaas A., Sekowski P., Mueller R. M. and Pobell F., *J Low Temp. Phys.*, **73** (1988) 101.

[42] Hechtfischer D. and Schuster G., PTB-Th-1, 2004.

[43] Casey A., Cowan B. P., Dyball H., Li J., Lusher C. P., Maidanov V., Nyéke J., Saunders J. and Shvarts Dm., *Physica B*, **329-333** (2003) 1556.

[44] Scribner R. A. and Adams E. D., *Rev. Sci. Instrum.*, **41** (1970) 287.

[45] Rusby R. L., Durieux M., Reesink A. L., Hudson R. P., Schuster G., Kühne M., Fogle W. E., Soulen R. J. and Adams E. D., *J. Low Temp. Phys.*, **126** (2002) 633.

[46] Greywall D. S. and Busch P. A., *J. Low Temp. Phys.*, **46** (1982) 451.

[47] Fogle W. E. and Soulen R. J. jr., *Toward and International Temperature Scale from* 0.65 *K to* 1 *mK*, workshop held at Leiden University, edited by Rusby and Mohandas (National Physical Laboratory) 1998, pp. 13-19.

[48] Ni W., Xia J. S., Adams E. D., Haskins P. S. and McKisson J. E., *J. Low Temp. Phys.*, **99** (1995) 167; **101** (1995) 305.

[49] Soulen R. J. and Fogle W. E., *Physics Today*, August issue (1997), 36.

[50] Schuster G., Hoffmann A. and Hechtfischer D., PTB-ThEx-21, 2001.

[51] Colwell J. H., Fogle W. E. and Soulen R. J. jr., *Temperature, its Measurement and Control in Science and Industry*, edited by Schooley J. F., Vol. **6** (American Institute of Physics, New York) 1992, pp. 101-106.

[52] Adams E. D., *Progress in Low Temperature Physics*, edited by Halperin W., Vol. **15** (Elsevier B. V.) 2005, pp. 423-457, and *Temperature, its Measurement and Control in Science and Industry*, edited by Ripple D. C., Vol. **7** (American Institute of Physics, New York) 2003, pp. 107-112.

[53] Rose-Innes A. C., private communication, 1966.

[54] Pekola J., Schoelkopf R. and Ullom J., *Physics Today*, May issue (2004) 41.

[55] Giazotto F., Heikkilä T. T., Luukanen A., Savin A. M. and Pekola J. P., *Rev. Mod. Phys.*, **78** (2006) 217.

[56] `www.nanoway.fi`

[57] Spietz L., Lehnert K. W., Siddiqi I. and Schoelkopf R. J., *Science*, **300** (2003) 1929.

This page intentionally left blank

High-temperature metrology and prospects for a new definition of the kelvin

J. Fischer

Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin
Abbestr. 2-12, 10587 Berlin, Germany

1. – Introduction: a brief course into history

The temperature of a body is its thermal state, regarded as a measure of its ability to transfer heat to other bodies. At present, this definition compels the attribution of larger numerical values to those bodies which have a higher ability to transfer heat to other bodies. Science took a long and difficult route, full of errors, to this contemporary definition of temperature.

Any historic "temperature scale" as those of Celsius and Fahrenheit depends upon the chosen thermometric working substance and its chosen property. A true thermodynamic temperature definition was not possible until 1848 when its foundations were laid by William Thomson, who later became professor for natural philosophy in the university of Glasgow, and assumed the title Lord Kelvin of Largs. He realised that for a special heat engine operating over a reversible Carnot cycle, the ratio of the heat Q_1 taken in at the higher temperature to that given out Q_2 at the lower temperature is

$$Q_1/Q_2 = T_1/T_2, \tag{1}$$

provided that T_1 and T_2 are so-called thermodynamic temperatures. Equation (1) is independent of the work W done and this is the essential point, this definition of the temperature is independent of any chosen thermometric working substance.

The definition of the quantity thermodynamic temperature has to be completed by assigning a numerical value to an arbitrary fixed point of temperature. In 1954 the General Conference on Weights and Measures (CGPM) adopted the current definition for the thermodynamic temperature scale proposed by Kelvin about a century ago. The temperature unit kelvin is defined as the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (TPW). By assigning an arbitrary value to the temperature of the TPW, at the same time the numerical value of the Boltzmann constant, k, is established. To demonstrate this, we consider an ideal gas as a system of microscopic particles. Starting with the relation $S = k \ln P$ (S entropy, P thermodynamic probability) of statistical mechanics the state equation is derived

$$pV_{\mathrm{m}} = N_{\mathrm{A}} kT, \tag{2}$$

where V_{m} is the molar volume and N_{A} the Avogadro constant, the number of particles per mole. The quantity kT is a characteristic energy which determines the energy distribution among the particles in the gas in thermal equilibrium.

However, the ideal gas is as fictitious as the reversible Carnot cycle. But using a noble gas at very low pressure, a thermometer can be constructed that really measures thermodynamic temperatures. In fact, the gas thermometer was the first thermometer to derive thermodynamic temperatures. Such kind of thermometers, whose basic relation between the measurand and T can be written down explicitly without having to introduce unknown, temperature-dependent constants, are generally called primary thermometers. Important methods to determine thermodynamic temperatures are reviewed in the next section. Any accurate realisation of the thermodynamic temperature is very time consuming and requires extreme metrological effort. A gas thermometer is only appropriate for use as a primary standard in fundamental laboratory measurements. On this ground and to harmonise the differing national temperature scales for the day-to-day practical use, so called International Temperature Scales have been developed by the Consultative Committee of Thermometry (CCT) and adopted by the CGPM. The International Temperature Scales reflect the most recent state of metrological accuracy and therefore they are replaced by new versions from time to time (see subsect. **8**˙1).

2. – Measurement of thermodynamic temperature

The primary thermometers which are currently under development or have produced results very recently are listed in fig. 1 (cf. ref. [1]). The envisaged or achieved temperature ranges are marked with grey bars. On top the optical thermometers based on Fourier spectroscopy (fts) developed at the National Research Council, (NRC), Canada [2] as well as on spectral [3-5] and total radiation thermometry [3] are listed. The latter will be described in detail in sects. **5** and **6**. Noise thermometry work [6-8] can be found in sect. **4**. At the National Physical Laboratory (NPL), UK, a primary thermometer on the basis of Rayleigh scattering is being developed [9]. At the bottom of the graph recent gas thermometric work is listed which is performed at Physikalisch-Technische Bundesanstalt (PTB), Germany, employing Dielectric-Constant Gas Thermometry (DCGT) [10] and at the French and Italian metrology institutes INM/LNE [11], INRIM, [12], the National

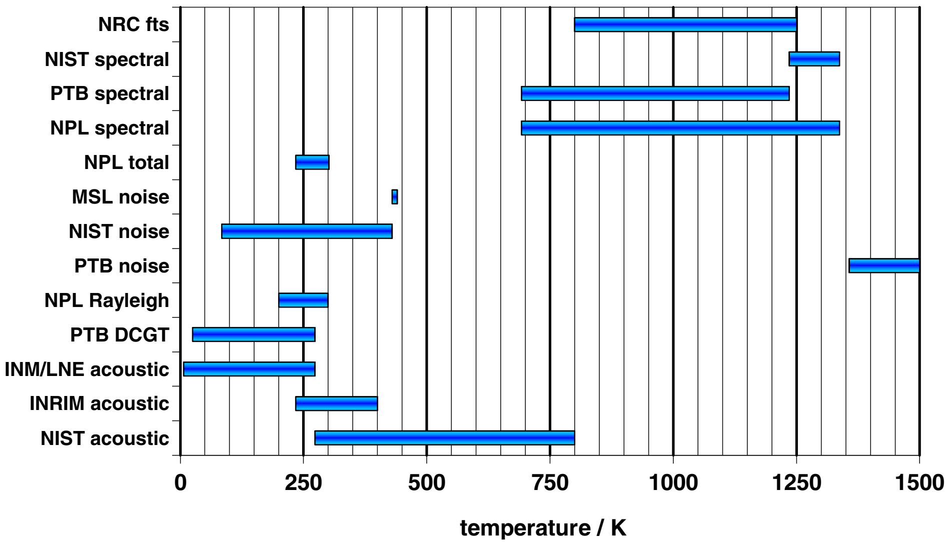


Fig. 1. – Primary thermometers recently used or currently under development with temperature ranges. For the abbreviations see the text above.

Institute of Standards and Technology (NIST), USA, [13], and the University College London (UCL), UK, [14] using Acoustic Gas Thermometry (AGT). Gas thermometers are discussed in the next section.

With a few exceptions thermodynamic temperatures are only determined to establish internationally agreed temperature scales (cf. subsect. **8**˙1). Only at very low temperatures thermometers are calibrated with thermodynamic methods of measurement. The determination of the Boltzmann constant applying primary thermometers is discussed in sect. **9** (for details, see ref. [15]).

3. – Gas thermometry

All three modern methods of gas thermometry, Constant-Volume Gas Thermometry (CVGT), AGT, and DCGT are based on different simple relations between the properties of an ideal gas and thermodynamic temperature (see fig. 2). CVGT is reviewed here only for historical reasons and to introduce the principle of a primary thermometer. For more details of AGT, see the contribution of Gavioso, this volume, p. 455.

Though many gases exhibit a nearly ideal behaviour at and above the TPW, in view of the desired level of accuracy, even at these temperatures, the small departures from the ideal behaviour must be carefully considered. This is done by measuring the relevant property in dependence on the density. Then, the ideal behaviour is deduced by an extrapolation to zero density applying an appropriate virial expansion. Two general facts should be emphasised. First, AGT and DCGT are based upon the variation with temperature of an intensive property of the gas (speed of sound u_0 and dielectric con-

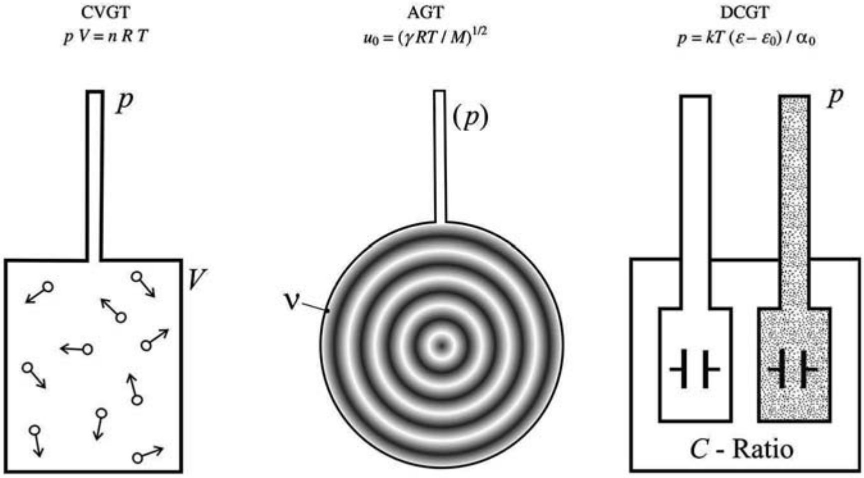

CVGT
$p V = n R T$
p
V
AGT
$u_0 = (\gamma RT / M)^{1/2}$
(p)
ν
DCGT
$p = kT(\varepsilon - \varepsilon_0) / \alpha_0$
p
C - Ratio

Fig. 2. – Principles of CVGT, AGT, and DCGT (for the symbols see text, the simple relations shown here are valid for an ideal gas).

stant ε, respectively), whereas primary CVGT requires a knowledge of the number n of moles of gas present in the gas bulb with volume V. Second, for AGT the dependence of the speed of sound on the pressure p is a second-order effect in the virial expansion.

3˙1. *Constant-volume gas thermometers.* – The classical gas thermometer is based on the equation of state for an ideal gas (2). At sufficiently low pressures and densities, the behaviour of a real gas can be described by the virial expansion ($B(T)$ and $C(T)$ second and third density virial coefficients, respectively, their values depend on temperature, $V_\mathrm{m} = V/n$ molar volume, R molar gas constant)

$$pV = nRT(1 + B(T)/V_\mathrm{m} + C(T)/V_\mathrm{m}^2 + \ldots). \tag{3}$$

Primary CVGT can be performed using two methods, the absolute and relative pV-isotherm CVGT, respectively. In the method of absolute pV-isotherm CVGT, the gas bulb at a constant but unknown temperature T is filled with a series of increasing amounts of gas to obtain a series of pressures p. pV/nR may be plotted as a function of $1/V_\mathrm{m}$ according to (3). The intercept of the resulting isotherm is the temperature T. For this method, it is not necessary to know the virial coefficients because the extrapolation to zero pressure is made by fitting a virial expansion to the experimental data. There are different ways of measuring the amount of gas. One method is to weigh a bulb of known volume with and without the gas sample and deduce the amount from the weight difference. This method has not been used in recent high-precision CVGT owing to the practical difficulties inherent in the weighing of low-density gases. It was used in historical measurements at the TPW to determine the molar gas constant R. In view of the careful analysis of these measurements in [16] considering also systematic effects related to gas sorption, to the measurement of the gas-bulb volume, and to the pressure dilatation of the bulb, it seems at present not possible to reach a level of relative uncertainty below about ten parts in 10^6 (10 p.p.m.).

Relative pV-isotherm CVGT is performed by measuring the pressures p_{r} and p of the same gas sample in the thermometer bulb at a known thermodynamic reference temperature T_{r} and the unknown temperature T. If the bulb volume V is assumed to be independent of temperature and pressure, in the case of an ideal gas, T is obtained from the equation

$$T = T_{\mathrm{r}} p / p_{\mathrm{r}}. \tag{4}$$

For a real gas, taking into account the second and third density virial coefficients as well as only terms up to the second order, this becomes

$$\frac{T}{T_{\mathrm{r}}} = \frac{p}{p_{\mathrm{r}}} \times \frac{1 + B(T_{\mathrm{r}})(p_{\mathrm{r}}/RT_{\mathrm{r}}) + \big(C(T_{\mathrm{r}}) - B(T_{\mathrm{r}})^2\big)(p_{\mathrm{r}}/RT_{\mathrm{r}})^2}{1 + B(T)(p/RT) + (C(T) - B(T)^2)(p/RT)^2}\,. \tag{5}$$

The volume of the gas bulb changes with temperature due to the thermal expansion of the bulb material. The thermal expansion must therefore be known as exactly as possible, as it is one of the main sources of error in gas thermometry. A change in the gas pressure also causes the bulb to dilate, and this change in volume must either be calculated, or the gas bulb must be surrounded by a second bulb filled with gas at the same pressure as the gas in the first bulb. In this case the dilation of the bulb due to pressure is negligible. The CVGT constructed by Guildner, Edsinger and Schooley was used for establishing the ITS-90 in the range from 0 °C to 660 °C [17, 18]. The estimated standard uncertainty ranges from a few mK near room temperature to about 10 mK near 660 °C.

3˙2. *Acoustic gas thermometers.* – For an ideal gas, the relation between the speed of sound u_0 and the thermodynamic temperature T is given by the equation ($\gamma = C_p/C_V$ adiabatic exponent, M molar mass of the gas)

$$u_0 = (\gamma RT/M)^{1/2}. \tag{6}$$

The pressure dependence of the speed of sound in a real gas may be expressed by a virial expansion

$$u^2 = u_0^2(1 + \alpha p + \beta p^2 + \ldots), \tag{7}$$

i.e. the influence of the pressure is of second order. The acoustic virial coefficients α and β can be expressed in terms of the density virial coefficients $B(T)$ and $C(T)$. As can be seen from eq. (6), the gas constant R and the molar mass M influence directly the result. Thus, the uncertainty of T depends on their uncertainties. The influence of R and M may be eliminated if the speed of sound is determined at the temperature of the TPW with the same thermometer and the same gas, *i.e.* if relative isotherms are measured. Applying (7), u_0 is determined by extrapolation to zero density and T is calculated using eq. (6).

Two methods have been used to measure the speed of sound [19, 20]. In older works, a fixed-frequency, variable-path, cylindrical acoustic interferometer was used. Nowadays, variable-frequency, fixed-path spherical resonators are preferred. Their figure of merit is about an order of magnitude higher than that of cylindrical resonators. Furthermore, boundary layer effects and the problems due to the excitation of different modes are essentially smaller. Colclough [21] published a paper on an acoustic thermometer which was used for measurements of T in the temperature range from 4.2 K to 20 K with an uncertainty of 1 mK to 4 mK using a cylindrical resonator. A similar instrument was used at a temperature of 273.16 K for a re-determination of the gas constant R [22].

Moldover *et al.* [23] assembled a 3 litre, steel-walled spherical resonator and used it during 1986 to re-determine the gas constant R with an estimated relative uncertainty of 1.7 p.p.m., a factor of 5 smaller than the uncertainty of the best previous measurement. They measured the acoustic resonance frequencies of the argon-filled resonator and the microwave resonance frequencies of the same cavity when evacuated. The microwave data was used to deduce the thermal expansion of the cavity. In 1986, Moldover and Trusler [24] used the same "gas-constant" cavity for temperature measurements close to room temperature and determined the temperature of the triple point of gallium (near 303 K). More recent measurements were conducted primarily during 1992 which led to new values of $T - T_{90}$ (T_{90} temperature according to the ITS-90) between 217 K and 303 K [25]. Ewing and Trusler [14] performed thermodynamic measurements in the range from 90 K to 300 K. The estimated standard uncertainty ranges from 0.9 mK to 1.3 mK.

Later, a new spherical resonator was developed at NIST including new acoustic transducers to cover the temperature range from 273 K to 800 K [13]. First results have been obtained for the melting point of gallium and the freezing points of indium and tin [26]. Independent acoustic measurements of the thermodynamic temperature have been performed by Benedetto *et al.* [12] between 234 K and 380 K and Pitre *et al.* [11] from 7 K to 24.5 K and from 90 K to 273 K using also spherical resonators. The results agree within the remarkably small combined uncertainty with those presented in refs. [24-26].

3˙3. *Dielectric-constant gas thermometers.* – The basic idea of DCGT is to replace the density in the state equation of a gas by the dielectric constant. The dielectric constant of an ideal gas is given by the relation $\varepsilon = \varepsilon_0 + \alpha_0 N/V$, where ε_0 is the exactly known electric constant, α_0 is the static electric dipole polarizability of the atoms, and N/V is the number density, *i.e.* the state equation of an ideal gas can be written in the form $p = kT(\varepsilon - \varepsilon_0)/\alpha_0$. DCGT avoids, therefore, the troublesome density determination of the conventional gas thermometry that is complicated by the gas contained in dead spaces not at the measuring temperature T and by gas adsorption in the system. Other advantages are that the pressure sensing tubes can be of any convenient size and that the thermometric gas can be moved in or out of the thermometer cell without the need to allow for the amount of the gas involved.

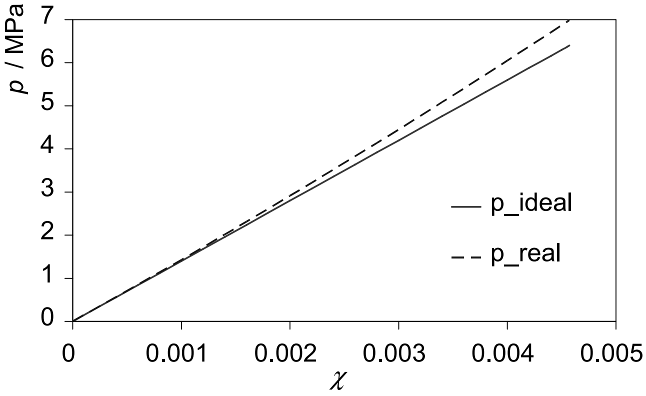


Fig. 3. – Pressure *vs.* dielectric susceptibility isotherms of helium at the TPW with (broken curve, p_{real}) and without (straight, full curve, p_{ideal}) consideration of the interaction between the atoms.

DCGT can be used for primary thermometry in two ways. Absolute measurements require to know the static electric dipole polarizability α_0 with the necessary accuracy. Nowadays this condition is fulfilled for helium, which became a model substance for evaluating the accuracy of *ab initio* calculations of thermophysical properties. Recent progress has decreased the uncertainty of the *ab initio* value of α_0 well below one part in 10^6 [27]. One disadvantage of helium is its relatively small polarizability, which is for instance smaller than that of argon by a factor of eight. On the other hand, primary thermometry does not require a value of α_0 if measurements are made both at the TPW (T_{TPW}) and at the measuring temperature T. The ratio of the two temperatures is given by $T/T_{\mathrm{TPW}} = (p/p_{\mathrm{TPW}})(\varepsilon_{\mathrm{TPW}} - \varepsilon_0)/(\varepsilon(T) - \varepsilon_0)$, where the quantities measured at the TPW have the corresponding index.

For a real gas, the interaction between the particles has to be considered by combining the virial expansions of the state equation and the Clausius-Mossotti equation. When neglecting higher-order terms and the dielectric virial coefficients this yields

$$p \approx \frac{\chi}{\frac{3A_\varepsilon}{RT} + \kappa_{\mathrm{eff}}} \left[1 + \frac{B(T)}{3A_\varepsilon}\chi + \frac{C(T)}{(3A_\varepsilon)^2}\chi^2 + \ldots \right], \tag{8}$$

where $\chi = \varepsilon/\varepsilon_0 - 1$ is the dielectric susceptibility, $A_\varepsilon = R\alpha_0/(3\varepsilon_0 k)$ is the molar polarizability, B and C are the second and third density virial coefficients considering the pair and triplet interactions, respectively, and κ_{eff} is the effective compressibility of a suitable capacitor used to measure the susceptibility χ. For determining $3A_\varepsilon/RT$ isotherms have to be measured, *i.e.* the relative change in capacitance $(C(p) - C(0))/C(0) = \chi + (\varepsilon - \varepsilon_0)\kappa_{\mathrm{eff}} p$ of the gas-filled capacitor is determined as a function of the pressure p of the gas: The capacitance C of the capacitor is measured with the space between its electrodes filled with the gas at various pressures and with the space evacuated so that $p = 0\,\mathrm{Pa}$. A polynomial fit to the resulting p *vs.* χ data points (cf. fig. 3 and eq. (8)), together with the knowledge of the dependence of the dimensions of the capacitor on p (κ_{eff}), yields $3A_\varepsilon/RT$. Because of the very small susceptibility (for instance 0.003 at 4 K and 0.05 MPa for helium), this technique causes extreme demands concerning the measurement of capacitance changes.

Primary DCGT using cylindrical capacitors has been performed by two groups in the temperature range from about 3 K to 27 K [28-30]. The results of both groups coincide with the scale NPL-75 established using CVGT within the combined uncertainty. This is an important confirmation of the thermodynamic accuracy of the NPL-75 and thus the ITS-90 because, apart from the pressure measurement, DCGT and CVGT have quite different error sources.

An evaluation of the results obtained so far considering especially the recent progress in the measurement of pressure and capacitance changes shows that DCGT has potential for both decreased uncertainty and increased application range. Measurements up to the TPW would be desirable because the density is measured *in situ*, *i.e.* the thermal expansion of the thermometer cell does not cause problems as in CVGT. Furthermore, as an inverse application of DCGT, measurements at the triple-point of water allow to determine A_ε/R (see eq. (8)) and thus the Boltzmann constant $k = (A_\varepsilon/R)\alpha_0/(3\varepsilon_0)$. The measuring system described in ref. [30] would allow to achieve a standard uncertainty of k of about 15 parts in 10^6, but it seems to be at least possible to build systems, which yield an uncertainty being an order of magnitude smaller [10]. Another inverse application of DCGT has been discussed in ref. [31]: realisation of a pressure standard near 1 MPa applying a toroidal cross capacitor filled with helium. Since the interaction between the helium atoms at the realised pressure has to be considered, this project requires values calculated *ab initio* for both the polarizability α_0 and the second density virial coefficient B.

4. – Noise thermometry

The noise thermometer is based on the temperature dependence of the mean square noise voltage, $\langle U^2\rangle$, developed in a resistor. Nyquist derived eq. (9) from thermodynamic calculations [32]

$$\left\langle U^2\right\rangle = 4kTR\Delta f \tag{9}$$

valid for frequencies $f \ll kT/h$, where R is a frequency-independent resistance, Δf the bandwidth, and h Planck's constant. From the statistical nature of the measured quantity, long measuring times arise which may be estimated from eq. (10)

$$\Delta T/T \approx 2.5/\sqrt{t\Delta f}, \tag{10}$$

with t being the measuring time. One of the main problems is the accurate measurement of the very small voltages developed avoiding extraneous sources of noise and maintaining constant bandwidth and gain of the amplifiers. For details, see refs. [19, 33, 34]. In the

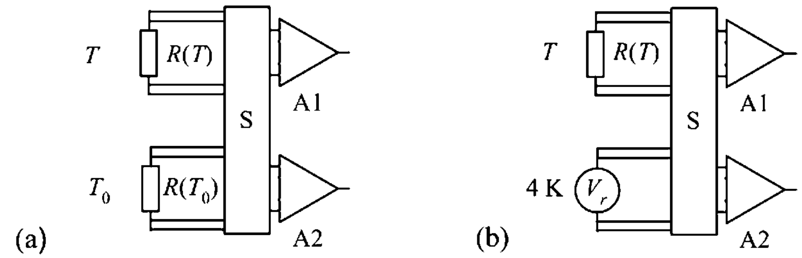


Fig. 4. – (a) Block diagram for the conventional relative method with switched-input noise correlator, S: switching, A1, A2: amplification and digitisation. (b) Block diagram for the new absolute method.

past, in the high-temperature range the uncertainties of noise thermometry have not been comparable to those of gas-based techniques due to limitations from the non-ideal performance of electronic detection systems. The most successful technique to date is the switched input digital correlator pioneered by Brixy and others [35, 8], see fig. 4(a). The correlator is implemented by digitizing the signals from the two channels and carrying out the multiplication and averaging function of the correlator by software. This eliminates the amplifier and transmission line noise superimposed on the thermal noise signal. In use, the thermometer switches between a reference noise source at a reference temperature T_0 and the noise source at the unknown temperature T. The switching removes the effects of drift in the gain and bandwidth of the amplifiers and filters. A relative standard uncertainty of $2 \cdot 10^{-5}$ at the zinc fixed point has been estimated.

The conventional switching approach does not contain sufficient free parameters to resolve the contradictory requirements for matching both the sensing resistances and the noise powers. Currently, a collaboration between NIST and MSL [7] explores a new approach using the perfect quantization of voltages from the Josephson effect. This approach keeps the proven elements of the switched correlator, but separates the roles of the temperature reference and the voltage reference. The sensing resistor in the reference arm of the comparator is replaced by an a.c. Josephson voltage standard. A block diagram is shown in fig. 4(b). The long-term goal of the project is to build a noise thermometer with an uncertainty of $1 \cdot 10^{-5}$ over the temperature range 83 K to 430 K. The first step was a proof of concept at the triple points of gallium and water [36]. Low temperature noise thermometry is reviewed in the contribution of Rusby in this volume, p. 393.

5. – Total radiation thermometry

It is only with the development of the cryogenic radiometer that accurate measurements have become possible of the total radiation emitted from a blackbody. The total radiant exitance $M(T)$ of a blackbody at a temperature T is given by

$$M(T) = \frac{2\pi^5 k^4}{15h^3c^2}T^4 = \sigma T^4, \tag{11}$$

where σ is the Stefan-Boltzmann constant and c the speed of light. To determine the total radiant exitance, it is necessary for practical reasons to make measurements over only a restricted solid angle rather than over a complete hemisphere. An aperture system must be interposed between the blackbody and the detector so that $M'(T) = gM(T)$, where g is the throughput of the optical system. Provided that the g is independent of temperature, we may write

$$\frac{M(T)}{M(T_{\text{TPW}})} = \frac{M'(T)}{M'(T_{\text{TPW}})} = \left(\frac{T}{T_{\text{TPW}}}\right)^4 \tag{12}$$

where $M'(T)$ and $M'(T_{\text{TPW}})$ are the quantities to be determined. Near room temperature a measurement of the ratio $M'(T)/M'(T_{\text{TPW}})$ to 1 part in 10^5 is sufficient to determine T to 1 mK or better. Quinn and Martin [37] and Martin *et al.* [38] have demonstrated that such measurements are possible and have obtained results between −130 °C and 100 °C. The system they used is illustrated in fig. 5. The blackbody radiator at a temperature T between 143 K (−130 °C) and 373 K (100 °C) irradiates an aperture system at liquid-helium temperatures, which allows a beam of thermal radiation to enter a second blackbody held initially at a temperature of 2 K. The absorbing blackbody acts as a heat flow calorimeter or cryogenic radiometer. The radiant power absorbed in the calorimeter leads to a rise in its temperature until the radiant power absorbed is balanced to a close approximation by the heat flow along a poorly conducting heat link to a heat sink maintained at a very stable temperature near 2 K. The temperature rise of the calorimeter, which amounts to about 3 K for a radiating blackbody temperature of 273 K, is monitored. When equilibrium has been reached, a shutter at liquid helium temperature is closed, cutting of radiation from the blackbody radiator. At the same time, sufficient electrical power is supplied to a heater on the calorimeter to keep it at the same temperature. Provided that a number of conditions are met, this easily measurable electrical power is a very precise equivalent of the thermal radiative power.

In carrying out measurements of the ratio $M'(T)/M'(T_{\text{TPW}})$, there are the following parameters of the system that had to be evaluated: the emissivity of the radiator, its effective temperatures and the absorptivity of the calorimeter, the diffraction effects at the apertures, the effects of scattering of thermal radiation from surfaces between the apertures, the absorption of thermal radiation at the aperture edges, the departures from ideal geometry of the apertures, the equivalence of radiant and electrical heating of the calorimeter, the uncertainty in the measurements of the electrical power applied to the calorimeter, and the energy transfer from radiator to calorimeter by residual gas. Although some of these parameters had to be known absolutely, others needed to be known only to the extent that their dependence on wavelength or temperature of the radiator was required. An absolute measurement of $M'(T_{\text{TPW}})$ for a determination of

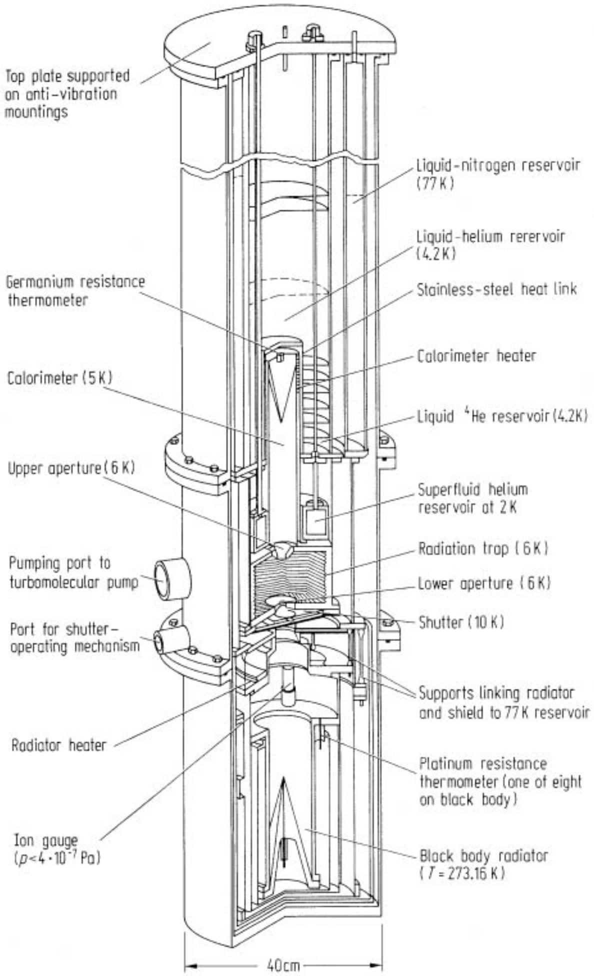

Top plate supported on anti-vibration mountings
Liquid-nitrogen reservoir (77 K)
Liquid-helium rerervoir (4.2 K)
Germanium resistance thermometer
Stainless-steel heat link
Calorimeter heater
Calorimeter (5 K)
Liquid ^{4}He reservoir (4.2K)
Upper aperture (6 K)
Superfluid helium reservoir at 2 K
Radiation trap (6 K)
Pumping port to turbomolecular pump
Lower aperture (6 K)
Port for shutter-operating mechanism
Shutter (10 K)
Supports linking radiator and shield to 77 K reservoir
Radiator heater
Platinum resistance thermometer (one of eight on black body)
Ion gauge ($p<4\cdot10^{-7}$ Pa)
Black body radiator ($T=273.16$ K)
40cm

Fig. 5. – Cryogenic radiometer after ref. [37].

the Stefan-Boltzmann constant, however, would require an absolute knowledge of all of them [37]. The result of the determination of the Stefan-Boltzmann constant had an uncertainty of about 1.3 parts in 10^4.

Recently NPL has built a new version of a total radiation thermometer which is expected to measure σ with an uncertainty of 0.001% and thermodynamic temperatures between the Hg and Sn fixed points [39] with uncertainties of around 0.5 mK. It should be noted that a determination of σ with an uncertainty of 0.001% corresponds to a determination of the Boltzmann constant with an uncertainty of 2.5 parts in 10^6, similar to that achieved with acoustic thermometry [23].

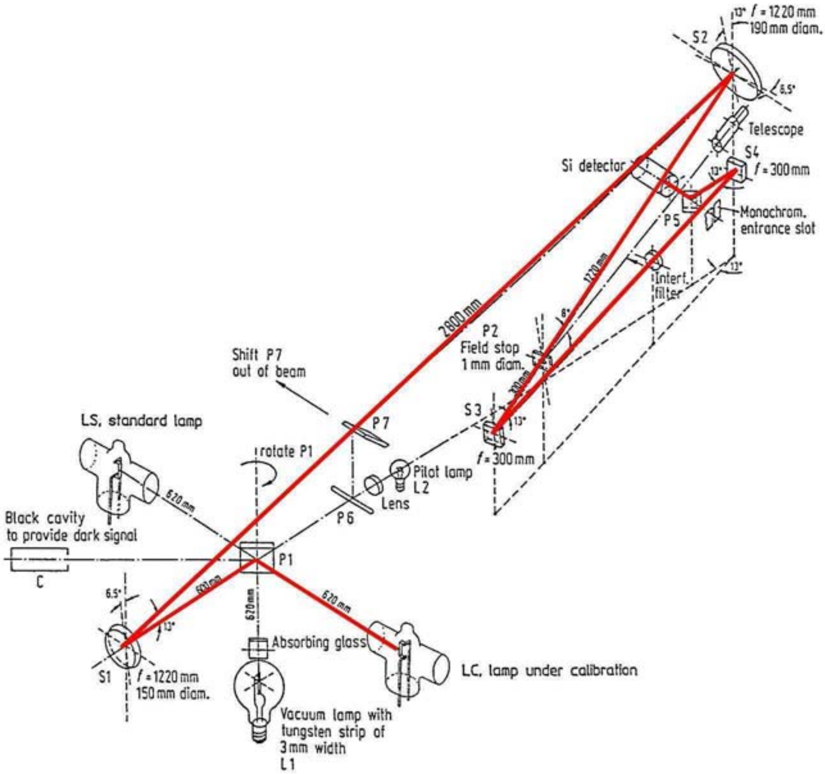


Fig. 6. – Optical layout of the spectral radiation thermometer of PTB. In place of the lamps LC, LS blackbodies B1 and B2 can be mounted.

6. – Spectral radiation thermometry

6˙1. *Relative mode referenced to known temperature.* – Equation (13) defines the ITS-90 in the temperature range above the silver point ($L_\lambda(\lambda, T)$ spectral radiance at a temperature T and wavelength λ, c_2 is the second radiation constant). However, the measurement of radiance ratios can be used as well to measure thermodynamic temperature if the temperature of one of the blackbodies is known in terms of thermodynamic temperature as a reference. This method has been applied to determine the temperatures of the Al, Ag, and Au fixed points of the ITS-90 and will be described in the following:

$$\frac{L_\lambda(\lambda, T_{90})}{L_\lambda(\lambda, T_{90,\mathrm{ref}})} = \frac{\exp\left[c_2/\left[\lambda T_{90,\mathrm{ref}}\right]\right] - 1}{\exp\left[c_2\left[\lambda T_{90}\right]\right] - 1}. \tag{13}$$

The spectral radiation thermometer used by Jung and Fischer [40-42] in the range 410 °C to 962 °C employed two simultaneously running blackbodies B1 and B2. They are observed alternatively by means of a rotatable plane mirror P1 (fig. 6). One of the black-

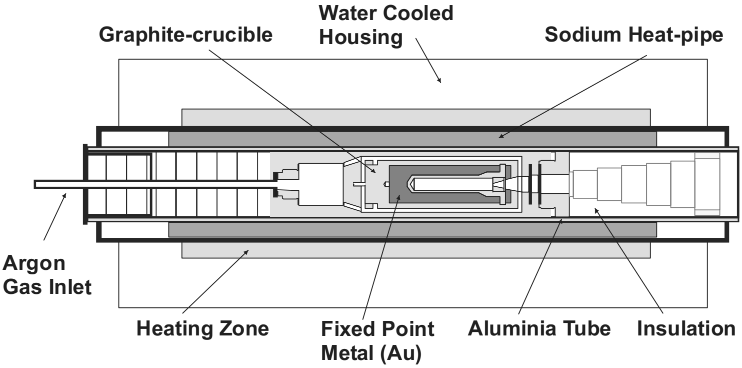


Fig. 7. – Schematic of the PTB gold fixed-point blackbody.

bodies has the unknown temperature, the other has the reference temperature. The radiation thermometer alternatively focused the apertures of both blackbodies onto a linear detector. An interference filter defined the wavelength. The unknown temperature is calculated from the photocurrent ratio and the reference temperature using Planck's radiation law and taking into account the spectral responsitivity of the radiation thermometer. During a first run of measurements, blackbody B1 had a reference temperature close to $T_{\mathrm{ref}} = 729\,\mathrm{K}$, a thermodynamic temperature as derived from gas thermometry [17, 18]. Blackbody B2, immersed in freezing aluminium, had the unknown temperature.

During the second run of measurements, blackbody B2, constantly running aluminium freezes, served as reference radiator. Unknown was the thermodynamic temperature T that corresponded to the temperature of blackbody B1. The latter was set to values between 410 °C and 630 °C while the corresponding thermodynamic temperatures came from the measured photocurrent ratios and the re-determined aluminium point. The results of Jung [40, 41], measured with two radiation detectors of different spectral sensitivity, agreed within one standard deviation. The more accurate result, obtained with the detector of best stability and linearity, gave rise to the SPRT reference function used for the ITS-90. These results are backed by the good agreement obtained at the BIPM [43] and the NPL [44]. Finally, during a third run the freezing points of silver and gold have been determined at PTB [42]. The apparatus was the same as in fig. 6 with the only exception that both blackbodies were of the freezing-point type according to fig. 7. The reference blackbody contained aluminium and the other one freezes of silver, respectively, gold.

6˙2. *Absolute mode.* – The measurement of radiance ratios removes the need to know the spectral responsivity of the radiation thermometer absolutely. However, to determine the spectral radiance without referencing to a source of known temperature requires an instrument which has a known absolute spectral response with a well-defined geometric viewing system. This system is called in its simplest version a filter radiometer and has a set of two view-defining apertures with accurately known dimensions. It is only in

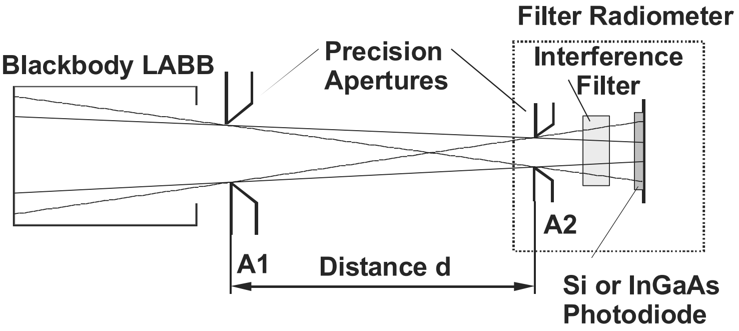


Fig. 8. – Optical layout of the PTB experiment with the large-area blackbody LABB, two precision apertures A1 and A2, and the filter radiometer to measure $T - T_{90}$.

relatively recent times that it has become possible to measure the absolute spectral responsivity of a filter radiometer with sufficient accuracy to compete with the method used in the measurement of radiance ratios. This became possible through the use of cryogenic radiometers as primary reference standard defining a scale of spectral responsivity. The first cryogenic radiometer was constructed at NPL [37] to measure the total radiation as described above. Cryogenic radiometers to establish scales of spectral responsivity use the spectral power of a monochromatic source such as a laser and subsequently calibrate the response of a transfer detector [3]. The additional uncertainty of the calibration of the filter transmission has to be taken into account, which cannot be lower than the uncertainty of the cryogenic radiometer used for calibration. Altogether, this results in best relative uncertainties at the 10^{-4} level for this type of radiation thermometry and measurements at temperatures of about 500 °C and above [3-5].

At the time of establishment of the ITS-90 there were no direct measurements of the thermodynamic temperature T above 729 K of sufficiently low uncertainty that could be used to anchor the scale. Instead, the fixed points of Al, Ag, Au, and Cu were determined through ratios of radiance using radiation thermometers. The temperature value assigned to the reference point at 729 K (456 °C) was the mean of two separate NIST gas thermometry experiments [45]. Since these two experiments differed by around 30 mK, this became the dominant uncertainty of ITS-90 realisations for all higher temperatures. This relatively large thermodynamic uncertainty of T_{90} propagates as T_{90}^2 and is thus around 50 mK at the gold point. To reduce these uncertainties, work is in progress applying acoustic thermometry [13, 26] and spectral radiation thermometry [3-5].

Using absolute filter radiometry measurements of $T - T_{90}$ in the temperature range 660 °C to 962 °C [46] and subsequently down to zinc-point temperatures [47, 5] have been performed at PTB. They employed a large-area blackbody formed by two concentric sodium heat pipes so that the filter radiometer does not need imaging optics whose transmittance is to be calibrated, see fig. 8. The temperature T_{90} of the blackbody was defined by three high-temperature standard platinum resistance thermometers. The obtained temperature differences $\Delta T = T - T_{90}$ are fitted by the relation

$$\Delta T = (T/T_{\mathrm{ref}})^2 \Delta T_{\mathrm{ref}}, \tag{14}$$

with $T_{\mathrm{ref}} = 729\,\mathrm{K}$ and $\Delta T_{\mathrm{ref}} = 22\,\mathrm{mK}$. This result indicates that the higher of the two reference temperatures derived from gas thermometry [17, 18] should be used for an improved temperature scale.

A different approach is applied at NIST [4] and at NPL [3] where imaging filter radiometers were developed. With these systems the thermodynamic temperatures of fixed-point blackbodies with their small cavity apertures can be measured directly at temperatures down to the zinc fixed point.

7. – Doppler broadening thermometry

Though Doppler broadening thermometry is a standard means of diagnostics for high-temperature plasmas [48], this method of temperature measurement is less well known in other fields of thermometry and so is described in some detail here. The method is based on the Doppler shift of the frequency of an electromagnetic wave in a moving frame of reference as compared to a frame at rest, and its theoretical basis is particularly transparent.

Consider a plane electromagnetic wave with frequency ν and wave vector $\boldsymbol{\kappa}$ in the laboratory frame of reference, say, and an atom or molecule moving with constant velocity $\boldsymbol{v} = c_0\boldsymbol{\beta}$ in this frame. As a result of the corresponding Lorentz transformation, the frequency observed in the atomic rest frame is Doppler-shifted. For atomic thermal velocities at temperatures of about 300 K, which are of interest in the present context, typical values of $\beta = v/c_0$ are some 10^{-6}, so that it may be sufficient to consider only the linear Doppler shift given by $\nu' = \nu(1 - \boldsymbol{\beta}\cdot\boldsymbol{\kappa}/\kappa)$, but it should be borne in mind that the relativistic quadratic Doppler effect gives rise to corrections at the p.p.m. level. Now consider the situation the other way round and assume that the atom has a sharp absorption resonance at a frequency which is ν_0 for an atom at rest, and that a tunable laser at rest in the laboratory is used to irradiate the atom moving with velocity $\boldsymbol{v}$. Then in a linear approximation, laser radiation will be absorbed if the laser is tuned to $\nu = \nu_0(1 + \boldsymbol{\beta}\cdot\boldsymbol{\kappa}/\kappa)$.

Finally, consider the laser beam to propagate along the x-axis through an absorption cell containing an ideal gas of such atoms or molecules with uniform temperature T. The Gaussian Maxwell probability density for v_x is proportional to $\exp[-(v_x/v_0)^2]$ with $v_0^2 = 2kT/m$ for atomic mass m. Around the absorption frequency ν_0 this translates into the corresponding Doppler-broadened absorption line profile with the Doppler width $\Delta\nu_{\mathrm{D}} = [2kT/(mc_0^2)]^{1/2}\cdot\nu_0$. It is this relation which allows the determination of the value of the Boltzmann constant by spectroscopic measurement of a Doppler-broadened absorption line profile and determination of its width.

In principle, the measurement can be done using standard laser-spectroscopic techniques. As a main advantage compared to absolute radiation thermometry, the Doppler

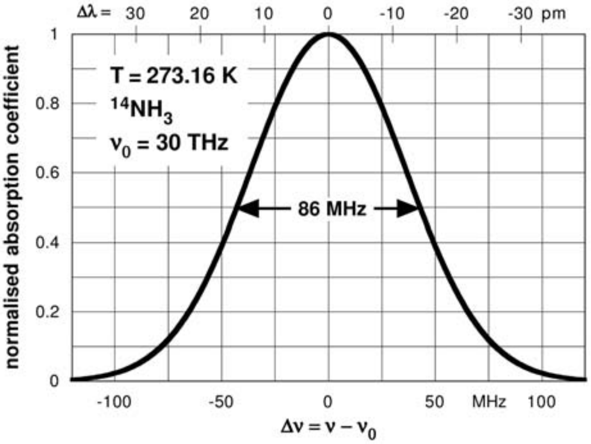


Fig. 9. – Doppler-broadened absorption line profile for ammonia $^{14}NH_3$ at T_{TPW} assuming a central frequency $\nu_0 = 30\,\mathrm{THz}$.

profile can be determined by relative radiation measurements since it is only its width which is of interest here. Moreover, laser frequencies can be controlled with extremely small uncertainties. However, at the 10^{-6} or 10^{-7} uncertainty level, various other sources of uncertainty will have to be investigated in detail. Apart from the quadratic Doppler effect mentioned above, these include, among many others, the effects brought about by interatomic interactions, notably the additional line broadening (collisional, transit time, and saturation broadening) and the reduction of Doppler broadening caused by a finite mean free path length (Dicke narrowing). Because of these, measurements may have to be performed at a series of pressure values and extrapolated to pressure zero. It should be noted that the extrapolation to zero pressure is problematic for reasons connected with the speed distribution of particles in monolayers on the cell's surface. Similarly, heating by the absorbed laser power may require an extrapolation to vanishing laser power if it cannot be neglected at all.

The possibility to obtain an accurate value of the Boltzmann constant has been demonstrated [49] in an experiment using an ammonia line probed by a CO_2 laser spectrometer close to 30 THz, see fig. 9. The absorption signal was recorded by splitting the laser beam in two, then propagating one of the two beams through the ammonia cell for spectroscopy while the other was used as a reference beam. The two beams were amplitude-modulated by two acousto-optic modulators at two different frequencies, then recombined and focused on a single photodetector. Two lock-in amplifiers were used for synchronous detection of the signal of each channel to determine the absorption signal numerically. The value of the Boltzmann constant deduced from the line width is equal to 1.3807(11) 10^{-23} J/K and agrees with that recommended by CODATA [50] within $3.6 \cdot 10^{-5}$ with a statistical relative uncertainty of $8 \cdot 10^{-4}$, limited mainly by the noise of the detected signal. Based on these preliminary results the authors [49] expect that the

Doppler broadening measurements can contribute to the determination of the Boltzmann constant with the required uncertainty (10^{-6}). As a goal to be reached within five years time, it seems to be challenging to decrease the noise by an order of magnitude and to increase the number of measurements by two orders of magnitude measuring over one month, which would yield a relative uncertainty of 10 p.p.m.

8. – Practical high-temperature metrology

8˙1. *The International Temperature Scale of 1990.* – The International Temperature Scales reflect the most recent state of metrological accuracy and therefore they are replaced with new versions from time to time. The International Temperature Scale of 1927 (ITS-27) was the first one to overcome the practical difficulties of the direct realisation of thermodynamic temperatures by gas thermometry and the first universally acceptable replacement for the differing existing national temperature scales. Finally on January 1 of 1990, the International Temperature Scale of 1990 (ITS-90) [51] came into force. We note that the thermodynamic temperature T, in units of K, may also be expressed as a Celsius temperature t according to

$$t/{}^{\circ}\mathrm{C} = T/\mathrm{K} - 273.15. \tag{15}$$

The ITS-90 accordingly defines both international kelvin temperatures T_{90} and international Celsius temperatures t_{90} by the corresponding relation:

$$t_{90}/{}^{\circ}\mathrm{C} = T_{90}/\mathrm{K} - 273.15. \tag{16}$$

Both the thermodynamic and the International Temperature Scale have the same units, the kelvin and the degree Celsius. Users sometimes prefer kelvin in the range below 273.15 K and degree Celsius above this point.

The thermodynamic basis of the ITS-90 is described in [45] and recommendations for its realisation are given in [52]. Later on it has been found that the scale deviates from thermodynamic temperatures more than originally assumed [53]. Nevertheless, compared with the former IPTS-68 and the EPT-76, it improves the reproducibility of temperature measurements, *e.g.* because it is smoother and prescribes more accurate instruments. The ITS-90 extends upwards from 0.65 K to the highest temperature practically measurable in terms of the Planck radiation law using monochromatic radiation. It is based on 17 well reproducible thermodynamic states of equilibrium, the defining fixed points: boiling points (3 K to 5 K with helium, 17 K, 20.3 K with hydrogen), triple points (equilibrium hydrogen, neon, oxygen, argon, mercury, water), melting point of gallium, and freezing points (indium, tin, zinc, aluminium, silver, gold, copper). To these states numerical values of the temperature T_{90} are assigned. These are values which have been determined by measurements of thermodynamic temperatures T in several national metrology institutes [45]. They are considered to be the best estimates at the time the scale was adopted. The defining fixed points are listed in table 1 in ref. [51] and shown in fig. 10 for temperatures above the TPW.

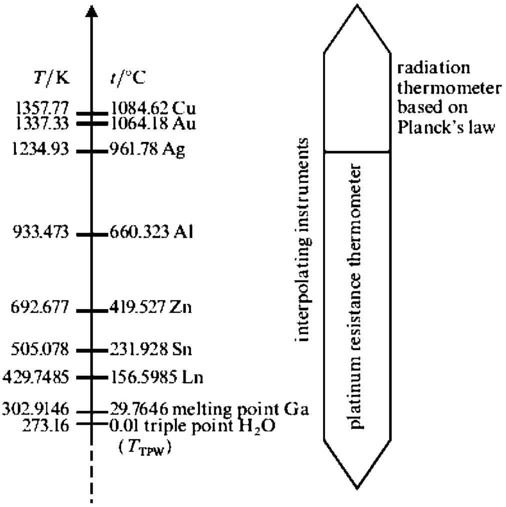


Fig. 10. – Schematic of the defining fixed points and interpolating instruments of the ITS-90 above 0.01 °C.

The interpolating instruments in the temperature range considered in this paper are the standard platinum resistance thermometer (SPRT, 14 K to 1235 K) and the radiation thermometer above. Thermocouples are not further used as interpolating instruments because of their poor reproducibility. For the first time, the ITS-90 comprises a number of ranges and sub-ranges, throughout each of which temperatures T_{90} are defined differently. Several of these ranges or sub-ranges overlap. At the highest metrological level, there are numerical differences resulting from using differing definitions, causing the so-called non-uniqueness [54]. However, in virtually all cases these differences are of negligible practical importance, see "Supplementary Information for the ITS-90" [52] and, *e.g.*, [55]. This has been achieved by two measures:

– An appropriate choice of the deviation functions (see below); in most of the sub-ranges linear or quadratic terms are sufficient [56].

– A fine adjustment applied to the temperature values of the fixed points of mercury, gallium, tin and aluminium. The magnitudes of the adjustments, mostly some tenths of a mK, are far below the uncertainties of the measured thermodynamic temperatures [56,45].

The temperature dependence of the resistivity of high-purity platinum is too complicated for a description based only on a calibration at the available fixed points of sufficient quality [57,19], but it is highly reproducible from sample to sample. Therefore, the ITS-90 prescribes two general reference functions for the ranges below and above the TPW representing a "typical" SPRT [51]. Values and derivatives of the two functions are continuous at the TPW. This allows to describe the characteristic of an SPRT by the sum of a reference function and an individual deviation function. The coefficients of the deviation function are deduced from the results of the calibration at the defining fixed points of the ITS-90. The calibration can be performed for eleven sub-ranges, which overlap more or less, using different deviation functions. To remove the influences of the dimensions of the platinum wire, both the reference and deviation functions are specified for the resistance ratio $W(T_{90}) = R(T_{90})/R(273.16\,\mathrm{K})$, where $R(T_{90})$ is the resistance at a temperature T_{90} and $R(273.16\,\mathrm{K})$ is the resistance at the TPW.

An acceptable SPRT must be made from pure, strain-free platinum and satisfy one of the relations to be checked at the triple point of mercury ($W(-38.8344\,^\circ\mathrm{C}) \leq 0.844235$) or the melting point of gallium ($W(29.7646\,^\circ\mathrm{C}) \geq 1.11807$). An acceptable SPRT to be used up to the freezing point of silver (961.78 °C) must also satisfy the relation: $W(961.78\,^\circ\mathrm{C}) \geq 4.2844$. The first two conditions guarantee a minimum purity of the platinum and the third is aimed to avoid SPRTs with excessive leakage currents at high temperatures. Three types of different designs cover the PRT range from 13.8 K to 962 °C, where the first two types have similar sensor elements: i) The capsule-type SPRT from 13.8 K to 273 K (sometimes 430 K). Characteristic data are: $R(0\,^\circ\mathrm{C}) = 25\,\Omega$, 5 mm diameter, 60 mm length, filled with 30 kPa helium at room temperature, the four platinum leads are taken out through a glass seal. ii) The conventional long-stem SPRT from −189 °C to 420 °C (sometimes 630 °C or 660 °C). Characteristic data are: $R(0\,^\circ\mathrm{C}) = 25\,\Omega$, 7 mm diameter, 600 mm length, filled with dry air. The sensors of 0.07 mm diameter platinum wire are supported by insulators of mica, alumina or silica. iii) The high-temperature long-stem PRT from 0 °C to 962 °C (suited to −189 °C). Characteristic data are: $R(0\,^\circ\mathrm{C})$ between $0.25\,\Omega$ and $2.5\,\Omega$, 7 mm diameter, 600 mm to 800 mm length, filled with 90% argon +10% oxygen at 20 kPa near room temperature. The sensors and leads are designed such as to minimise the strain on heating and cooling and the leakage through the insulation resistance. Platinum wire up to 0.4 mm diameter and insulators and sheaths of quartz are used to achieve this.

From 961.78 °C, the freezing point of silver, up to the highest practicably measurable temperatures T_{90} is defined in terms of the ratio of the spectral radiances $L_\lambda(T_{90})$ and $L_\lambda(T_{90,\mathrm{ref}})$ of two blackbodies (eq. (13)). One of them has the temperature T_{90} to be determined. The other has the reference temperature $T_{90,\mathrm{ref}}$ which stands for one of the freezing points of silver, gold or copper. The ratio is measured by means of a spectral radiation thermometer. The temperature T_{90} is calculated from the measured ratio using Planck's law for monochromatic radiation [51]. Appropriate designs of the apparatus and good current practice of their application are extensively described by the "Supplementary Information for the ITS-90" [52], in a monograph [58] and to some extent by the textbook [19]. Recently, the following overviews about practical temperature metrology have been published [59-62].

8·2. *Melting and freezing points of metals.* – In the case of the freezing and melting points of metals, the solid and liquid phases and a gas, *e.g.* argon or air, are in thermodynamic equilibrium. The reference pressure for the freezing and melting points is 101.325 kPa. In the past, often sealed fixed-point cells were used and are recommended in ref. [52]. But since the pressure cannot be measured in these cells, the pressure of the gas over the metal may be incorrect. Thus, for optimal realisation of the fixed-point temperatures, the fixed-point cells of the pure metals Ga, In, Sn, Zn, Al, and Ag must not be sealed; each cell must be equipped with a valve or some other arrangement so that the gas over the metal can be controlled and its pressure measured [63]. Also, the metal must always be kept under an inert gas atmosphere.

For the realisation of the metal fixed points, the influence of impurities on the temperature and shape of the melting or freezing curves causes usually the main uncertainty component. The influence is of the order of 0.5 mK per p.p.m., which is about one order of magnitude larger compared with the effects on the triple-point temperatures of the cryogenic gases. Thus, metal samples of a purity of 99.9999% or better have to be used if the desired uncertainty is smaller than 1 mK. This demand on the purity must be fulfilled for the overall impurity content. Since in the purity certificates of the suppliers usually only analysis results for a few impurity elements are given, the realisation of the metal fixed points at the highest level requires to analyse in detail the individual fixed-point samples applying appropriate techniques. The situation is even more complicated because the effect of the impurities depends on their crystallographic behaviour at low concentrations in the host material [64]. Thus, for estimating the uncertainty component caused by impurities, reliably, specific methods must be applied and three methods are proposed in ref. [55] and are recommended in ref. [65]. The state-of-the-art level of accuracy of the fixed-point realisation for contact thermometry can be proved by the results of key comparisons [66-68]. The standard deviations of the results range from 0.2 mK for gallium to 4 mK for silver, *i.e.* the estimates given in ref. [52] of at most 1 mK for silver are too optimistic. Detailed uncertainty budgets, which are in accordance with the results of the key comparisons, are for instance presented in ref. [55], and recommendations for realising the metal fixed points in ref. [63].

8·3. *Eutectic points of metal-carbon compositions.* – At extremely high temperatures, most materials become so reactive that the choice of a crucible material, which does not affect the fixed-point material, becomes severely limiting. There are many types of metals with melting temperature higher than copper, the highest fixed point of the ITS-90, but none had been used successfully as practical fixed point because contamination by the graphite crucible would cause depression of the melting and freezing temperature. The use of other crucible material, such as alumina, had been successful only for a short-term experiment for Pd and Pt freezing point measurement, with a temperature of 1555 °C and 1768 °C, respectively. Melting and freezing plateaus have been observed for alumina (melting temperature 2053 °C) in a tungsten crucible [69], but the low emissivity of tungsten is still a problem for radiation thermometry.

The use of metal-carbon eutectics as the fixed-point material offers a solution to the combined problem of the graphite crucible and the fixed-point materials. Graphite, the crucible material, is already a component of the fixed-point alloy, and cannot cause contamination. As shown in the binary phase diagram for Ni-C in fig. 11 [70], the liquid phase takes its lowest temperature at the eutectic point, and there cannot be any further depression of the freezing point by carbon contamination. The molten metal, at slightly higher temperature than the eutectic point, would always be slightly richer in

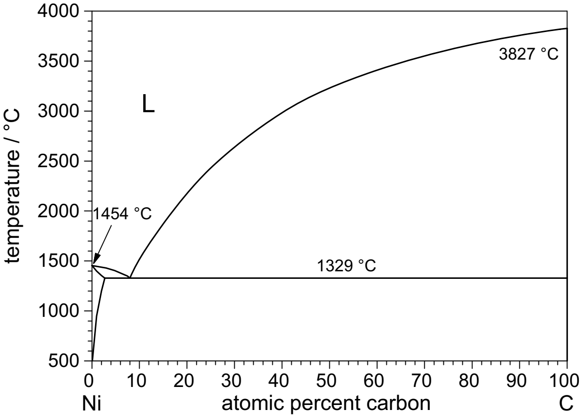


Fig. 11. – Binary phase diagram of the eutectic system Ni-C, L denotes the liquid phase.

C content than the eutectic. However, solidification of graphite during cooling would reduce the carbon content and the molten metal would reach eutectic composition when freezing commences. Thus, reproducible plateaus are observed. There is no way to adjust the composition ratio because carbon could endlessly be supplied from the crucible. Furthermore, graphite crucibles form blackbodies of high emissivity suitable for radiation thermometer calibration, but also there is no concern about reaction with the surrounding furnace material, which is also graphite.

Possible high-temperature fixed points using metal-carbon eutectics are listed in fig. 12. The melting temperatures are conveniently spaced out of the temperature range from the copper point up to 2500 °C. Metal-carbon eutectics, with no special precautions or procedures and with a wide variation in the cooling rates for the preceding freezes, have shown melting plateau repeatability of 20 mK [71] in some cases. On the other hand, the reproducibility between different cells with different crucible designs, material sources, and material purity can be considerably worse [72] and needs further investigation. Evidently, material purity plays the major role [71] as the metal-carbon eutectics appear to be more susceptible to impurities than pure-metal fixed points.

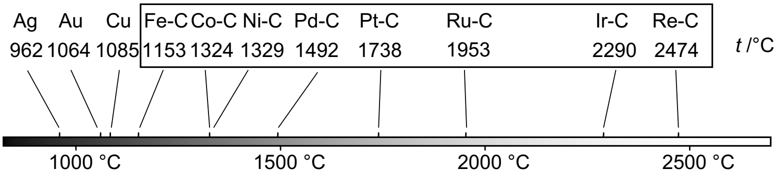


Fig. 12. – Approximate melting temperatures of metal-carbon eutectic compositions. The three highest defining fixed points of the ITS-90 are shown on the left, outside the box.

The eight metals in the box of fig. 12 are the ones that show the simplest binary phase diagrams with carbon and are in consideration to become supplementary fixed points of the ITS-90. For most of the others, carbides are formed, which makes the phase diagram complex. For instance, Ti gives TiC, with a melting temperature of 3067 °C and forms an eutectic with Ti and with graphite. The possible material combinations are Ti-TiC eutectic in a Ti (or TiC) crucible, and TiC-C eutectic in a C (or TiC) crucible. The former is not easy to realise as reaction between the crucible material and the furnace material is hard to prevent. On the other hand, TiC-C eutectic in a graphite crucible requires no additional technical improvements and is of more interest because the eutectic point has a temperature above the 3000 K mark. Other candidates are ZrC-C eutectic at 2927 °C and HfC-C eutectic at 3180 °C [73].

In the ITS-90, the defining fixed point with the highest temperature is the copper point. Above this, no fixed point is available that can be widely used as a practical calibration device. High-temperature fixed points using metal carbon eutectics may change this situation. The eutectics can be held in graphite crucibles, which makes the cells practical devices for use at very high temperatures. These fixed points can benefit the future high-temperature standards in various ways. They may play the role of transfer standards replacing the standard tungsten strip lamps in thermometry or standard incandescent lamps in source-based radiometry. The temperature scale may be realised with much smaller uncertainty by interpolation at a selected number of these fixed points [74]. Extension of this technique to above 3000 K is possible with metal carbide-carbon eutectics [73]. However, the temperature values to be assigned to these new fixed points have to be determined beforehand with sufficiently low uncertainty by primary thermometry.

9. – New definition of the kelvin

Testing experimentally the fundamental laws of physics means in practice the precise determination of the fundamental constants appearing in the laws. Since they are ultimately related to the physical units the precise experimental realisation of the latter is an unavoidable prerequisite for the progress in the development of our physical understanding of nature. Precision experiments relating the (space- and time-independent) fundamental constants to the system of units guarantee stability and reproducibility. The essence of current activities is that prototypes, which may vary uncontrollably with time and location, are replaced by abstract experimental prescriptions that relate the units to the constants. This ensures that the requirement of invariance with space and time is fulfilled. This approach is shown for the definition of the kelvin and the Boltzmann constant. The unit of temperature T, the kelvin, is presently defined by the temperature of the TPW. Thus, the kelvin is linked to a material property. Instead, it would be advantageous to proceed in the same way as with other units: to relate the unit to a fundamental constant and fix its value. By this, no temperature value and no measurement method would be favoured. For the kelvin, the corresponding constant is the Boltzmann constant k, because temperature always appears as thermal energy kT in fundamental laws of physics.

As shown by the recently published CODATA values for fundamental physical constants [50], the 2002 recommended value of the Boltzmann constant k with $u_{\mathrm{r}}(k) = 1.8 \cdot 10^{-6}$ is to a very large extent determined by the NIST result [23] and, therefore, is not yet regarded as sufficiently corroborated to replace the present definition of the kelvin. The important point to note here is that the measurement uncertainty of any value of k would be transferred to the value of T_{TPW}, if that k value were taken to be the exact value of the Boltzmann constant and used to define the kelvin. Hence, if the 2002 CODATA recommended value would be fixed as the exact value of k tomorrow, the best estimate of T_{TPW} would still be 273.16 K. However, this value would no longer be exact (as it is now, as a result of the current definition of the kelvin), but would become uncertain by $u_{\mathrm{r}}(T_{\mathrm{TPW}}) = 1.8 \cdot 10^{-6}$, which corresponds to 0.49 mK. The issue to be decided by the thermometry community is whether or not this uncertainty is acceptable, and if not, how small an uncertainty is required.

Another primary thermometer applicable for determining k is based on the Stefan-Boltzmann law and measures the total radiation without spectral selection (see sect. **5**). This method has been developed at NPL [37] and gave a standard-uncertainty estimate for k of $u_{\mathrm{r}} = 3.2 \cdot 10^{-5}$. A project proposal of NPL anticipates that the new absolute radiation thermometer of NPL can be operated with sufficiently small uncertainty and contribute to an improved value of the Boltzmann constant as well [39].

Doppler-broadening thermometry utilizing radiation measurements has only recently been proposed for the purpose of determining the Boltzmann constant [49] and is presently under investigation at the University Paris North, France, with respect to the uncertainty that can possibly be achieved. Though it is a standard means of diagnostics for high-temperature plasmas [48], this method of temperature measurement is less well known in other fields of thermometry and so is described in some detail in sect. **7**. In a joint project of the universities of Milan and Naples [75], Italy, a similar experiment is under development. They will determine the Doppler broadening of an absorption line of water vapour probed by a diode laser based spectrometer system in the near infrared.

In fig. 13, the history of the determinations of the gas constant $R = kN_{\mathrm{A}}$ is shown. In 1972 Batuecas [16] reviewed the last determinations based on absolute CVGT. The determination by AGT of Colclough *et al.* [21] is described in subsect. **3**˙2, the determination by total radiation thermometry (TRT) of Quinn and Martin [37] in sect. **5**. Considering the aforementioned uncertainty estimates, the two most promising methods for reduction of the uncertainty of k currently are dielectric-constant gas thermometry (DCGT, subsect. **3**˙3) and acoustic gas thermometry (AGT, subsect. **3**˙2). It seems to be possible that the DCGT method at PTB or the AGT work of different groups will have been advanced so far by the end of 2009, that they can contribute to an improved value of k or R with similar relative uncertainty as that obtained by Moldover *et al.* [23] with the AGT in 1988. Thus, an improved value of the Boltzmann constant proposed for the definition of the kelvin would ideally have been determined by at least these two fundamentally different methods and be corroborated by other —preferably optical— measurements with larger uncertainty. We shall assume here that the experiments currently underway to measure R or k [76] will achieve a relative standard uncertainty by

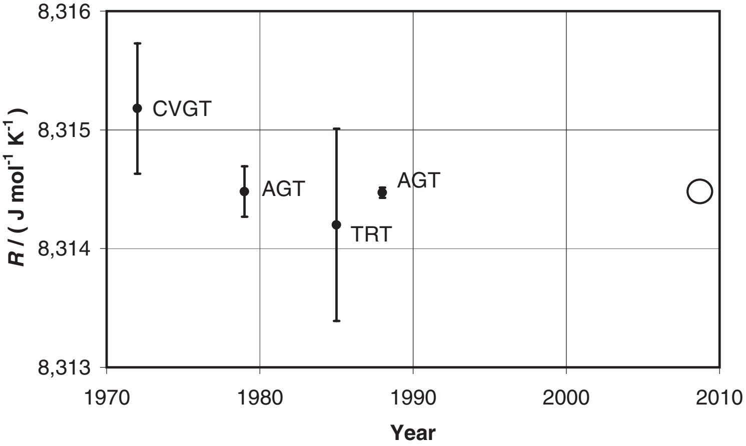


Fig. 13. – History of determinations of the gas constant $R = kN_A$ and prospects for the near future. For clear visibility, the error bars indicate the uncertainties at the confidence level of 99%.

the end of 2009, which is a factor of about two smaller than the current u_r of approximately $2 \cdot 10^{-6}$, so that $u_\mathrm{r}(T_\mathrm{TPW})$ will be reduced to about $1 \cdot 10^{-6}$, corresponding to about 0.25 mK, and that this will be small enough for the redefinition of the kelvin to be adopted by the 24th General Conference on Weights and Measures (CGPM) in 2011.

In the discussion about the new definition of the kelvin, it should also be recognised that the "practical" International Temperature Scale of 1990, ITS-90, is a defined temperature scale which assigns an exact temperature value T_{90} to each defining fixed point. Hence, the ITS-90 value of the TPW temperature will remain 273.16 K, that is, $T_\text{TPW-90} = 273.16\,\mathrm{K}$ exactly. The value and uncertainty of T_TPW would only need to be taken into account if for some critical reason one has to know how well the thermodynamic temperature scale is represented by the ITS-90 at a particular temperature or in a particular temperature range. In fact, although the consistency of T_TPW as realised by different TPW reference cells can be as low as 50 μK (and even less if the isotopic composition of the water used is taken into account), the uncertainties of the thermodynamic temperatures of all other defining fixed points, which are the basis for all practical thermometry, are significantly larger [77]. In contrast to other units, the uncertainty of the realisation of the kelvin varies greatly with temperature: at 1300 K, for instance, the uncertainty is roughly 100 times greater than at T_TPW. Hence, the fact that T_TPW will not be exactly known but have a standard uncertainty of 0.25 mK will have negligible practical consequences.

To put the new definition of the kelvin into practice, a *mise-en-pratique* has already been recommended to the CIPM by the CCT [78]. The *mise-en-pratique* will allow direct determination of thermodynamic temperatures particularly at temperatures far away from the triple point of water in parallel to the realisation described in the International

Temperature Scale. In the high-temperature range this will considerably reduce the uncertainty of the realisation of the kelvin for many purposes for which the need to refer back to the TPW is anomalous, such as radiation thermometry.

In ref. [79], two alternative new definitions are suggested for each of the units kilogram, ampere, kelvin and mole to be chosen by the CGPM in 2011. The first type explicitly defines a unit in terms of a particular quantity of the same kind as the unit and, through a simple relationship implied by the definition itself or one or more basic laws of physics, implicitly fixes the value of a fundamental constant; we call these "explicit-unit definitions". The second type explicitly fixes the value of a fundamental constant and, through a simple relationship implied by the definition itself or one or more laws of physics, implicitly defines a unit; we call these "explicit-constant definitions". It should be understood that the alternative definitions for the same unit are in fact equivalent; they are only different ways of stating the same definition, and in no way should the choice of words be regarded as final at this stage. The "explicit-unit definition" for the kelvin, as proposed for the first time in [15], could be as simple as "the kelvin is the change of thermodynamic temperature that results in a change of the thermal energy kT by exactly $1.3806505 \cdot 10^{-23}$ joule." This definition is comparable with the current definitions of other base units. The "explicit-constant definition" has no evident direct physical meaning and reads "the kelvin, unit of thermodynamic temperature, is such that the Boltzmann constant is exactly $1.3806505 \cdot 10^{-23}$ joule per kelvin."

REFERENCES

Below, *Temperature Its Measurement and Control in Science and Industry* is abbreviated as *TMCSI* and *International Symposium on Temperature and Thermal Measurements in Industry and Science (TEMPMEKO)* as *ISTIS*.

[1] Rusby R. L., Moldover M. R., Fischer J., White D. R., Steur P. P. M., Hudson R. P., Durieux M. and Hill K. D., Working Group 4 report to CCT May 2005, *BIPM Com. Cons. Thermométrie* **23** (Document CCT/05-19) 2005.

[2] Steele A. G. and Rowell N. L., *ISTIS*, edited by Fellmuth B., Seidel J. and Scholz G., Vol. **8** (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 383-388.

[3] Fox N. P., *ISTIS*, Vol. **8** edited by Fellmuth B., Seidel J. and Scholz G., (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 27-35.

[4] Allen D. W., Saunders R. D., Johnson B. C., Gibson C. E. and Yoon H. W., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 577-582.

[5] Taubert D. R., Hartmann J., Hollandt J. and Fischer J., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 7-12.

[6] White D. R., Mason R. S. and Saunders P., *ISTIS*, Vol. **8**, edited by Fellmuth B., Seidel J. and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 129-34.

[7] Benz S. P., Martinis J. M., Nam S. W., Tew W. L. and White D. R., *ISTIS*, Vol. **8**, edited by Fellmuth B., Seidel J. and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 37-44.

[8] Edler F., Kühne M. and Tegeler E., *Metrologia*, **41** (2004) 47.

[9] De Podesta M. and Edwards G., *ISTIS*, Vol. **9**, edited by Zvizdic D. (Laboratory for Process Measurement, Faculty of Mechanical Engineering and Naval Architecture, Zagreb) 2005, pp. 85-90.

[10] Fellmuth B., Fischer J., Gaiser C. and Haft N., *ISTIS*, Vol. **9**, edited by Zvizdic D. (Laboratory for Process Measurement, Faculty of Mechanical Engineering and Naval Architecture, Zagreb) 2005, pp. 73-78.

[11] Pitre L., Moldover M. R. and Tew W. L., *Metrologia*, **43** (2006) 142.

[12] Benedetto G., Gavioso R. M., Spagnolo R., Marcarino P. and Merlone A., *Metrologia*, **41** (2004) 74.

[13] Ripple D. C., Defibaugh D. R., Moldover M. R. and Strouse G. F., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 25-30.

[14] Ewing M. R. and Trusler J. P. M., *J. Chem. Thermodyn.*, **32** (2000) 1229.

[15] Fischer J., Fellmuth B., Seidel J. and Buck W., *ISTIS*, Vol. **9**, edited by Zvizdic D. (Laboratory for Process Measurement, Faculty of Mechanical Engineering and Naval Architecture, Zagreb) 2005, pp. 12-22.

[16] Colclough A. R., *Gas constant, X-ray interferometry, nuclidic masses, other constants, and uncertainty assignments*, in *Precision Measurement and Fundamental Constants II, NBS Special Publication **617***, edited by Taylor B. N. and Phillips W. D. (1984), pp. 263-75.

[17] Guildner L. A. and Edsinger R. E., *J. Res. Natl. Bur. Stand. Sect. A*, **80** (1976) 703.

[18] Edsinger R. E. and Schooley J. F., *Metrologia*, **26** (1989) 95.

[19] Quinn T. J., *Temperature*, 2nd edition, *Monographs in Physical Measurement* (Academic Press, London) 1990.

[20] Pavese F. and Molinar G., *Modern Gas-Based Temperature and Pressure Measurements* (Plenum Press, New York and London) 1992.

[21] Colclough A. R., *Proc. R. Soc. London, Ser. A*, **365** (1979) 349.

[22] Colclough A. R., Quinn T. J. and Chandler T. R. D., *Proc. R. Soc. London, Ser. A*, **368** (1979) 125.

[23] Moldover M. R., Trusler J. P. M., Edwards T. J., Mehl T. J. and Davis R. S., *J. Res. Natl. Bur. Stand.*, **93** (1988) 85.

[24] Moldover M. R. and Trusler J. P. M., *Metrologia*, **25** (1988) 165.

[25] Moldover M. R., Boyes S. J., Meyer C. W. and Goodwin A. R. H., *J. Res. Natl. Inst. Stand. Technol.*, **104** (1999) 11.

[26] Strouse G. F., Defibaugh D. R., Moldover M. R. and Ripple D. C., *TMCSI*, Vol **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 31-36.

[27] Łach G., Jeziorski B. and Szalewicz K., *Phys. Rev. Lett.*, **92** (2004) 233001.

[28] Gugan D. and Michel G. W., *Metrologia*, **16** (1980) 149.

[29] White M. P. and Gugan D., *Metrologia*, **29** (1992) 37.

[30] Luther H., Grohmann K. and Fellmuth B., *Metrologia*, **33** (1996) 341.

[31] Moldover M. R., *J. Res. NIST*, **103** (1998) 167.

[32] Nyquist H., *Phys. Rev.*, **32** (1928) 110.

[33] Crovini L. and Actis A., *Metrologia*, **14** (1978) 69.

[34] White D. R., Mason R. S. and Saunders P., *ISTIS*, Vol. **8**, edited by Fellmuth B., Seidel J. and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 129-134.

[35] Brixy H., Hecker R., Oehmen J., Rittinghaus K. F., Setiawan W. and Zimmermann E., *TMCSI*, Vol. **6**, edited by Schooley J. F. (American Institute of Physics, New York) 1992, pp. 993-996.

[36] Nam S. W., Benz S. P., Martinis J. M., Dresselhaus P., Tew W. L. and White D. R., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C., (American Institute of Physics, Melville, NY) 2003, pp. 37-42.

[37] Quinn T. J. and Martin J. E., *Philos. Trans. R. Soc. London A*, **316** (1985) 85.

[38] Martin J. E., Quinn T. J. and Chu B., *Metrologia*, **25** (1988) 107.

[39] Martin J. E. and Haycocks P. R., *Metrologia*, **35** (1998) 229.

[40] Jung H. J., *Metrologia*, **20** (1984) 67.

[41] Jung H. J., *Metrologia*, **23** (1986) 19.

[42] Fischer J. and Jung H. J., *Metrologia*, **26** (1989) 245.

[43] Bonhoure J. and Pello R., *BIPM Com. Cons. Thermometrie*, **15** (Document CCT/84-21) 1984.

[44] Coates P. B., Andrews J. W. and Chattle M. V., *Metrologia*, **21** (1985) 31.

[45] Rusby R. L., Hudson R. P., Durieux M., Schooley J. F., Steur P. P. M. and Swenson C. A., *Metrologia*, **28** (1991) 9.

[46] Stock M., Fischer J., Friedrich R., Jung H. J., Werner L. and Wende B., *ISTIS*, **6** (1996) 19-24.

[47] Hartmann J., Taubert D. R. and Fischer J., *ISTIS*, Vol. **8**, edited by Fellmuth B., Seidel J and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 377-82.

[48] Griem H. R., *Plasma Spectroscopy* (McGraw-Hill, New York) 1964.

[49] Daussy C., Briaudeau S., Guinet M., Amy-Klein A., Hermier Y., Bordé C. J. and Chardonnet C., *Proceedings of the 17th International Conference on Laser Spectroscopy*, edited by Hinds E., Ferguson A. and Riis E. (World Scientific) 2005, pp. 104-111.

[50] Mohr P. J. and Taylor B. N., *Rev. Mod. Phys.*, **77** (2005) 1.

[51] Preston-Thomas H., *Metrologia*, **27** (1990) 3; 107.

[52] Preston-Thomas H., Bloembergen P. and Quinn T. J., *Supplementary Information for the International Temperature Scale of 1990* (BIPM, Sèvres, Pavillon de Breteuil) 1990.

[53] Rusby R. L., Hudson R. P., Moldover M. R., Fischer J., White D. R., Steur P. P. M., Reesink A. L., Durieux M. and Fogle W. E., 2003 Working Group 4 report to CCT May 2003 *BIPM, Com. Cons. Thermométrie* **22** (Document CCT/03-26) 2003.

[54] Mangum B. W., Bloembergen P., Chattle M. V., Fellmuth B., Marcarino P. and Pokhodun A. I., *Metrologia*, **34** (1997) 427.

[55] Fellmuth B., Fischer J. and Tegeler E., *Uncertainty budgets for characteristics of SPRTs calibrated according to the ITS-90*, *BIPM Com. Cons. Thermométrie* **21** (Document CCT/01-02) 2001.

[56] Crovini L., Mangum B. W., Kemp R. C., Jung H. J., Ling Shankang and Sakurai H., *Metrologia*, **28** (1991) 317.

[57] Nicholas J. V., *ISTIS*, Vol. **7**, edited by Dubbeldam J. F. and de Groot M. J. (IMEKO / NMi Van Swinden Laboratorium, Delft) 1999, pp. 100-105.

[58] Mangum B. W. and Furukawa G. T., *NIST Tech. Note* **1265** (1990).

[59] Bentley R. E. (Editor), *Handbook of temperature measurement* (Springer, Singapore) 1998.

[60] Michalski L., Eckersdorf K., Kucharski J. and McGhee J., *Temperature Measurement*, 2nd edition (John Wiley & Sons Ltd., New York) 2001.

[61] Nicholas J. V. and White D. R., *Traceable Temperatures. An Introduction to Temperature Measurement and Calibration*, 2nd edition (John Wiley & Sons Ltd., New York) 2001.

[62] Bernhard F. (Editor), *Technische Temperaturmessung* (Springer-Verlag, Berlin) 2004.

[63] Mangum B. W., Bloembergen P., Chattle M. V., Fellmuth B., Marcarino P. and Pokhodun A. I., *Optimal realization of the defining fixed points of the ITS-90 that are used for contact thermometry*, *BIPM Com. Cons. Thermométrie*, **20**, (Document CCT/2000-13) 2000.

[64] Mangum B. W., Bloembergen P., Fellmuth B., Marcarino P. and Pokhodun A. I., *On the influence of impurities on fixed-point temperatures*, *BIPM Com. Cons. Thermométrie*, **20** (Document CCT/99-11) 1999.

[65] Ripple D., Fellmuth B., De Groot M., Hermier Y., Hill K. D., Marcarino P., Pokhodun A., Matveyev M. and Bloembergen P., *Methodologies for the estimation of uncertainties and the correction of fixed-point temperatures attributable to the influence of chemical impurities*, *BIPM Com. Cons. Thermométrie*, **23** (Document CCT/05-08) 2005.

[66] Steele A. G., Fellmuth B., Head D. I., Hermier Y., Kang K. H., Steur P. P. M. and Tew W. L., *Metrologia*, **39** (2002) 551.

[67] Mangum B. W., Strouse G. F., Guthrie W. F., Pello R., Stock M., Renaot E., Hermier Y., Bonnier G., Marcarino P., Gam K. S., Kang K. H., Kim Y.-G., Nicholas J. V., White D. R., Dransfield T. D., Duan Y., Qu Y., Connolly J., Rusby R. L., Gray J., Sutton G. J. M., Head D. I., Hill K. D., Steele A., Nara K., Tegeler E., Noatsch U., Heyer D., Fellmuth B., Thiele-Krivoj B., Duris S., Pokhodun A. I., Moiseeva N. P., Ivanova A. G., de Groot M. J. and Dubbeldam J. F., *Metrologia*, **39** (2002) 179.

[68] Nubbemeyer H. G. and Fischer J., *Metrologia*, **39** (2002) Techn. Suppl. 03001.

[69] Sakate H., Sakuma F. and Ono A., *Metrologia*, **32** (1995) 129.

[70] Massalski B. T. (Editor), *Binary Alloy Phase Diagrams*, Vol. **1**, (ASM Int., Materials Park Ohio) 1990.

[71] Sasajima N., Yamada Y., Zailani B. M., Fan K. and Ono A., *ISTIS*, Vol. **8**, edited by Fellmuth B., Seidel J. and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 501-506.

[72] Machin G., Yamada Y., Lowe D., Sasajima N., Sakuma F. and Fan K., *ISTIS*, Vol. **8** edited by Fellmuth B., Seidel J. and Scholz G. (VDE Verlag GmbH, Berlin) ISBN 3-8007-2676-9, 2002, pp. 851-856.

[73] Sasajima N., Yamada Y. and Sakuma F., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 279-284.

[74] Bloembergen P., Yamada Y., Yamamoto N. and Hartmann J., *TMCSI*, Vol. **7**, Editor-in-Chief Ripple D. C. (American Institute of Physics, Melville, NY) 2003, pp. 291-296.

[75] EUROMET project 885 progress report 2006.

[76] Fellmuth B., Gaiser Ch. and Fischer J., *Meas. Sci. Technol.*, **17** (2006) R145.

[77] Fischer J. and Fellmuth B., *Rep. Prog. Phys.*, **68** (2005) 1043.

[78] Recommendation T 3 (2005) to the CIPM: *Creation of a "mise-en-pratique" of the definition of the kelvin*, *BIPM Com. Cons. Thermométrie*, **23** (Document CCT/05-32) 2005.

[79] Mills I. M., Mohr P. J., Quinn T. J., Taylor B. N. and Williams E. R., *Metrologia*, **43** (2006) 227.

Metrological applications of acoustic and microwave resonators

R. M. Gavioso

INRIM - Istituto Nazionale di Ricerca Metrologica, Dipartimento di Acustica
Strada delle Cacce 91, 10135 Torino, Italy

1. – Introduction

A major development of the theory and practice of acoustic and microwave resonators has been achieved in recent years. These scientific instruments may now be used to provide an extremely accurate estimate of the thermodynamic and electrical properties of dilute gases. When combined in a single experiment, the features of a microwave and acoustic resonator prove capable to challenge, in terms of achievable accuracy, the currently adopted primary standards for the physical quantities temperature and pressure.

Within this present frame, a measurement of speed of sound in helium or argon effected at the temperature of the triple point of water candidates as the most promising method for a new determination of the molar gas constant R and the Boltzmann constant k at a 1 ppm level. If this is achieved in a near future, it will support the adoption of a new definition of the unit of temperature, the kelvin, based on an exactly defined value of these constants.

Recently, a significant effort has also been pursued in improving the theoretical calculation of some physical properties of monoatomic gases *ab initio*, with particular attention to helium. As a result of this improvement, the calculated helium properties may now serve as a reference for the purpose of calibrating many different instruments, from primary thermometers to viscometers.

2. – Basic description of some physical properties of dilute monoatomic gases

2˙1. *Thermophysical and transport properties.* – The equation of state and the transport properties of fluids, both in the gaseous or liquid state, are determined by the forces exerted between molecules. In the simplest possible approximation the forces are zero, the molecules have no volume and the ideal gas law $p\rho = RT$ governs the state of the fluid. However, even at the moderate density corresponding to ambient conditions, the deviations from this simple law get to be appreciable. The well physically founded power series expansion known as *virial equation of state*:

$$p/\rho RT = 1 + B(T)\rho + C(T)\rho^2 + \dots \tag{2.1}$$

gives a much more accurate description of the thermodynamic state of the fluid over an extended range of temperature and pressure. Here the functions $B(T), C(T), \dots$ depend only on temperature and take the name of second, third, .. density virial coefficients. Statistical mechanics provides a link between the virial coefficients of different order and the forces exerted within clusters composed by an increasing number of molecules. As an example, the second virial coefficient may be expressed as

$$B(T) = 2\pi N_A \int_0^\infty \left[1 - \exp[-\varphi(r)kT]\right] r^2 \mathrm{d}r, \tag{2.2}$$

where N_A is the Avogadro constant, k the Boltzmann constant and $\varphi(r)$ is the potential energy of interaction between two molecules, which is assumed to be a function of the intermolecular separation r only.

By analogy with the virial eq. (2.1), the speed of sound in a real gas may be represented by a power series of the pressure

$$u^2 = u_0^2\left[1 + (\beta_a/RT)p + (\gamma_a/RT)p^2 + \dots\right], \tag{2.3}$$

where, u_0 is the speed of sound at zero pressure and the *acoustic virial coefficients* $\beta_a(T), \gamma_a(T), \dots$, are related to the corresponding density virials in eq. (2.1) by second order differential equations. For example

$$\beta_a(T) = 2B + 2(\gamma^0 - 1)T\left(\frac{\mathrm{d}B}{\mathrm{d}T}\right) + \frac{(\gamma^0 - 1)^2}{\gamma^0} T^2 \frac{\mathrm{d}^2 B}{\mathrm{d}T^2}\,. \tag{2.4}$$

Accurate measurements of speed of sound as a function of pressure along one isotherm can be analyzed with the expansion (2.3) to determine the first two or three acoustic virial coefficients and u_0. The most relevant metrological applications based on a determination of the speed of sound at zero pressure will be discussed below in sects. **4** and **5**.

If the interaction potential energy is known for a fluid, all the thermodynamic and transport properties of the fluid may be in principle calculated. Quantum mechanics and the fundamental constants allow to obtain $\varphi(r)$ at discrete small values of r and asymptotic forms of $\varphi(r)$ at large r for the "simple" He-He interaction. Some of these results are characterized by a considerable accuracy and provide a reliable estimate of the associated uncertainty. This accuracy may be maintained, with negligible additional computational uncertainty, throughout the complex quantum statistical mechanical calculations which are used to obtain the macroscopic thermophysical properties of interest from the interpolated potential. With this method the function $B(T)$ for helium was recently calculated with a relative uncertainty between 0.4% and 0.2% in the temperature range between 100 K and 300 K [1]. As an example illustrating a practical application of this approach, the viscosity of helium was recently used to improve the accuracy of both the thermal conductivity and the viscosity of argon with relative uncertainties below 0.1% [2].

It is worth noticing that, at least for helium, the recent improvement in the calculation of $B(T)$ and the differential relationship in eq. (2.4) provide an important independent check of the pressure dependence of $u(p)$, thus evidencing possible systematic errors which may affect the experimental apparatus used for speed of sound measurements.

The different methods which might be used to obtain a complete equation of state for the fluid of interest from speed of sound data, when these are combined with other experimental quantities or results from molecular theory, have been reviewed elsewhere [3].

2˙2. *Electrical properties.* – The thermodynamic and electrical properties of a dilute gas are linked together by the Clausius-Mossotti equation. Once more this may be conveniently virially expanded as a function of density:

$$\wp = \frac{1}{\rho}\frac{\varepsilon_{\mathrm{r}}-1}{\varepsilon_{\mathrm{r}}+2} = A_\varepsilon + b_\varepsilon\rho + c_\varepsilon\rho^2 + \dots . \tag{2.5}$$

Here $\wp$ is the polarizability, ε_{r} is the relative dielectric constant of the gas, A_ε is the molar polarizability and the coefficients in the power series: $b_\varepsilon, c_\varepsilon$, .. are the second, third, .. dielectric virial coefficients. For a polar substance, having a permanent dipole moment, eq. (2.5) should be modified to keep into account the temperature dependence of the polarizability, which is expressed by the Debye equation

$$A_\varepsilon(T) = A_\varepsilon + \frac{N_A\mu^2}{9\varepsilon_0 kT}\,, \tag{2.6}$$

where μ is the dipole moment. Equation (2.5) and the virial equation of state (2.1) may be combined to obtain

$$\frac{\varepsilon_{\mathrm{r}}-1}{\varepsilon_{\mathrm{r}}+2} = \frac{A_\varepsilon p}{RT}\frac{\left(1+b_\varepsilon p/RT+\dots\right)}{\left(1+Bp/RT+\dots\right)}\,. \tag{2.7}$$

Equation (2.7) suggests different metrological applications: i) by measuring $\varepsilon_{\mathrm{r}}(p,T)$ and trusting the pressure instrumentation one may realize a primary dielectric constant gas thermometer (DGCT); ii) as discussed in more detail in sect. **5**, the same measurement of $\varepsilon_\mathrm{r}(p,T)$ using standard calibrated thermometers may realize a primary pressure standard; iii) finally, if both p and T are measured, the measured values for ε_r may be compared with *ab initio* calculations from theory. As the result of a substantial effort pursued in the calculation of A_ε for helium, which now keeps into account both relativistic and QED corrections, its relative uncertainty has been reduced by a factor 50 in the last ten years, leading to the remarkable result: $A_\varepsilon = (0.5172542 \pm 0.0000001)\,\mathrm{cm}^3{\cdot}\mathrm{mol}^{-1}$ [4]. To the best of current knowledge the values of the second and third dielectric virial coefficients b_ε and c_ε for helium are, respectively, obtained from theory [5] and direct capacitance measurements [6].

3. – Basic theory and operation of acoustic and microwave resonators

Both acoustic and microwave resonators have been widely used in science during the last fifty years. To quote but a few relevant metrological applications which were achieved in the past: accurate measurements of the speed of light in vacuum in a cylindrical cavity by Essen in 1950 [7], and the determination of the gas constant by means of a cylindrical interferometer in 1979 [8]. A major successive development of the ultimate accuracy achievable with these instruments took place around 1985, when Moldover and Mehl worked out a complete theoretical model for an acoustic spherical resonator and showed its agreement with experimental practice at the level of 1 ppm [9]. The same authors later extended their model to describe that the same level of accuracy could be obtained in the description of the electromagnetic field confined within a cavity of spherical geometry and developed applications which combined acoustic and/or microwave measurements for applications ranging from primary thermometry [10] to an original primary pressure standard [11]. As a later further development of these experimental techniques, the use of resonators having an intentionally perturbed geometry from that of an ideal sphere showed that the precision achievable in the determination of microwave resonance frequencies may reach the level of a few ppb [12].

3˙1. *Acoustic and microwave frequencies in a spherical cavity.* – The solution of the wave equation for a steady acoustic pressure field excited in a loss-less gas confined within a spherical cavity of ideal geometry and infinite acoustic impedance are given by

$$f^\mathrm{a}_{ln} = z^\mathrm{a}_{ln}(u/2\pi a), \tag{3.1}$$

where the indexes (l,n) identify the order and symmetry of different normal modes of the cavity, u is the speed of sound, z_{ln} are exact numerical constants and a is the inner cavity radius. Among modes of different symmetry, the $(0,n)$ are named purely radial and benefit of properties which make them particularly suitable for speed of sound measurements. Particularly, they are the only non-degenerate modes and they are characterized by exceptionally high quality factors $Q = f/2g$, where g is the resonance halfwidth.

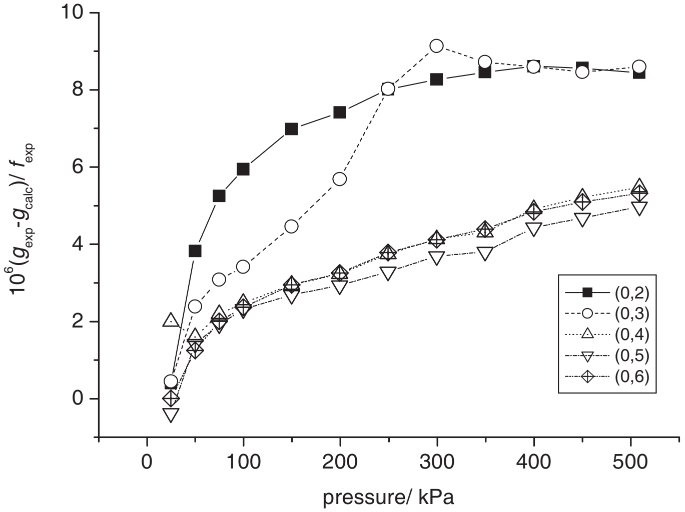


Fig. 1. – Excess acoustic halfwidths ($g_{\rm exp} - g_{\rm calc}$) in a stainless-steel spherical resonator filled with argon in the pressure range 25 to 500 kPa [13].

The electromagnetic resonance frequencies within a perfectly conducting cavity of spherical geometry are

$$f_{ln}^{\rm m} = z_{ln}^{\rm m}(c/2\pi a), \tag{3.2}$$

where c is the speed of light. Among these solutions we distinguish two subsets named TM (transverse magnetic or electric) and TE (transverse electric or magnetic). Differently from the acoustic scalar pressure field, all the solutions in eq. (3.2) are degenerate, the degree of the degeneracy being equal to $(2l + 1)$ and thus at least triple for TM1n and TE1n modes.

The previous ideal models do not take into account the major perturbating effects which take place in a real case and may be considered a valid approximation at the level of a few parts in 10^4. For the acoustic modes the major perturbative effects include: visco-thermal energy losses taking place within a narrow layer close to the inner solid cavity surface; coupling of the gas and shell motion as a function of pressure; the effect of holes machined through the resonator wall. For microwave resonances the only effect which is usually taken into account is the skin effect which is proportional to the resistivity of the metal comprising the inner wall of the cavity. In both cases, an important test of the accuracy of these models in describing the perturbative phenomena which take place within the resonator, consists in comparing the experimentally determined halfwidths with those independently calculated from the model. The results of such a comparison are illustrated in fig. 1 for the first five purely radial modes measured in argon, at 273.16 K in a pressure range between 25 and 500 kPa, within a stainless-steel spherical resonator having an inner diameter of 12 cm.

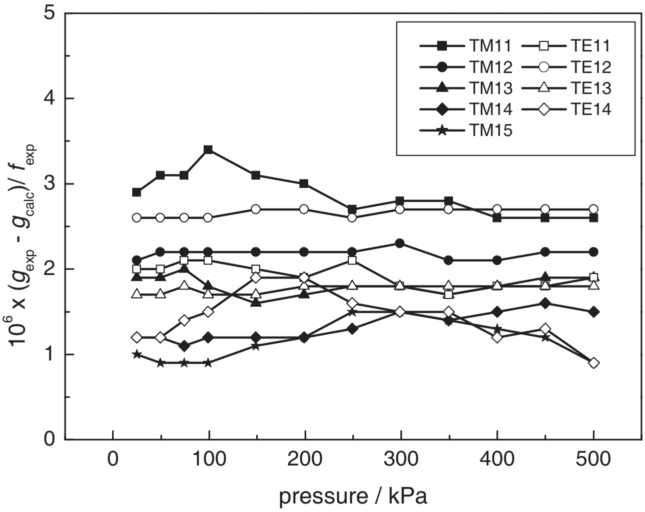


Fig. 2. – Excess electromagnetic halfwidths ($g_{\mathrm{exp}} - g_{\mathrm{calc}}$), scaled by corresponding average frequency, for nine triply degenerate TM1n and TE1n modes in argon [13].

Tipically, the excess acoustic halfwidths scaled by the corresponding frequency amount to a few ppm and approach zero at the lowest pressures, where possible systematic errors would mainly affect an extrapolation of the experimental results towards the zero-pressure limit.

Figure 2 reports a comparison between experimental and calculated halfwidths for nine triply degenerate electromagnetic modes. Also in this case the order of the agreement is usually at the level of a few ppm of the resonance frequency, setting an upper limit to the systematic error which might affect the experimental data.

3˙2. *Effects of perturbed geometry on achievable precision and accuracy.* – Besides the perturbative effects considered in the previous section, in the most common practical case one deals with a resonant cavity affected to some extent by geometrical imperfections caused by fabrication defects which may limit both the precision and accuracy of acoustic and microwave measurements. Perturbation theory has been used to show that the frequency of the acoustic radial modes and the average frequency of microwave multiplets is independent of volume-preserving geometrical deformations of a spherical cavity to first order of the perturbation [14]. For some particular geometries like prolate and oblate spheroids, ellipsoids, cavities comprised of hemispheres having unequal diameters, the effect has been calculated for the radial acoustic modes at the second order of the perturbation [15]. Resonators fabricated with intentionally perturbed geometry may be used in order to increase the precision which can be achieved in the measurement of the average frequency of degenerate microwave modes. As an example of the effectiveness of this strategy, we consider a particular kind of geometrical perturbation which consists in the misalignment of the two hemispheres comprising a spherical cavity. Figure 3

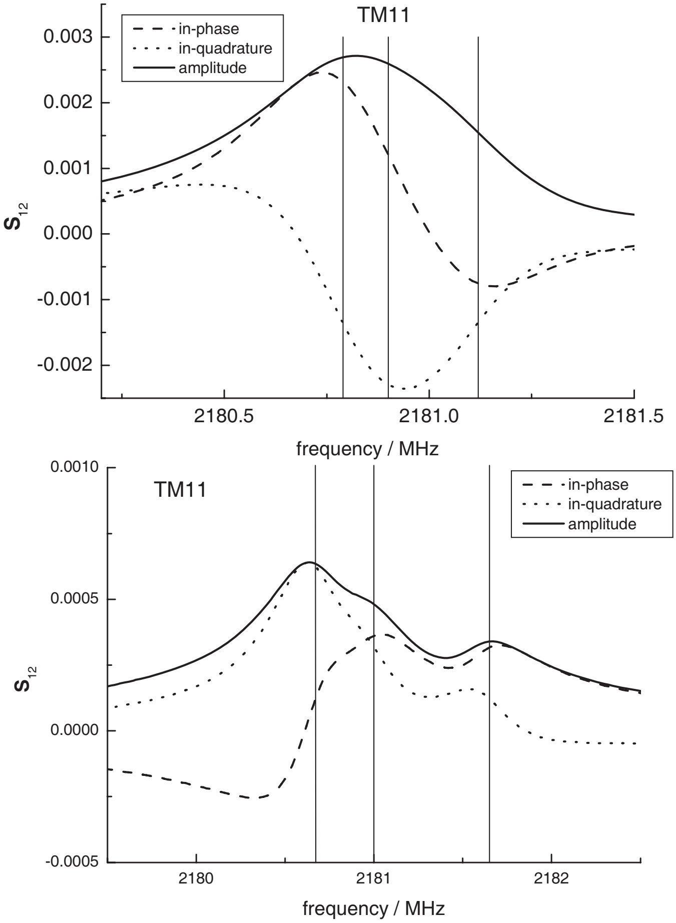


Fig. 3. – Electromagnetic spectrum recorded in the vicinity of the TM11 mode for two different misalignments of the hemispheres comprising a spherical cavity. Top: $\Delta x = 5 \cdot 10^{-4} a$; Bottom: $\Delta x = 1.5 \cdot 10^{-3} a$ [13].

shows the difference in the spectrum as measured in vicinity of the triplet TM11 when the vertical axes of the two hemispheres of a spherical cavity of radius $a = 6\,\mathrm{cm}$ are misaligned by approximately $5 \cdot 10^{-4} a$ and $1.5 \cdot 10^{-3} a$. In both cases the experimental data were recorded using a network analyzer which swept over 201 discrete frequency values spanning a range of 1.5 or 3 MHz.

The data were then fitted with the sum of three complex lorentzian functions plus a linear background for a total of 16 adjustable parameters:

$$u + \mathrm{i}v = \sum_{m=0,\pm1} \frac{2\mathrm{i} f g_{ln}^{m} \mathbf{A}_{ln}^{m}}{\left(\mathbf{F}_{ln}^{m^2} - f^2\right)} + \mathbf{B} + \mathbf{C}\left(\mathbf{F}_{ln}^{m} - f\right). \tag{3.3}$$

In eq. (3.3) $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are complex constants and $\mathbf{F} = f + \mathrm{i}g$ are the complex degenerate

Table I. – *Typical precision achievable in fitting eq.* (3.3) *to experimental data for mode TM11 in two differently perturbed geometries. The absolute uncertainties of the resonance frequencies and the halfwidths of the three components of the mode TM11 are listed.*

	$\Delta x = 5 \cdot 10^{-4} a$			$\Delta x = 1.5 \cdot 10^{-3} a$		
singlet	f/MHz	g/MHz	$\sigma f, g$/MHz	f/MHz	g/MHz	$\sigma f, g$/MHz
TM11(-1)	2180.7878	0.2344	0.1318	2.1806522	2.398	0.0006
TM11(0)	2180.8033	0.2326	0.3093	2.1809653	2.441	0.0017
TM11(+1)	2181.1170	0.2474	0.0014	2.1816131	2.427	0.0008

resonance frequencies of the mode under study. The substantial difference in the precision achievable for the two misalignments is reported in table I.

As evidenced in table I, the ultimate precision achievable in the measurement of the frequency of microwave triplets depends on the separation of their single components and would be favoured by the high-quality factors attainable in cavities fabricated or plated with high-conductivity metals like copper or gold.

4. – Speed of sound as a thermodynamic temperature standard

Recently, a major contribution to primary thermometry came from acoustic measurements. Measurements of speed of sound in argon and helium have been performed by different research groups to determine the thermodynamic temperature spanning the range from 7 to 500 K [10, 16-19]. A promising experimental work is currently underway at NIST aiming to extend the range covered by the acoustic method up to 800 K [20]. In the overlapping temperature ranges these results have been found to be consistent within the remarkably small combined uncertainties. When used to test the accuracy of the currently adopted International Temperature Scale of 1990 (ITS-90), the same results evidence systematic deviations between the Scale and the thermodynamic temperature which are significant even close to the Triple Point of Water (TPW) where the Scale is exactly defined. While the collection of these results might be used for a new revision of the Scale in reason of their high consistency and accuracy, they rather suggest that the best possible advantage to primary thermometry would consist in a new definition of the unit of temperature, the kelvin, based on the adoption of an exact value of the Boltzmann constant k. In the following of this section, after a brief recall of the basic physical relationship between speed of sound and thermodynamic temperature, we illustrate the recent results obtained by primary acoustic thermometers; finally we discuss the perspective of a possible future improvement of an acoustic/microwave determination of the molar gas constant R and the Boltzmann constant k.

4·1. *Relationship between speed of sound and thermodynamic temperature.* – The speed of a longitudinal acoustic wave u propagating through a fluid may be expressed as

$$u^2 = \big(\partial p/\partial \rho\big)_S, \tag{4.1}$$

where p and ρ are the small acoustic perturbations to the equilibrium pressure and density of the fluid and S denotes entropy. Elementary considerations of thermodynamics and kinetic theory allow to obtain from eq. (4.1) and the equation of state of a perfect gas $p = \rho RT$:

$$u^2 = \gamma^0 \frac{kT}{m} = \gamma^0 \frac{RT}{M}\,, \tag{4.2}$$

where γ^0 is the ratio of the ideal-gas heat capacities, m and M are, respectively, the molecular and molar mass and T the thermodynamic temperature. Considering the kinetic energy contained in molecular motion,

$$kT = \frac{1}{3} m v_{\mathrm{rms}}^2, \tag{4.3}$$

where v_{rms}^2 is the mean square molecular speed, it follows that

$$u^2 = \frac{\gamma^0}{3} v_{\mathrm{rms}}^2, \tag{4.4}$$

i.e. the speed of sound and the mean speed of the molecules are of the same order of magnitude.

4·2. *Physical meaning of the Boltzmann constant and primary acoustic thermometry.* – Currently, a higher precision may be achieved in reproducing (with secondary thermometers and fixed points) the thermodynamic states corresponding to particular temperatures than in measuring the statistical quantities, like the mean kinetic energy, which characterize those states. Mainly for this reason, the current definition of temperature is adopted, stating the kelvin to be equal to 1/273.16 of the thermodynamic temperature of the triple point of water T_{w}. Given this definition, the Boltzmann constant serves the purpose of linking mechanical and thermal quantities. Recalling eq. (4.4), the energy E contained in a single mechanical degree of freedom can be expressed as

$$E = \frac{1}{2} k T_{\mathrm{w}}. \tag{4.5}$$

The macroscopic counterpart of the Boltzmann constant, the universal gas constant R is then defined as

$$R = 2EN_{\mathrm{A}}/T_{\mathrm{w}}. \tag{4.6}$$

Thus, an experimental determination of k or R must be performed on a system which is in equilibrium at 273.16 K. Correspondingly, such an experiment would be suitable for the realization of a primary thermometer, once values for k or R have been fixed to be a particular value. For a thermometer to be considered primary, it must be possible to write down an equation of state explicitly, without having to introduce unknown temperature-dependent constants. This is in fact the case for an acoustic thermometer measuring speed of sound in a real gas in the low-pressure limit. Examining eq. (4.2) it is evident that the opportune choice is a monoatomic gas (argon, helium) for which γ^0 has the constant value 5/3 over a very large temperature range. The kelvin thermodynamic temperature T of a gas can then be determined from the zero-pressure limit of the ratio of speed of sound measurements at T and T_w

$$\frac{T}{273.16\,\mathrm{K}} = \lim_{p\to 0}\left(\frac{u^2(p,T)}{u^2(p,T_\mathrm{w})}\right). \tag{4.7}$$

In principle eq. (4.7) could be used to calibrate thermometers directly on the thermodynamic temperature scale. Alternatively, the comparison of "acoustic temperatures" with the indications of standard platinum resistance thermometers can estimate the difference $(T - T_{90})$ as a function of T. It should be stressed that the ratio method represented by eq. (4.7) gives T/T_w without the knowledge of R or γ/M being required, thus eliminating the uncertainties associated with those quantities.

4·3. *Review of recent results obtained by primary acoustic thermometry.* – By combining eqs. (3.1), (3.2) and (4.7) it is evident that the ratio between the acoustic and microwave frequencies measured as a function of temperature and pressure within the same cavity would be independent of the thermal expansion and elastic compression of the cavity dimensions:

$$\frac{T}{273.16\,\mathrm{K}} = \lim_{p\to 0}\left[\left(\frac{f_a(p,T)}{\langle f_\mathrm{m}(p,T)\rangle}\right)^2 \times \left(\frac{\langle f_\mathrm{m}(p,T_\mathrm{w})\rangle}{f_a(p,T_\mathrm{w})}\right)^2\right]. \tag{4.8}$$

The ratio u/c can be determined to an accuracy, mainly limited by the acoustic measurements, of a few parts in 10^6. Would it came possible in the future to improve this accuracy to one part in 10^7, then one could imagine to fix a conventional value for the gas constant and then measure the triple point of water like any other temperature. Starting in 1999, the measurement of ratios such as in eq. (4.8) have altogether determined the differences between the acoustic thermodynamic temperature and ITS-90 which are illustrated in fig. 4.

The extraordinary level of consistency of these results, is evidenced by a comparison with the thermodynamic temperature measurements shown in fig. 5, which were used as input data for the definition of ITS-90 [21]. The relevance of th recent acoustic results supports and encourages a new definition of the kelvin in terms of a fixed value of the Boltzmann constant [22], rather than the revision of the scale. As a major advantage, the

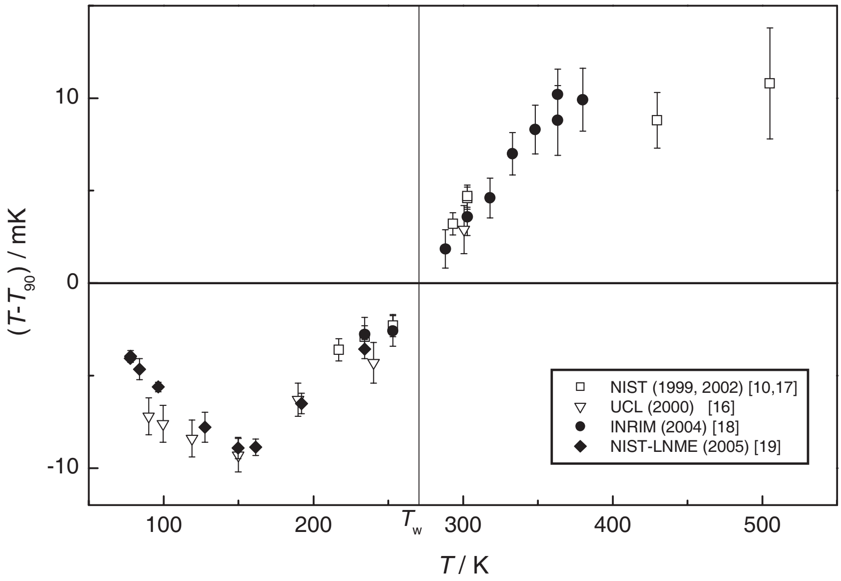


Fig. 4. – Comparison of recent primary acoustic thermometry data [10, 16-19].

accuracy of the practical realization of the unit of temperature would be more uniform over an extended temperature range. When such a definition will finally be adopted the triple point of water would loose its special status and would be assigned an uncertainty. However, the fixed points will maintain their practical value, mainly for standard thermometers calibration purposes.

4˙4. *Redetermination of the molar gas constant R and the Boltzmann constant k.* – The last and most accurate determination of R was achieved at NIST in 1988 [23] by measuring the speed of sound in a sample of monoisotopic ^{40}Ar using a spherical resonator. The volume of the cavity was inferred by careful weighing of the resonator when filled with a standard batch of mercury. The overall relative uncertainty of the measurement was 1.7 ppm. After twenty years this result still represents the state of the art. Particularly, no innovative method has yet been proposed or tested which would significantly improve the accuracy in the determination of the acoustic frequencies of the resonator, which represent the final limit of accuracy achievable. Speculating about a possible experimental scheme that, without being an exact replica of the 1988 experiment, might take to a comparable accuracy in the final result, we consider the possibility to use microwaves for characterizing the dimensions of a resonator. Taking advantage of the exact definition of the speed of light in vacuum, the speed of sound at zero pressure and thus the value of R might be determined from the following expression:

$$R = \frac{M}{\gamma_0 T_\mathrm{w}} {c_0}^2 \lim_{p\to 0} \left(\frac{f'_\mathrm{a}(p)}{\langle f'_\mathrm{m}(p)\rangle} \right)^2 , \tag{4.9}$$

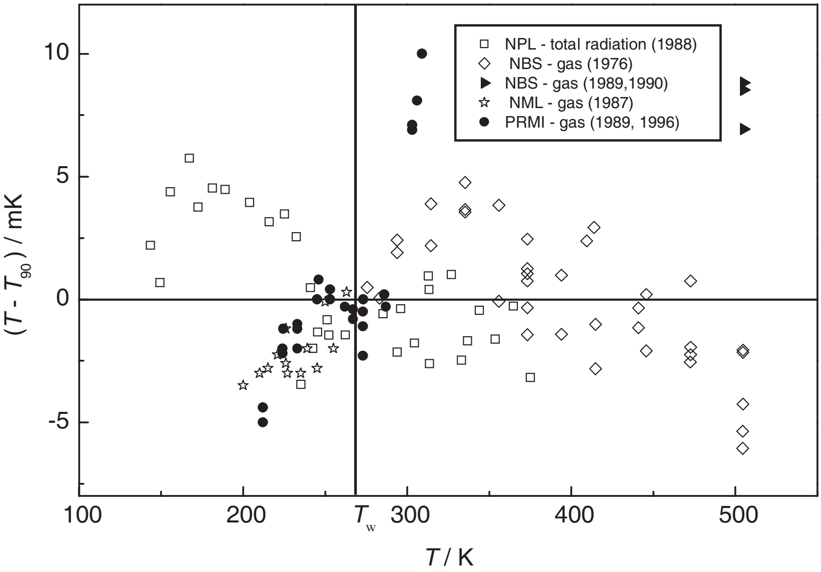


Fig. 5. – Differences $(T - T_{90})$ between the thermodynamic temperature and ITS-90 determined from different primary acoustic thermometers [21].

where both the acoustic $f'_{\rm a}(p)$ and the average microwave frequencies $\langle f'_{\rm m}(p)\rangle$ have been corrected for known perturbations as described in sect. **2**. In this scheme no absolute dimensional characterization is necessary, as implied by the absence of dimensional quantities in eq. (4.9). In practice one would have to measure the frequencies of a certain number of acoustic modes, preferably radial, and a comparable number of microwave modes, preferably triply degenerate TM1n and TE1n. For the purpose of maintaining a high precision in the determination of the average microwave frequencies, a geometrical perturbation of small or predictable effect should be applied to the shape of the resonator; however, the amplitude of the perturbation should be maintained within $1 \cdot 10^{-3}$ of the cavity radius, if its currently unpredictable second-order effects on both the acoustic and electromagnetic field should be maintained within 1 ppm. We might also conjecture that, among the whole set of the possible combinations of acoustic and microwave modes which give the squared frequency ratios in eq. (4.9), those given by a combination of radial acoustic modes $(0, n)$ and TE1n modes might be favoured as the ideal eigenvalues z_{ln} for these modes are the same.

While performing preliminary tests of misalignment as a suitable geometrical perturbation for the purpose of measuring R, at INRIM we used eq. (3.2) to calculate the vacuum radius of a misaligned resonator maintained at 273.16 K using microwave data from nine TM1n and TE1n modes, and obtained the results shown in fig. 6. A systematic, approximately linear frequency dispersion of the radii calculated from different modes is apparent, though remarkably spanning a range of only ±3.5 ppm of the radius.

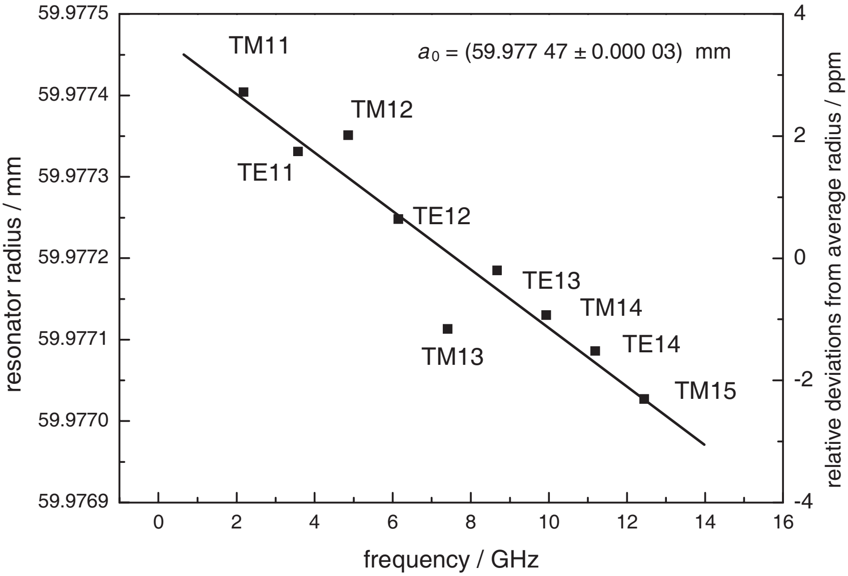


Fig. 6. – Frequency dependence of the radii determined from nine different microwave modes in a misaligned resonator.

When the data are fitted to zero frequency, according to the conjecture that the linear dependence is a consequence of the misalignment, the precision of the fit is approximately 30 nm or 0.5 ppm of the cavity radius.

5. – Speed of light for an atomic pressure standard

At present, the most accurate pressure standards realized and maintained at the major National Metrological Institutes are of two kinds: i) the *mercury manometer*, an apparatus of considerable realization and operation complexity which is best suited for use around or below atmospheric pressure, allows to measure absolute pressures in the order of 100 kPa with a relative standard uncertainty of about $3 \cdot 10^{-6}$; ii) gas-operated *pressure balances*, best suited for operation up to a few tens of MPa, which consist of finely-machined pistons mounted vertically in close-fitting cylinders. The pressure required to support a piston and the associated ring-weights is calculated from the mass of the piston and weights, gravitational acceleration and the effective area of the piston-cylinder assembly. Estimates of the effective area of the piston based on dimensional characterizations and finite-element modelling assess the uncertainty of such instruments to be in the order of 10 ppm below 2 MPa and between 10 to 30 ppm in the range from 2 to 7 MPa.

The high level of precision and accuracy currently achievable in the determination of the resonance frequencies of a quasi-spherical electromagnetic resonator and the remarkably low uncertainty which characterizes the theoretical *ab initio* calculation of

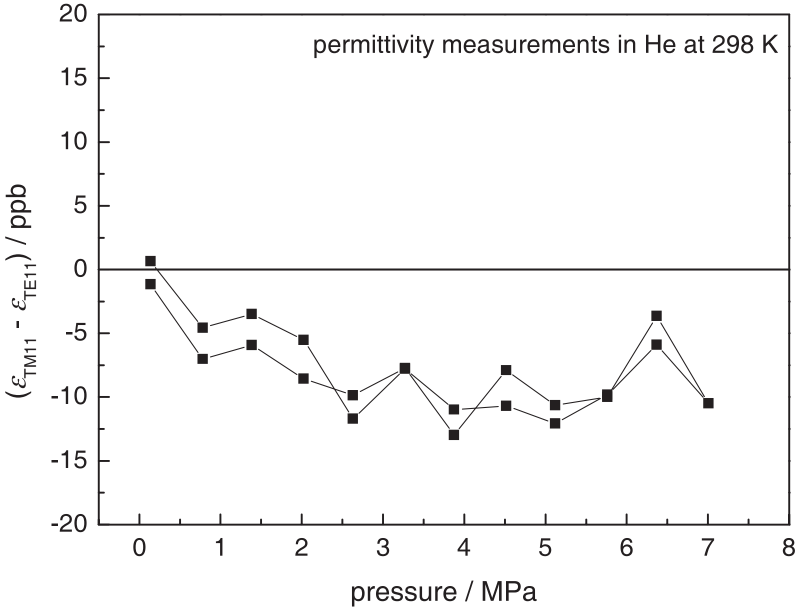


Fig. 7. – Difference in the permittivity measured from two different microwave modes in an electromagnetic resonator.

the electric, transport and thermophysical properties of ^{4}He, suggest and motivate the realization of a new competitive pressure standard based on a measurement of the permittivity of He.

5·1. *Working equations for an electric primary pressure standard.* – Equation (2.7) may be used to express pressure as a function of the thermodynamic and electric properties of a monoatomic gas:

$$p = \frac{\varepsilon_r - 1}{\varepsilon_r + 2}\frac{RT}{A_\varepsilon}\left(1 + \frac{B - b_\varepsilon}{A_\varepsilon}\frac{\varepsilon_r - 1}{\varepsilon_r + 2} + \frac{C - c_\varepsilon - 2Bb_\varepsilon + 2b_\varepsilon^2}{A_\varepsilon}\left(\frac{\varepsilon_r - 1}{\varepsilon_r + 2}\right)^2 + \dots\right). \tag{5.1}$$

For the purpose of realizing a primary pressure standard based on eq. (5.1) and the use of a pure sample of ^{4}He, the quantities A_ε, B and b_ε may be obtained by *ab initio* calculations, C and c_ε from literature measurements and the dielectric constant ε_r would be measured in an instrument maintained at constant temperature. For this electrically based pressure standard to compete with conventional standards, the dielectric constant of helium, which amounts to only ~ 1.004 at 7 MPa, would need to be measured with parts-per-billion precision. The most suitable instruments for such a demanding metrological challenge are quasi-spherical resonators.

By measuring microwave frequencies in vacuum and when the cavity is filled with gas at

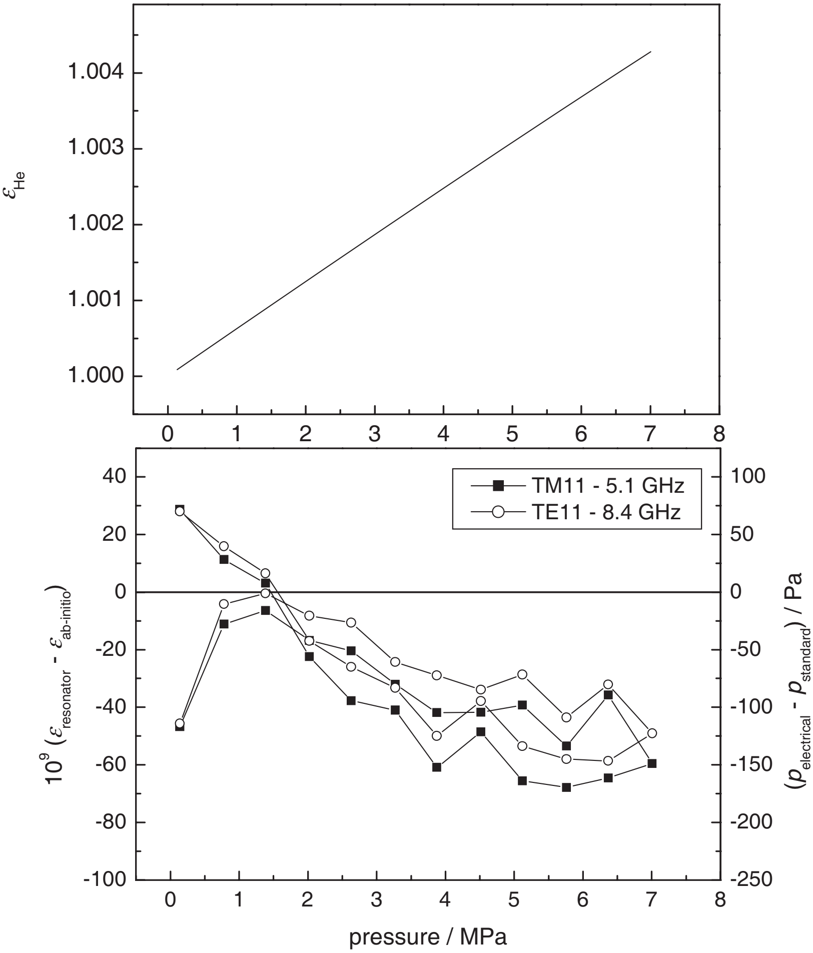


Fig. 8. – Left axis: comparison of experimental and theoretical values of the permittivity of He in the range 0 to 7 MPa; Right axis: comparison of the pressure determined using an electromagnetic resonator and a conventional pressure standard.

different pressures, we may write down the following working equation:

$$\varepsilon_r = \frac{1}{\mu_r}\left(\frac{\langle f'_{\rm m}(0)\rangle}{\langle f'_{\rm m}(p)\rangle(1-(k_T/3)p)}\right)^2, \tag{5.2}$$

where μ_r is the magnetic permeability of helium, which may also be calculated from theory with useful accuracy and k_T is the isothermal compressibility of the solid material comprising the cavity, which may be measured by independent measurements using resonant ultrasound spectroscopy (RUS).

5·2. *Current precision and accuracy achievable with a primary pressure standard.* – As an example of the potential of quasi-spherical resonators in achieving the precision which is needed for a primary pressure standard, fig. 7 shows the difference in the permittivity as measured using two different microwave modes, namely TM11 and TE11, in a 2.5 cm inner radius maraging steel quasi-spherical cavity whose inner surface was electroplated with copper [12]. Remarkably, the results from the two different modes differ at most by 13 ppb, while the gas pressure is varied between 0 and 7 MPa.

In fig. 8 the $\varepsilon_r(p,T)$ results for two helium runs at ambient temperature within the same cavity are compared to *ab initio* values determined using eq. (5.1) and the pressures measured by a conventional standard (piston gauge). It should be noted that the systematic negative difference displayed on the graph would not be accounted for by the presence of impurities in the sample of helium.

Alternatively, by trusting the results of the *ab initio* calculated properties of helium, the same fig. 8 itself, displays the difference in Pa between a conventional pressure standard calibrated at NIST and the proposed electrical standard; the pressure deviations which are shown on the right axis of fig. 8, amount to approximately (150 ± 230) Pa at 7 MPa or fractionally $(22 \pm 35) \cdot 10^{-6}$ when all the contributions from the calculated properties and the experimental determination of $\varepsilon_r(p,T)$ are taken into account.

REFERENCES

[1] Hurly J. J. and Moldover M. R., *J. Natl. Bur. Stand.*, **105** (2000) 667.

[2] May E. F., Moldover M. R., Berg R. F. and Hurly J. J., *Metrologia*, **43** (2006) 247.

[3] Trusler J. P. M., *Physical Acoustics and Metrology of Fluids* (Adam-Hilger, Bristol) 1991, chapt. 1-4.

[4] Lach G., Jeziorski B. and Szalewicz K., *Phys. Rev. Lett.*, **92** (2004) 233001.

[5] Rizzo A., Hattig G., Fernandez B. and Koch H., *J. Chem. Phys.*, **117** (2002) 2609.

[6] White M. P. and Gugan D., *Metrologia*, **29** (1973) 37.

[7] Froome K. D. and Essen L., *The velocity of light and radio waves* (Academic, New York) 1969.

[8] Colclough A. R., Quinn T. J. and Chandler T. R. D., *Philos. Trans. R. Soc. London*, A **368** (1979) 125.

[9] Moldover M. R., Mehl J. B. and Greenspan M., *J. Acoust. Soc. Am.*, **79** (1986) 253.

[10] Moldover M. R., Boyes S. J., Meyer C. W. and Goodwin A. R. H., *J. Natl. Bur. Stand.*, **104** (1999) 11.

[11] May E. F., Pitre L., Mehl J. B., Moldover M. R. and Schmidt J. W., *Rev. Sci. Instrum.*, **75** (2004) 3307.

[12] Gavioso R. M., May E. F., Schmidt J. W., Moldover M. R. and Wang Y., *Proceedings of the 2006 Conference on Precision Electromagnetic Mesurements*, edit by Levi F. and Pisani M. (CLUT, Torino) 2006, pp. 30-31.

[13] Gavioso R. M., Benedetto G., Giuliano Albo P. A. and Spagnolo R., unpublished results.

[14] Mehl J. B., *J. Acoust. Soc. Am.*, **71** (1982) 1109.

[15] Mehl J. B., *J. Acoust. Soc. Am.*, **79** (1986) 278.

[16] Ewing M. B. and Trusler J. P. M., *J. Chem. Thermodynamics*, **32** (2000) 1229.

[17] STROUSE G. F., DEFIBAUGH D. F., MOLDOVER M. R. and RIPPLE D. C., *2003 Temperature: Its Measurement and Control in Science and Industry, 8th International Temperature Symposium*, Chicago (IL) (edited by RIPPLE D. C.) vol. **7**, 2003, pp. 31-36; (*AIP Conf. Proc.* **684**) (2002).

[18] BENEDETTO G., GAVIOSO R. M., SPAGNOLO R., MARCARINO P. A. and MERLONE A., *Metrologia*, **41** (2004) 74.

[19] PITRE L., MOLDOVER M. R. and TEW W. L., *Metrologia*, **43** (2006) 142.

[20] RIPPLE D. C., DEFIBAUGH D. F., MOLDOVER M. R. and STROUSE G. F., and *2003 Temperature: Its Measurement and Control in Science and Industry, 8th International Temperature Symposium*, Chicago (IL) (edited by RIPPLE D. C.) vol. **7**, 2003, pp. 25-30; (*AIP Conf. Proc.* **684**) (2002).

[21] PRESTON-THOMAS H., *Metrologia*, **27** (1990) 3.

[22] MILLS I. M., MOHR P. J., QUINN T. J., TAYLOR B. N. and WILLIAMS E. R., *Metrologia*, **43** (2006) 227.

[23] MOLDOVER M. R., TRUSLER J. P. M., EDWARDS T. J., MEHL J. B. and DAVIS R. S., *Phys. Rev. Lett.*, **60** (1988) 249.

This page intentionally left blank

Mass metrology: Underlying assumptions, present best practice, new frontiers

R. S. Davis

Bureau International des Poids et Mesures, Pavillon de Breteuil, 92312 Sèvres Cedex, France

1. – Introduction

It is usually taken for granted that there is a well-defined physical quantity called mass that can be measured in terms of an internationally recognized unit known as the kilogram. However, our concept of mass has been evolving during the past 100 years, starting with the notions of special and general relativity, continuing with quantum electrodynamics, up to the general acceptance of quantum chromodynamics. The definition of the kilogram predates the earliest of these developments and thus there are strong intellectual arguments and growing practical arguments to change it. Yet the work of mass metrologists continues, being influenced by remarkable technological advances. This paper attempts to make sense of the present situation and to describe the science of mass metrology in the early 21st century.

We begin with some modern notions of mass but our goal is to explain why mass metrologists are generally content with Newtonian physics. We continue with a discussion of the present definition of the kilogram, its strengths and weaknesses, and what is required of a new definition reflecting our current understanding of physics. The central part of the paper discusses the determination of the mass of an unknown object in principle, the effective use of modern balances, and the specifications of commercially available mass standards.

The intersection of mass metrology with the burgeoning developments of micro- and nano-electromechanical systems is a fascinating topic and one that will no-doubt see great technological advances in the near future. The short discussion given at the end of the paper is intended to stimulate the reader's interest in this field.

2. – What do mass metrologists measure?

2˙1. *The relation between mass metrology and contemporary physics.* – In the following sections, we will find that 19th century physics (Newton, Maxwell, etc.) is largely sufficient to describe what mass metrologists measure. In particular, we assume that mass is a conserved quantity (the conservation of mass is sometimes referred to as Newton's "zeroth law"). Nevertheless, we know that mass is not conserved in general and that the discourse of contemporary physics is much more likely to be in terms of energy and momentum rather than force, mass and acceleration. We now discuss briefly why mass metrology is largely immune to modern notions of physics and where we can expect to encounter trouble by ignoring the physics of the last century.

These subjects have been taken up by Wilczek in a series of articles published in *Physics Today.* Two articles discuss the nature of mass [1,2] and the remaining three [3-5] deconstruct the familiar $F = ma$, concluding that while this equation does not express an "ultimate truth", it is nevertheless "extraordinarily useful". A large part of its utility lies in the fact that the zeroth law is an excellent approximation in many domains, including chemistry and mass metrology. Why this is so, can be argued from basic principles of Quantum Electrodynamics (QED) and Quantum Chromodynamics (QCD). Very briefly:

i) Most of the mass of atoms is in their nuclei and the nuclei that we normally deal with have negligible decay rates. Their stability is understood. Going back one step, QCD attributes the rest mass of nucleons to E/c^2, where E is the energy of massless gluons and quarks [1].

ii) The nuclei that we deal with are in their ground state. That is, there is a large energy gap to the first nuclear excited state.

iii) Electrons, which make up a small but significant fraction of atomic masses, do not decay nor do they have internal structure. The electron rest mass, m_{e}, is a "universal constant" [4].

iv) Chemical reactions involve *outer* electrons, which can be shown to have non-relativistic velocities of order αc [6], where α is the fine-structure constant, roughly 1/137. Thus the mass equivalent of the electron's kinetic energy is a small fraction, about α^2, of its rest mass and the electron rest mass is a small fraction of an atomic mass.

When one considers that the principal violation of the zeroth law in chemical reactions comes from changes in the kinetic energy of outer electrons, it becomes plausible that this "law" is violated to no more than parts in 10^9 by chemical reactions. We see confirmations of this limit in tables of covalent bond energies and of the cohesive energies of crystals, to name but two obvious examples.

What about gravitational force? In mass metrology, we also take for granted the weak equivalence principle. In the laboratory, the gravitational force on an object of mass m

is simply $F = mg$, where g is the local acceleration of gravity. The quantity F defines the weight of the object. We tacitly assume that

$$g = G\frac{M_\mathrm{E}}{R_\mathrm{E}^2} + \text{centrifugal} + \dots, \tag{1}$$

where G is the Newtonian gravitational constant [7], M_E is the mass of the Earth and R_E is its radius. The centrifugal term depends on the Earth's latitude, angular velocity and radius but not on its mass—and yet all terms on the right side of eq. (1) are multiplied by the same value of m to obtain F. The passive gravitational mass, m_1, which is appropriate to the first term, is equivalent to the inertial mass, m_2, appropriate to the second so that $m = m_1 = m_2$. Thus we combine all the terms on the right-hand side of (1) into one value of g which, in fact, can be measured to very high accuracy [8]. Furthermore, it makes no difference what "an object of mass m" is made of, *i.e.* how the total mass is partitioned among constituent rest masses, various binding energies, and internal kinetic energies. The total mass m is the sum of the rest masses and the E/c^2 terms. We cannot distinguish between the two in any weighing experiment, although ever more precise tests of these assumptions remain interesting [9]. A brief discussion of the analog of eq. (1) in general relativity, and its implications for the concept of mass, can be found in refs. [1] and [4].

2˙2. *Traditional mass metrology.* – Weighing instruments are shown on the covers of the two most recent monographs on mass metrology [10, 11]. The first of these illustrations is a scene from a medieval apothecary's shop while the second is a well-known funereal drawing that comes down to us from ancient Egypt. Rather than suggesting that weighing is a matter of life and death, these images were no doubt chosen as a reminder that weighing is deeply rooted in human society.

We will not deal further with historical developments (to which ref. [10] devotes an interesting chapter) but instead move directly to consider modern analytical weighing. We will consider that an object to be weighed can be placed on a transducer, which we will refer to as "the balance". The output of this transducer is read in an arbitrary unit, which is proportional to force in newtons. Furthermore, the balance is sensitive only to force in the vertical direction, where "vertical" is defined as the axis of the local acceleration of gravity. Other than these restrictions, the balance is sensitive to any vertical force applied to it. To model this mathematically, we have

$$m_\mathrm{X} g - \rho_\mathrm{F} V_\mathrm{X} g = k I_\mathrm{X}, \tag{2}$$

where

m_X : mass of an object, X;

g : the gravitational acceleration acting at the centre of gravity of X;

ρ_F : the density of a fluid (usually ambient air) surrounding X;

I_X : the balance reading when X is placed on the pan;

k : the conversion factor between the balance unit and the appropriate unit in the International System of units (SI).

The second term on the left-hand side of eq. (2) is due to the application of Archimedes' Principle. If the weighing is done in vacuum, then the term involving ρ_F is negligible. The density of air at 101325 Pa, 293.15 K and 50% relative humidity is $1.199\,\mathrm{kg\,m^{-3}}$. It is useful to know that moist air is a reasonable approximation to an ideal gas. A simple formula for air density, generally more than adequate, is found in ref. [12]. The same reference also gives the CIPM-81/91 formula for the density of moist air, which is much more elaborate. The volume of X, V_X, will generally be a weak function of temperature and an even weaker function of pressure. In Newtonian physics, the mass of X is independent of its temperature.

Mass and volume are extensive quantities, giving a measure of the size of the object X. Their ratio is the density of X. Density is an intrinsic quantity that is a material property, independent of size. (This is strictly true only if the material is homogeneous.)

Because mass, volume and density have the relation

$$\rho_X = \frac{m_X}{V_X}\,, \tag{3}$$

eq. (2) can be rewritten in other useful ways:

$$m_X\left(1-\frac{\rho_F}{\rho_X}\right) = \left(\frac{k}{g}\right) I_X, \tag{4}$$

$$V_X\left(\rho_X-\rho_F\right) = \left(\frac{k}{g}\right) I_X. \tag{5}$$

Assuming we can estimate the density of air and that we know the volume or density of X, we still cannot determine m_X without knowing the ratio k/g. To obtain this additional piece of information, it is usual to include a known mass, S, in the measurements. In analogy to eq. (4),

$$m_S\left(1-\frac{\rho_F}{\rho_S}\right) = \left(\frac{k}{g}\right) I_S, \tag{6}$$

so that

$$\frac{m_X}{m_S} = \frac{I_X}{I_S}\frac{\left(1-\frac{\rho_F}{\rho_S}\right)}{\left(1-\frac{\rho_F}{\rho_X}\right)}\,. \tag{7a}$$

Using the formalism of eq. (2), we may also express the result as

$$m_X - m_S = \frac{k}{g}\big(I_X - I_S\big) + \rho_F\big(V_X - V_S\big), \tag{7b}$$

where the factor k/g can be inserted from eq. (6) or, as we will see in subsubsect. **4**˙2.2, the factor can be determined by means of a built-in calibration weight. The response of real balances drifts with time and this must be accounted for in the weighing scheme. Note that, in order to use eq. (7a), the balance readings must be made with respect to $I = 0$ and so the unloaded balance must be capable of reading zero. Equation (7b) does not require this constraint; the unloaded balance need not be capable of reading zero (sometimes referred to as an "underload" condition) so long as the balance is capable of giving sensible results for I_X and I_S. When used this way, the balance is sometimes referred to as a comparator (see subsect. **4**˙1). Due to physisorbed gas, the mass difference might depend weakly on the difference in surface area between S and X but this term is usually omitted for measurements made under normal atmosphere at approximately 50% relative humidity.

Note: i) k does not appear in eq. (7a) because we have assumed that the balance reading is linear between I_X and I_S and ii) g does not appear in (7a) because we have assumed that the centres of gravity of X and S are at identical height above the balance pan, or at least close enough to ignore the gradient of gravity at the surface of the earth: $\partial(\ln(g))/\partial z \approx -3\times 10^{-7}\,\mathrm{m}^{-1}$, where z increases with height. (It is an amusing exercise to derive the value of the gravitational gradient from the model of a non-rotating, spherical earth of radius R_E, combined with the knowledge that 1 m was originally defined to represent $10^{-7} 2\pi R_\mathrm{E}/4$.)

If m_S is known in units of the SI, then m_X can be determined from eq. (7a) or (7b). Often, however, it is just the ratio of two local masses that needs to be known, in which case the unit of mass is irrelevant. Such a situation is common in the preparation of gravimetric mixtures of chemicals or gases. We shall also see below that the unified atomic mass scale is another case where ratios to the mass of a ^{12}C atom are more useful than traceability to a macroscopic unit of mass.

Equation (7b) is written in terms of differences. This feature may been exploited in several ways: i) The value of ρ_F can be determined experimentally by measuring the difference in balance readings between two mass standards of equal mass and surface but significantly different volumes [13]. ii) The balance factor k/g can be made independent of ρ_F if S and X have different masses but the same volume and surface [14]. iii) Surface effects can be studied if S and X have the same mass and volume but significantly different surface areas [15].

Note that the densities of aqueous liquids are about 800 times the density of air and that the densities of stainless-steel mass standards are about 8 times the density of water. The correction term involving ρ_F (when F represents ambient air) is small, though often significant, when weighing liquids and solids. The correction term is negligible if X and S have the same density (eq. (7a)) or volume (eq. (7b)).

If the density (or volume) of X must be determined, one method is to compare the balance reading of X when it is surrounded air and then by water, both of known density.

Using relations analogous to eq. (5), it is straightforward to show that

$$\frac{\rho_{\mathrm{X}} - \rho_{\mathrm{a}}}{\rho_{\mathrm{X}} - \rho_{\mathrm{w}}} = \frac{I_{\mathrm{X,a}}}{I_{\mathrm{X,w}}}\,, \tag{8}$$

where the subscripts "a" and "w" refer to air and distilled water. Thus the density of X may be determined from a measured ratio of balance readings and the handbook densities of air and distilled water. The last depends primarily on temperature and secondarily on pressure, dissolved air and isotopic composition [16].

Several important questions have been overlooked in this section and still need to be addressed:

- Traceability: How do we know the mass of S *a priori*?

- Non-linearity: Can we assume that real balances are linear devices? This is equivalent to justifying our assumption that k/g is independent of I.

- Interferences: If the balance is an indiscriminate force transducer, what additional effects can bias the results of eqs. (7a) or (7b)?

These and related considerations will be addressed in the following sections.

3. – The kilogram and the International System of Units

The previous section discussed mass as a physical *quantity*. In this section we look at the kilogram, which is the *unit* of mass in the International System of Units (SI) [17]:

"The kilogram is the unit of mass. It is equal to the mass of the international prototype of the kilogram."

The definition dates from 1901 and the artefact in question, the international prototype is a polished cylinder of platinum alloyed with 10% iridium. To minimize its surface area, the height and diameter are both approximately 39 mm. It is maintained at the International Bureau of Weights and Measures (BIPM) and, as we shall see, has been used infrequently since its commissioning in 1889. It is now accepted that the definition cited above refers to the mass of the international prototype immediately after cleaning and washing by means of a specified process [17]. Additional information about the effects of cleaning may be found in ref. [18].

It is often remarked that the kilogram is the last of the seven base units of the SI still defined by an artefact. However, the kilogram is implicit in the definitions of three other base units: the ampere, mole, and candela. Let us recall the first two of these:

"The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 metre apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per metre of length." [17]

Note that the ampere definition refers to the newton and this has SI dimensions of $\mathrm{kg\,m\,s^{-2}}$. Thus the present definition of the kilogram is also implicit in all derived electrical units: volt, ohm, coulomb, farad, etc. Similarly,

"the mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon 12."

Simply put, the numerical value of the mass in kilograms of any entity X is given by the ratio R:

$$R = \frac{m_{\mathrm{X}}}{m_{\mathcal{K}}}\,, \tag{9}$$

where $\mathcal{K}$ is used to denote the international prototype. This definition imposes a hierarchical system, since access to the international prototype is highly restricted. Thus the ratio is achieved by a chain of distributed calibrations that can be shown schematically as

$$\{m_{\mathrm{X}}\}[\mathrm{kg}] = \left\{\frac{m_{\mathrm{X}}}{m_n}\right\}\left\{\frac{m_n}{m_{n-1}}\right\}\cdots\left\{\frac{m_2}{m_1}\right\}\left\{\frac{m_1}{m_{\mathcal{K}}}\right\}[\mathrm{kg}], \tag{10}$$

where the curly brackets signify the numerical value of the quantity within and the square brackets indicate the associated unit. In mass metrology, X may represent secondary mass standards, typically in the ranging from 1 mg to 5000 kg [12]. Propagation of the mass unit to multiples and submultiples of 1 kg can be accomplished used sophisticated weighing schemes and associated data analyses [19, 20]. Every ratio on the right-hand side of eq. (10) represents a measurement with an associated uncertainty. Ratios close to unity and which require only small corrections to remove systematic biases have the smallest uncertainties. For this reason, copies of the international prototype form the first link in the calibration chain. To date, almost 100 copies, or "prototypes", have been manufactured by the BIPM. Six of these are the "official" copies that are stored with the international prototype. Most of the other prototypes have been distributed as "national prototypes" to Member States of the Metre Convention in order to provide world-wide traceability to the mass of the international prototype, and the remaining prototypes are used by the Mass Section of the BIPM to provide calibrations to Member States during periods when the international prototype is not available for use.

An obvious problem for the 21st century is that the SI masses of certain fundamental constants, such as the electron mass and the mass of an atom of ^{12}C, are given in terms of a man-made artefact. The international prototype also helps to determine the SI values of other constants, such as the Planck constant, h, and the fundamental charge, e. Logic favors the definition of artefact properties in terms of fundamental constants. However, to carry out such a programme, one needs, for example, to measure the mass of a 1 kg artefact in terms of the mass of a single atom of ^{12}C. In this case, $R \approx 2 \times 10^{-26}$ and, as we show below, we would like to measure R to an accuracy of several parts in 10^{-8} ! Remarkably, it seems clear that two classes of experiments are on course to achieve this goal, as described elsewhere in this volume [21, 22].

Historically, the mismatch between atomic masses and the international prototype led to the creation of the atomic mass scale. Here the unit of mass is the "unified atomic mass unit", u, which is defined as 1/12 the mass of an atom of ^{12}C at rest and in its ground state. Clearly, u is outside the SI. The conversion factor is defined as

$$\frac{[\mathrm{u}]}{[\mathrm{kg}]} = \frac{1}{12}\left\{\frac{m_{^{12}\mathrm{C}}}{m_{\mathcal{K}}}\right\}. \tag{11}$$

Note that the conversion factor is a measurable quantity. At present, its relative standard uncertainty is given by CODATA as 0.17×10^{-6}, which is orders of magnitude greater than, for example, the uncertainty with which the electron mass is known in atomic mass units [6]. However, the two complementary methods of relating a kilogram artefact to the atomic mass unit at present give discrepant results [6, 21-23], although this problem should be resolved in the near future.

Of course the properties of fundamental constants cannot possibly depend on an artefact such as the international prototype. If one pays careful attention to the correlations among the measured constants, then the influence of the international prototype is removed, provided that the unit of mass has been stable. The atomic mass unit is useful precisely because the ratios of atomic masses do not depend on the international prototype. Using these units effectively removes the covariance terms due to traceability to the international prototype. However, working with large covariances can be a tedious process, which would become largely unnecessary were the kilogram to be redefined in terms of an appropriate fundamental constant. This approach has been emphasized by Mills *et al.* [24] who envisage a new SI, completely based on fundamental constants, perhaps as early as 2011.

There remains the question of the stability of the present definition of the mass unit. By definition, the mass of the international prototype is always 1 kg. Nevertheless, if an experiment of the type implicit in eq. (11) showed that the ratio [u]/[kg] were changing with time, one would know that the mass of the artefact is not stable. With this knowledge in hand, one would be forced to redefine the kilogram in terms of u (in this example). Evidence of this type has not yet been produced, although it has been sought in historical data [25]. Nevertheless, there is convincing internal evidence to show that the relative masses of the prototypes are drifting amongst themselves [23], as shown in fig. 1. The graph shows data obtained from 16 national prototypes, each of which was placed into service in 1889 and each of which was brought back to the BIPM for comparison to the international prototype about 60 years and 100 years later. The median dispersion relative to the mass of the international prototype is of the order of 0.2 μg/year (2×10^{-10}/year in relative mass). We emphasize that this is internal evidence. The *measured* drift of the entire ensemble of prototypes with respect to a fundamental constant of mass is still consistent with zero, within present experimental uncertainties. Another way of stating this is that there is no experiment that can link an artefact of mass 1 kg to a fundamental constant to an uncertainty of 5 μg so that measurements over a period of 25 years would determine changes in mass as small as 0.2 μg/year relative to the fundamental constants. However, the situation is improving rapidly.

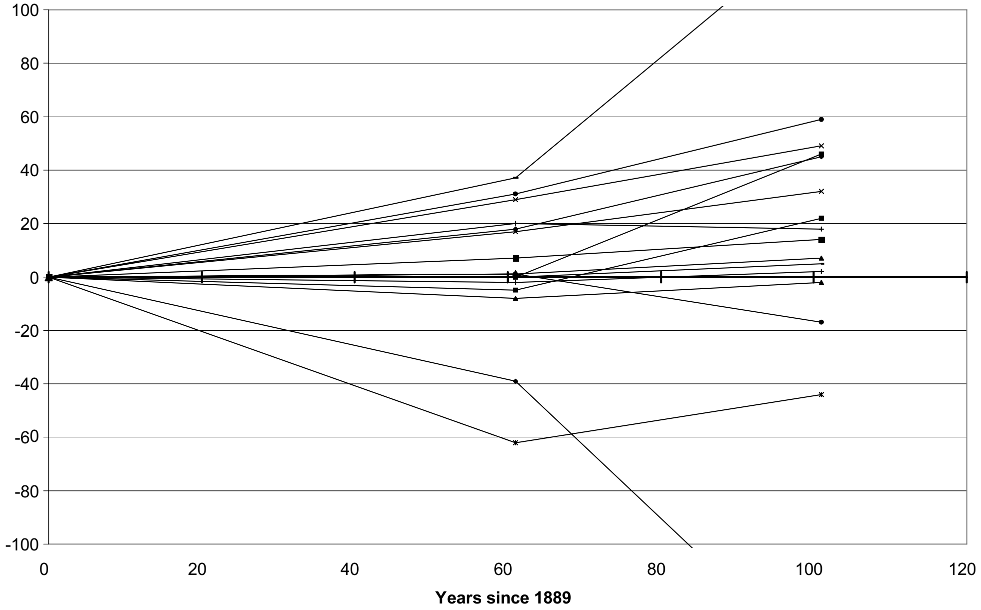


Fig. 1. – Mass with respect to the international prototype of 16 national prototypes. The vertical axis represents $(\{\frac{m_{\mathrm{X_i}}}{m_{\mathcal{K}}}\}_t - \{\frac{m_{\mathrm{X_i}}}{m_{\mathcal{K}}}\}_0)/10^{-9}$ where $\mathrm{X_i}$ represents the i-th of the 17 artefacts, including the international prototype [23].

Of course the internal consistency of the prototypes could mask instabilities that are common to all. For example, surface contamination was removed from the prototypes in order to obtain the data shown in fig. 1 [14, 15, 23]. The method used is solvent cleaning followed by steam washing. The second of these operations is shown in fig. 2. In addition to mass growth due to reversible surface contamination, it has been suggested that traces of mercury vapour, still present in older laboratories, could lead to irreversible growth in the mass of platinum-iridium artefacts [26]. Clearly, both mass metrology and physics will benefit from a non-artefact definition of the kilogram provided that the definition can be realized with sufficient accuracy for 1 kg objects.

4. – How modern balances work

An ideal balance would work as follows: The object of interest is placed on the balance pan and the accurate mass of the object is immediately available for input to a computer or visual display. Modern balances approach this goal. However, the fact that balances respond to the vertical force applied to the pan rather than to the mass placed on the pan means that, even if the balance were a perfect force transducer for vertical forces, air density must be taken into account. In addition, thermal effects and extraneous electrical or magnetic forces must be minimized. Less than perfect transducers must be engineered to be immune to side forces and off-centre loading of the pan. Non-linearity and hysteresis

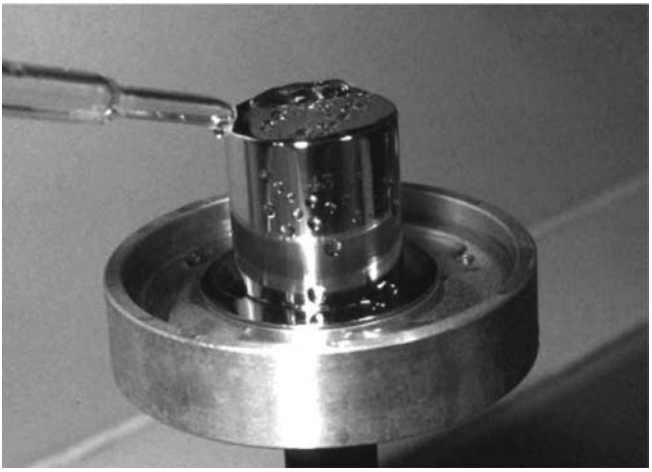


Fig. 2. – A copy of the international prototype being steam cleaned to remove surface contamination. Copies of the international prototype are made of the same alloy, have the same nominal dimensions and are manufactured to have the same mass (within a tolerance of ±1 mg) as the international prototype. The mass each copy is determined by a calibration having a relative standard uncertainty of order 5×10^{-9} [23]. Credit: BIPM photo.

in the transducer must be eliminated or compensated. Finally, if the end result is to be a mass in SI units, the transducer output must be traceable to the kilogram.

We will first consider the ultimate performance that might be achieved by a balance designed to compare the mass of two 1 kg weights. We then turn to the more practical aspects of weighing with commercially available balances.

4˙1. *Ultimate performance.* – Balances are required in order to realize the calibration chain shown schematically in eq. (10). In subsect. **2**˙2, we treated the balance as a kind of transducer whose response is an electrical signal proportional to the vertical force on the loading pan. In fact, this is a serviceable description of modern electronic balances.

The best 1 kg balances are "mass comparators", which means that they can measure the *difference* between two 1 kg objects but the transducer range is limited to 1 g or less. A careful and complete analysis of the fundamental limits to mass comparison using such a device has been given by Speake [27]. The quantum-mechanical limit is derived economically and shown to be entirely negligible. We will present the more practical problem of anelasticity and show how it contributes to $1/f$ noise. Anelasticity is an ubiquitous phenomenon [28] and thus it should interest anyone attempting a precise mechanical measurement. We will also discuss the immunity of the balance to noise of various frequencies. The reader should also consult the useful review paper by Quinn [29].

4˙1.1. Anelasticity. A traditional single-pan balance is shown schematically in fig. 3. The beam performs damped harmonic motion about an equilibrium angle, θ_0. Although there is no damper shown in the figure, there will always be loss mechanisms in the pivots

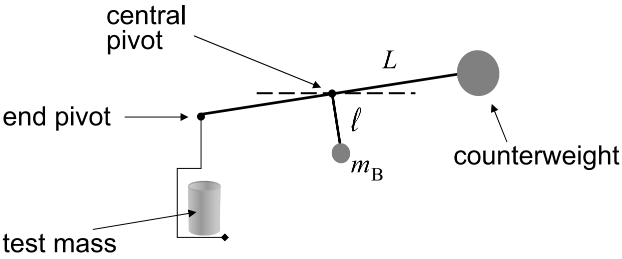


Fig. 3. – Schematic view of traditional one-pan balance without servocontrol. We take the simple case of equal arms of length L and further assume that the pivots are aligned with the centre of mass of the counterweight and that the centre of mass of the beam is located a distance ℓ below the centre of the beam. For simplicity, we also assume that the pan and suspension are massless. Thus the balance beam is horizontal when the mass of the test object placed on the pan equals the mass of the counterweight.

as well as viscous damping due to the ambient atmosphere. This harmonic-oscillator model is heuristic but contains useful features of the real mechanical behaviour of mass comparators. A simplified transfer function of the balance shown in fig. 3 is given by

$$T = -\omega^2 J + \mathrm{i}\omega\gamma + k_0 + k_\mathrm{f}, \tag{12a}$$

where $T(\omega)$ is a driving torque, J is the total moment of inertia of the balance and load about the central pivot, γ is a constant parameter that models velocity-dependent damping processes (which may be unintended and, therefore, very small), and k_0 is the mechanical stiffness of the balance. The last is easily estimated from the equilibrium angle of fig. 3 to be

$$k_0 = \frac{\Delta m}{\Delta\theta} gL \approx m_\mathrm{B}\ell g. \tag{12b}$$

In our model, the balance beam is horizontal ($\theta = 0$) when $\Delta m = 0$. Equation (12b) thus describes the equilibrium angle $\theta_0 = \Delta\theta$ due to a small imbalance in mass between the test mass and counterweight, if k_f were negligible. The term k_f is due to the elastic properties of the pivots. In modern balances, these pivots are flexure strips [29] and k_f is may be estimated from the dimensions of the strip and the elastic modulus E of the strip material. For such strips, k_f is *not* negligible compared to k_0. When $T = 0$, one can observe the "natural frequency", f_0, of the balance. From eq. (12a), this is given by

$$f_0 \cong \frac{1}{2\pi}\sqrt{\frac{k_0 + k_\mathrm{f}}{J}}\,.$$

All materials have the property known as anelasticity (also known as internal friction). As a consequence the elastic modulus, E, of any material used to construct balance pivots

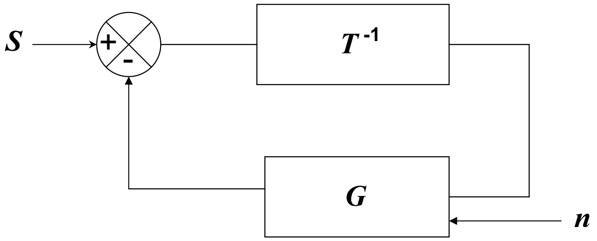


Fig. 4. – Block diagram of a servocontrol loop (after Speake [27]).

contains a small imaginary term: $E = E_0(1+\mathrm{i}\phi)$, where $\phi \ll 1$. Thus a small component of the elastic response to a periodic driving frequency $f = \omega/(2\pi)$ is 90° out of phase with the driving frequency. This in turn implies that $k_\mathrm{f} = k_\mathrm{r} + \mathrm{i}k_\mathrm{i}$, where k_r and k_i are parameters and $k_\mathrm{r} \gg k_\mathrm{i}$. Finally, eq. (12a) must be modified to take account of this extra source of damping:

$$T = -\omega^2 J + \mathrm{i}\omega\left[\gamma + \frac{k_\mathrm{i}}{\omega}\right] + k_0 + k_\mathrm{r}. \tag{13}$$

We have taken k_i to be a constant because experiment shows that ϕ, and therefore k_i, is essentially independent of frequency (at least over the wide spectrum that interests us). This is a key feature of anelasticity [28]. The imaginary, or damping, term in eq. (13) can be related to the spectral density of noise torques through the fluctuation-dissipation theorem. The term in γ determines the strength of the white-noise spectrum [27]. We recall that white noise can be reduced by averaging over time. However, the term in k_i/ω determines the strength of a $1/f$ noise spectrum.

4˙1.2. Effect of servocontrol and detector noise. We now add servocontrol to the mechanical balance shown in fig. 3 [27,30]. The control loop is shown schematically in fig. 4. An idealized transfer function, $T(\omega)$, of the unservoed balance has already been given in eq. (12a). A signal $S(\omega)$, which is a torque imbalance, changes the beam angle from zero to $\theta(\omega)$; the servocontrol block $G(\omega)$ includes an electromechanical force transducer that responds by damping the balance oscillations and producing a compensating torque that drives the balance beam back to its reference position, $\theta = 0$. The final balance reading is directly proportional to the final compensation torque. The electromechanical stiffness added by the servocontrol, k_sc, is generally much greater than k_0, k_r. These stiffnesses are additive, as springs acting in parallel, and thus the stiffness of the servoed balance is much greater than it is without servocontrol.

Now suppose that the angle detector in G receives a noise signal, $n(\omega)$. Then for frequencies much below the natural frequency of the mechanical balance, the signal-to-noise ratio, *SNR*, is given by $SNR = S(\omega)/(k_0 + k_\mathrm{r})$. This is the usual frequency regime for precision 1 kg mass comparators. Speake has shown that this consideration puts a very tight limit on detector noise. For frequencies much higher than the natural

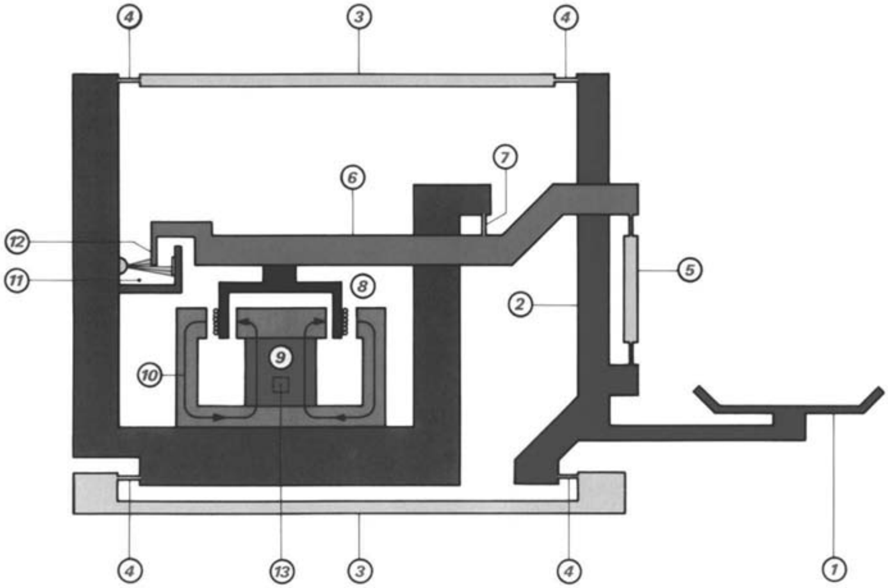


Fig. 5. – Sketch of a modern analytical balance. The user sees only the pan (1). Of particular interest is the mechanical structure consisting of: the balance beam (6), the central pivot (7), the flexible parallelogram structure defined by pivots (4), and the servocontrol mechanism consisting of: the angle detector (11); the current-carrying coil (8); the permanent magnet (9); the magnetic circuit (10); the temperature sensor (13). Credit: Courtesy of Mettler-Toledo

frequency of the balance, the $SNR = S(\omega)/(I\omega^2)$, the inertia of the balance acting as a noise filter. Note that k_{sc} does not appear in these calculations of *SNR*.

4'2. *Commercial analytical balances.* – While few scientists will build state-of-the-art mass comparators, many will work with analytical balances of the type shown schematically in fig. 5. Only the pan is accessible to the user, the balance mechanism itself being a "black box". For ease of access, the pan is not suspended as in fig. 3. Instead, as the beam rotates, the pan is guided by a parallelogram structure so that it moves only in the vertical direction. The motion of the beam is tightly servocontrolled so that the parallelogram operates over a very small distance. Also noteworthy in fig. 5 is the presence of a large permanent magnet. By sending electrical current i through the coil, the servocontrol system produces the necessary electromagnetic force to maintain the balance beam at horizontal. The electromagnetic force equals iDB where B is the magnetic flux density in the air-gap of the permanent magnet and D is a geometrical term that is essentially the total length of current-carrying wire that is perpendicular to B. It is a property of permanent magnets that the flux density B is a weak function of temperature, which explains the thermometer in fig. 5. The core metallic components of analytical balances can now be made from a monolithic block of metal, as shown in fig. 6. In order to produce the balance reading, the servocontrol current is generally converted to a voltage and then digitized. We now discuss some important aspects of this remarkable technology.

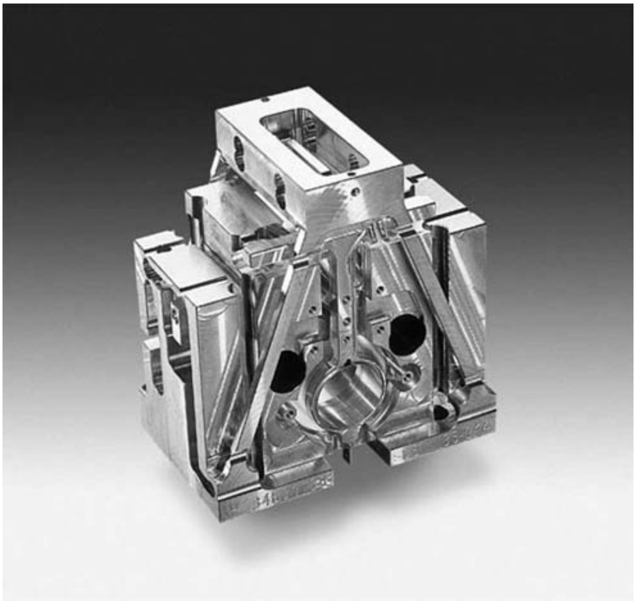


Fig. 6. – The mechanical structure of a modern analytical balance formed from a block of metal. As shown, the block is rotated 90° from its assembled position so that it is resting on its rear, horizontal surface. Credit: Courtesy of Sartorius AG

4˙2.1. Centring. If, instead of being constrained to vertical motion by the parallelogram, the plane of the pan rotated with the balance beam, then the balance response would depend on where an object is placed on the pan. This is called "centring error". The maximum centring error is specified by the manufacturer. However, the sensitivity of the balance to off-centre loading can be tested by the user who should, in any event, make an effort to place all loads on the centre of the balance pan. Though less convenient to use, balances of the type shown in fig. 3 can have completely negligible centring errors [29].

4˙2.2. Scale calibration. The electrical output of the balance is usually given in units of mass. In order to do this, the best analytical balances have one or two built-in calibration weights. The user can then request an automatic calibration of the balance, the result of which has certain important consequences. As an example, let us consider an analytical balance with 100 g capacity and a scale that reads from zero to 100. Let us ignore the fact that the balance appears to be reading in grams and assume only that its reading is in terms of some arbitrary unit. Now suppose that the balance has an internal standard S whose mass is precisely known to be 100.002 g. The balance scale is first set to zero. The standard weight is loaded and the internal circuitry of the balance then forces the balance reading to be exactly 100.002. What has happened can be shown though a slight modification to eq. (6):

$$\{100.002\}[\mathrm{g}]\left(1-\frac{\rho_{\mathrm{a1}}}{\rho_{\mathrm{S}}}\right)=\left(\frac{k}{g}\right)\{100.002\}[\mathrm{unit}], \tag{14}$$

where the air density at the time of the internal balance calibration is ρ_{a1}. The left-hand side of eq. (14) represents a corporal mass standard that is acted on by gravity and affected by air buoyancy. To be more explicit, we have shown the measurement units within square brackets and the numerical value of the associated with the unit within curly brackets. Here [unit] represents the arbitrary balance unit, which must be converted to the SI unit, gram (units, such as g, are not italicized to distinguish them from quantities, such as the local acceleration of gravity, g). Once the calibration has been completed, the conversion factor k/g has been fixed to be equal to $(1-\rho_{\mathrm{a1}}/\rho_{\mathrm{S}})[\mathrm{g}]/[\mathrm{unit}]$. Now when an object X is placed on the balance pan, eq. (4) becomes

$$m_{\mathrm{X}}\left(1-\frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{X}}}\right)=I_{\mathrm{X}}\left(1-\frac{\rho_{\mathrm{a1}}}{\rho_{\mathrm{S}}}\right), \tag{15}$$

where ρ_{a} is the air density when X was placed on the balance and both sides of the equation are in terms of the SI unit, gram. Usually it can be assumed that $\rho_{\mathrm{a}}=\rho_{\mathrm{a1}}$ and thus eq. (15) is the same as if X had been exactly balanced by a corporal weight of mass I_{X} and density ρ_{S}. Sometimes these balances are referred to as “direct-reading” but one should be aware that $m_{\mathrm{X}} \neq I_{\mathrm{X}}$, even after the balance has been calibrated using an internal or external mass standard. Actually, the balance manufacturer uses the conventional mass and density of the internal standard, as described below in sect. **5**, so that ρ_{S} should be taken to be $8000\,\mathrm{kg/m^3}$. As with eq. (7a), we have not yet dealt with the effect of non-linearity. We have also ignored the effect of the height difference between the centres of gravity of S and X (see subsect. **2**˙2).

If the internal calibration weights of the balance are insufficiently accurate, then the user may calibrate the scale using his/her own calibrated weight. If the balance software permits this, then eq. (15) still applies. If not, then eq. (6) should be used to determine k/g.

4˙2.3. Nonlinearity. In mathematics, a linear function F has the feature of additivity:

$$F(\mathrm{A})+F(\mathrm{B})=F(\mathrm{A}+\mathrm{B}), \tag{16}$$

for any A and B. If we consider $F(\mathrm{X})$ to be the operation of putting a mass X onto the balance with result $F(\mathrm{X})=I_{\mathrm{X}}$, then eq. (16) describes what we mean by a linear balance. Once again, we simplify by ignoring effects that are usually negligible: differences in the centres of gravity of A and B, changes in air density during the course of the measurements. Nonlinearity is usually determined in the following way. Start with a calibrated set of weights having the same density, such that the largest of these, with mass $m_{\max}$, is near the maximum capacity of the balance. Let the weighing result for this weight be $I_{\max}$. Plot this point on a graph whose horizontal axis starts at zero and continues to $m_{\max}$ and whose vertical axis starts at zero and continues to $I_{\max}$. Connect the points (0,0) and $(m_{\max}, I_{\max})$ by a straight line. Now measure other weights or combinations of weights in the set in order to plot a set of at least nine additional points with approximately equal spacing along the horizontal axis. Deviations from the straight line describe the nonlinearity of the balance. Often the nonlinearity is significant over the full range of the balance but insignificant over, say, 1/10 the balance range. In this case, it is best to compare the balance reading of the object of interest, X, with that of a standard weight, S1, whose mass is within $m_{\max}/10$ of that of X. The result is then derived from eq. (7b), which we now rewrite in terms of the density of S1 and the calibrated value for k/g:

$$m_{\mathrm{X}} = m_{\mathrm{S1}}\left(1 - \frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{S1}}}\right) + (I_{\mathrm{X}} - I_{\mathrm{S1}})\left(1 - \frac{\rho_{\mathrm{a1}}}{\rho_{\mathrm{S}}}\right) + \rho_{\mathrm{a}} V_{\mathrm{X}}. \tag{17}$$

By minimizing the magnitude of $(I_X - I_{S1})$, the influences of the balance calibration, including scale nonlinearity, are also minimized. However this strategy no longer exploits all the convenience features of the balance. The best course is to determine the target uncertainty for m_{X} and then to use the most convenient procedure that will achieve it.

4˙2.4. Magnetic interactions. As seen in fig. 5, the balance pan is not far from a large permanent magnet. While this magnet is reasonably well shielded (the shielding is better than might be inferred from the schematic drawing), there are inevitably stray fields in the vicinity of the balance pan. These fields are small and would cause no problem so long as the objects placed on the pan are not themselves magnetized and do not have high magnetic susceptibility. Magnetic susceptibility is a material property of the alloy, whereas magnetization can vary from sample to sample depending on exposure to strong magnetic fields, cold working, heat treatment, etc. [31]. The best commercially available standard weights are made of “nonmagnetic” alloys of stainless steel. Nevertheless, the magnetic properties of these alloys can vary significantly. For this reason, an international recommendation proposes limits to the magnetization and magnetic susceptibility of mass standards as well as means for testing whether the limits have been achieved [12]. The weighing of ferrous or magnetic materials is obviously problematic.

4˙2.5. Thermal effects. In writing a basic weighing model, such as that described in eq. (2), it is tacitly assumed that the only effect of ambient air is that of buoyancy (Archimedes’ Principle). This is true under equilibrium conditions but serious errors can occur in the presence of air convection. Such errors have been studied under the condition that the objects to be weighed are not in thermal equilibrium with the balance environment and when the walls of the balance enclosure do not have the same temperature. These effects can lead to systematic errors that are greater than the standard deviation of the measurements as modeled analytically [32] and by finite element analysis [33]. Weighing under vacuum obviously eliminates the possibility of air convection as well as the need for a buoyancy correction.

5. – Legal metrology and conventional mass

5'1. *What scientists should know about conventional mass.* – Scientists are often surprised to learn that commercially available mass standards are manufactured to represent something called "conventional mass" and, although the units of conventional mass are the SI units kilograms, grams, etc., the quantity that they describe is not exactly the same as the quantity m in the formula $F = ma$. The latter is sometimes referred to as "true mass" or mass in the Newtonian sense, in order to avoid ambiguity. In this section we give a motivation for the use of conventional mass and describe how metrologists and other scientists can live comfortably with this convention.

We begin by rewriting eq. (4) but now assuming that the fluid surrounding the object is laboratory air of density ρ_a:

$$m_\mathrm{r}\left(1-\frac{\rho_\mathrm{a}}{\rho_\mathrm{r}}\right)=k'I_\mathrm{r}. \tag{18}$$

We use the subscript r for "reference" because, as we shall see, the concept of conventional mass can be useful for measurement scientists but this utility rarely extends beyond the mass assigned to commercial, stainless-steel mass standards when they are used as reference standards.

Conventional mass exploits the correlation between m_r and ρ_r when air buoyancy cannot be neglected. As we have seen above, the density of air near sea level is approximately $1.2\,\mathrm{kg/m^3}$. The International Organization of Legal Metrology (OIML) defines this value as the "conventional" air density and given the symbol ρ_0 [12, 34]. Instead of its real density, ρ_r, let us now assign the OIML conventional density, ρ_c, to r with the value $\rho_\mathrm{c} = 8000\,\mathrm{kg/m^3}$ at 20 °C. Since the right-hand side of eq. (18) is a measured result which is correct for an arbitrary air density, our choices of conventional air density and stainless-steel density impose the following definition of the conventional mass of r, $m_\mathrm{c,r}$:

$$m_\mathrm{c,r}\left(1-\frac{\rho_0}{\rho_\mathrm{c}}\right)=m_\mathrm{r}\left(1-\frac{\rho_0}{\rho_\mathrm{r}}\right). \tag{19}$$

Thus the left-hand side of eq. (19) can always be substituted for the right-hand side in equations such as eq. (18). Note that $(1-\rho_0/\rho_\mathrm{c})$ is simply the number 0.999850. High-quality weights used for science are generally OIML class E_1, E_2 or F_1 where the manufacturing tolerances and physical properties of the weights are best for E_1 and are successively relaxed for E_2 and F_1 [12]. It is often not appreciated that the tolerances quoted by manufacturers are differences between the *conventional* mass and the nominal mass of the weight.

The mass tolerance determines the allowed density of the alloy used to construct the weight. This is because eq. (19) is just an approximation to the general case $\rho_\mathrm{a} \neq \rho_0$. Only if $\rho_\mathrm{r} = \rho_\mathrm{c}$ will the equality shown in eq. (19) be maintained irrespective of the air density. In general, use of conventional mass and density instead of their "true" counterparts leads to an error $\Delta m_{\mathrm{w,r}}$:

$$\text{(20a)} \qquad \Delta m_{\mathrm{w,r}} = m_{\mathrm{r}}\left(1 - \frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{r}}}\right) - m_{\mathrm{c,r}}\left(1 - \frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{c}}}\right).$$

The requirement of [12] is that for ρ_{a} within $\rho_0 \pm 0.1\rho_0$, $\Delta m_{\mathrm{w,r}}$ must be within 1/4 of the specified tolerance for the weight. The only way to meet the requirement is to use stainless-steel alloys whose densities are sufficiently close to $8000\,\mathrm{kg/m^3}$. For instance, the tolerance for a 1 kg weight of class E_2 is ± 1.5 mg, leading to the published requirement $7810\,\mathrm{kg/m^3} < \rho_{\mathrm{r}} < 8210\,\mathrm{kg/m^3}$, which has been rounded. It follows from the definition of $m_{\mathrm{c,r}}$ (eq. (13)) that

$$\text{(20b)} \qquad \Delta m_{\mathrm{w,r}} = m_{\mathrm{r}}\left(\rho_0 - \rho_{\mathrm{a}}\right)\left(\frac{1}{\rho_{\mathrm{t}}} - \frac{1}{\rho_{\mathrm{c}}}\right)\left(1 + \frac{\rho_0}{\rho_{\mathrm{c}}} + \left(\frac{\rho_0}{\rho_{\mathrm{c}}}\right)^2 + \ldots\right).$$

The final factor on the right-hand side of eq. (20b) is a power series in ρ_0/ρ_{c} but this can be truncated to 1 if the tolerance limits for density have been respected.

The simplification arising from this formalism is that, if the user does not require uncertainties smaller than the combined uncertainty of the tolerance and $\Delta m_{\mathrm{w,r}}$, then it is reasonable to treat all weights in a set as if they have the density $8000\,\mathrm{kg/m^3}$ and to use their nominal values as the estimate of their conventional mass. If this uncertainty is too large, then the weights require further calibration. At this stage, it is often preferable to have the standard weights calibrated in terms of mass in the Newtonian sense instead of in terms of their conventional mass. Since the former requires a value for ρ_{r}, the conventional mass can always be calculated by the user if desired.

Generally, scientific metrology requires mass in the Newtonian sense ("true" mass) as the end result, as in eq. (17). Note that this equation can be rewritten as either

$$\text{(21a)} \qquad m_{\mathrm{X}} = m_{\mathrm{S}}\left(1 - \frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{S}}}\right) + k'\big(I_{\mathrm{X}} - I_{\mathrm{S}}\big) + \rho_{\mathrm{a}} V_{\mathrm{X}},$$

or

$$\text{(21b)} \qquad m_{\mathrm{X}} = m_{\mathrm{S,c}}\left(1 - \frac{\rho_{\mathrm{a}}}{\rho_{\mathrm{c}}}\right) + \Delta m_{\mathrm{w,S}} + k'\big(I_{\mathrm{X}} - I_{\mathrm{S}}\big) + \rho_{\mathrm{a}} V_{\mathrm{X}}.$$

In both cases, the properties of the unknown X are their Newtonian properties. Note that we have used the symbol k' to represent the experimental value of k/g and we assume that the difference is balance readings is small enough so that k' may be calibrated either in terms of "true" or conventional mass. Either eq. (21a) or eq. (21b) can be used. Using the same input information, the same uncertainty can be obtained using either formalism. However, eq. (21a) is the more straight-forward for highest-accuracy work since $\Delta m_{\mathrm{w,S}}$ is a term that converts the conventional mass and buoyancy correction back to their Newtonian counterparts. For some applications, the weights themselves are used in vacuum, as in the case of piston gauges operated in the absolute mode [35]. Here the ambient air density is 100% smaller than ρ_0 and thus the use of conventional mass is inappropriate. Each piston weight should be calibrated as the object X in eqs. (21a) or (21b).

The 2004 edition of OIML Recommendation 111-1 [12] is a long and useful document that contains a wealth of information: descriptions of the various classes of weights that are commercially available, methods to evaluate uncertainties, equations for the density of ambient air (including a formula to estimate the air density when only the elevation of the laboratory is known), methods to determine the volume or density of mass standards, methods to evaluate the magnetic properties of stainless steel alloys, calibration of electronic balance readings, and waiting times to achieve thermal equilibrium are among them.

5˙2. *Another way to view conventional mass.* – There is another useful way to look at conventional mass. Consider the standard S with mass m_S and volume ρ_S that appears in eq. (21a). Assume that S has been calibrated by a standards laboratory that has provided a certificate stating the "true" mass and corresponding uncertainty, the alloy density and corresponding uncertainty, and the air density $\rho_{\mathrm{a}1}$ at which the calibration was carried out. This value of $\rho_{\mathrm{a}1}$ was used to compute m_S from balance readings, assuming that the density of S is ρ_S but recognizing that this density value has a known uncertainty $u(\rho_\mathrm{S})$. The "true" mass of S will thus have an uncertainty component $m_\mathrm{S}\rho_{\mathrm{a}1}u(\rho_\mathrm{S})/{\rho_\mathrm{S}}^2$ which is due to $u(\rho_\mathrm{S})$. The corresponding component for the uncertainty of the conventional mass $m_{\mathrm{c,S}}$ appears to be much smaller, $(\rho_{\mathrm{a}1}-\rho_0)u(\rho_\mathrm{S})/{\rho_\mathrm{S}}^2$. Thus the uncertainty in alloy density has a relatively small effect on conventional mass compared to "true" mass.

However, appearances can be deceiving. We are assuming in this discussion that mass standards are ultimately used to calibrate something of scientific interest, which we have denoted as X in eq. (21a). By simple substitution of the weighing equation used by the calibrating laboratory to determine the value m_S, it can be shown that $u(m_\mathrm{X})$ has an uncertainty component $(\rho_\mathrm{a}-\rho_{\mathrm{a}1})u(\rho_\mathrm{S})/{\rho_\mathrm{S}}^2$ due the uncertainty in the density of S. Of course the uncertainty cannot be different if one casts the weighing equation for X in terms of the conventional mass of S as in eq. (21b). Some purchasers of standard weights ask for a calibration of weight densities small enough to make $m_\mathrm{S}\rho_{\mathrm{a}1}u(\rho_\mathrm{S})/{\rho_\mathrm{S}}^2$ negligible for their work. However, for normal laboratory environments, it is only the contribution $(\rho_\mathrm{a}-\rho_{\mathrm{a}1})u(\rho_\mathrm{S})/{\rho_\mathrm{S}}^2$ that must be acceptably small in order to determine the mass of X.

6. – Nanograms to yoctograms ("there's plenty of room at the bottom")[1]

6˙1. *Mass and nanometrology.* – In a real sense, the vast and growing field of nanometrology was already foreseen by Richard Feynman in his 1959 address "There's

[1] This section is a revised version of a keynote address given by the author to the 19th Conference of IMEKO TC3, Cairo, Egypt, 19-23 February 2005.

plenty of room at the bottom" [36]. Mass metrology has not been immune to this revolution [37]. The smallest mass standard commonly used in metrology is 1 mg. The smallest resolution of mass comparators commonly used in mass metrology is $0.1\,\mu$g. However, recent advances in technology have created new fields and opportunities for mass measurements far smaller than these limits.

We start with the atomic force microscope (AFM) and its "calibration". This device is based on a flexible cantilever, typically $100\,\mu$m long. The free end is attracted to (or repelled by) the object under study and the deflection is measured to a precision of picometres. The cantilever has a spring constant that is typically about 0.1 N/m. The challenge is to determine this constant, either by calibration or by other means. The problem is analogous to determining the sensitivity of a sensitive balance. The focus here will be on traceability to the kilogram of AFM calibrations.

The next type of device to be considered is that reported by Craighead and colleagues at Cornell University (USA) [38,39]. It is also a cantilever but, in this case, it is only $4\,\mu$m long and is used to determine the inertial mass of microscopic objects. The developers have demonstrated that their device is suitable for mass determinations in the attogram range (1 attogram = 10^{-18} g). The Cornell group have now configured their device to detect a change in mass due to the presence of a specific virus.

Finally, we recall W. Paul's elegant determination of the gravitational mass of the neutron [40]. Paul was able to create a magnetic "spring" within a neutron storage ring and determine how far this spring stretches due to the added weight of a neutron. The sensitivity of the measurement was 0.1 yg (1 yoctogram = 10^{-24} g). His result is consistent with the accepted value of the inertial mass of the neutron. Although this work was carried out about 25 years ago, it is not known to many mass metrologists.

6˙2. *Cantilever systems.* – The AFM and the most well-known Craighead devices are cantilevers (fig. 7). In its operational mode, the tip of the AFM is acted on by the force due to a proximate surface, thereby causing the cantilever to bend. The free end of a horizontal cantilever will deflect in the vertical direction when a vertical force F is applied. If the deflection, Δz is small, then the cantilever can be described as a simple spring with constant k:

$$F = k\Delta z. \tag{22}$$

Assuming that Δz can be accurately measured, calibration of the device consists of determining k. There is in fact no requirement that the cantilever be horizontal. However, if it is horizontal, we have the possibility that F can be produced by the gravitational weight of a standard mass, in complete analogy to the calibration of k/g described above in subsect. **4**˙2.2.

Cantilevers can also operate in a dynamic mode whereby the free end of the cantilever is made to oscillate at its resonant frequency f_0, giving us a second equation:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}}\,, \tag{23}$$

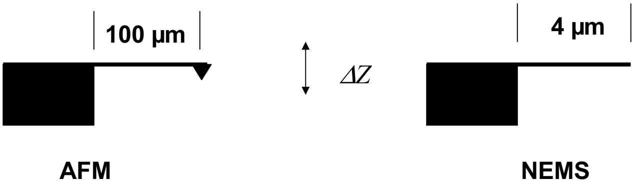


Fig. 7. – Schematic drawing of cantilevers used for mass (and force) measurements below 0.1 μg (1 nN).

where m is the mass at the tip of a "massless" cantilever. If the cantilever cannot be considered to be massless, then a change in mass of Δm will lead to a new resonant frequency f_1 so that

$$\Delta m = \frac{k}{(2\pi)^2}\left[\frac{1}{f_1^2} - \frac{1}{f_0^2}\right]. \tag{24}$$

Note that the same spring constant k appears in both eqs. (23) and (24).

6·3. *Calibration of AFMs using calibrated mass standards (deadweights).* – There are a number of interesting reviews on this subject (*e.g.*, refs. [41,42]). There are three main strategies for determining k:

a) Theoretical calculation. For instance, a finite element analysis can determine k based on the dimensions of the cantilever and the physical properties of the material(s) from which it is made.

b) Static deflection. A known force can be applied to the tip of the cantilever and k can be determined from eq. (22).

c) Dynamic vibrational response. The value of k can be determined from eqs. (23) or (24).

In this discussion of mass metrology, we take as our only example the second strategy, nevertheless aware that the other strategies cannot be ignored in practice. In particular, we describe briefly recent results from Pratt and his colleagues [43] at the National Institute of Standards and Technology (NIST). Their apparatus is shown in fig. 8. From the schematic diagram, it can be seen that a force along the vertical axis can be applied in two ways. The first is by adding a "deadweight" of either 20 μN or 200 μN to the top of the vertical column. (The term "deadweight" traditionally applies to mass standards used to apply known gravitational forces up to the order of meganewtons. The analogy is exact for the device shown in fig. 8.) The second method is based on a capacitive transducer whose inner and outer electrodes are shown in fig. 8, can be calibrated from first principles. Recall that the electrostatic potential energy U that is stored in a capacitor

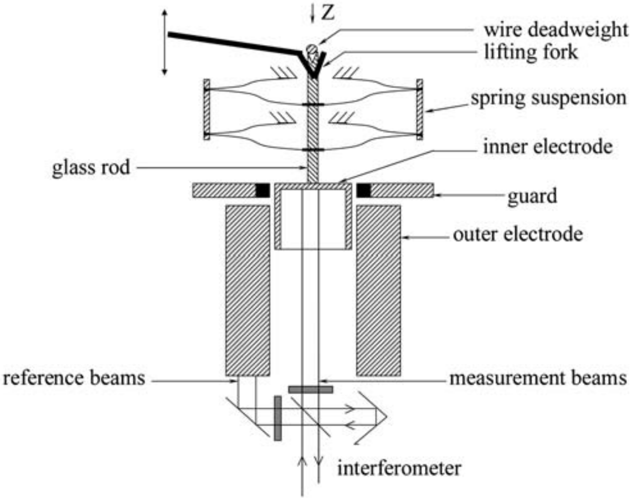


Fig. 8. – Schematic diagram of NIST apparatus. The diameter of the inner electrode is 15 mm. (Provided by the Small Force Metrology Laboratory, NIST.)

is given by

$$U = \frac{1}{2} C V^2, \tag{25}$$

where V is the voltage across the capacitor plates. The vertical force of this system is given by the gradient of U in the Z direction and, finally, a change in force is given by

$$\Delta F = \frac{1}{2} \frac{\partial C}{\partial Z} \left(V_1^2 - V_2^2 \right). \tag{26}$$

The derivative of C must be determined in a separate experiment. This ideal capacitor does not depend on the polarity of the two voltages because each is squared. In practice, the right-hand side of eq. (26) also contains a linear term in the voltage difference, due to patch effects [43]. The patch effects are eliminated by using voltages of alternating polarity. The authors argue plausibly that, if the electrostatic force deduced from eq. (26) agrees with that derived from a deadweight, then this indicates that the transducer is working properly. This approach is analogous to earlier attempts to determine the SI volt in terms of the base units [44].

The transducer may then used to calibrate AFMs. After considerable effort the NIST apparatus has demonstrated agreement with deadweights to better than 0.1% when operating with amplitude-modulated a.c. voltage. A similar approach is under active development by Choi and colleagues [45]. As already mentioned, one should also be aware of other calibration strategies, *e.g.*, ref. [46].

6˙4. *Oscillating cantilevers.* – A general introduction to this subject has been given by Lavrik and Datskos [47,48]. These authors point out that, since the weight of objects of molecular size is too small to be measured, one determines the inertial mass as, for example, in eq. (24). As mentioned in subsect. **2**˙1, the equivalence of gravitational mass (used to calculate weight) and inertial mass has been tested to accuracies far beyond the needs of mass metrology. According to ref. [48], a general rule is that the smallest change in mass detectable at room temperature is roughly 10^{-6} times the mass of the cantilever. This limit is determined by thermal noise in the oscillation frequency of the cantilever. Therefore the cantilever dimensions must be reduced to achieve greater precision. However, reduction in size below optical wavelengths makes detection of the oscillation frequency difficult.

It is arguable that the performance of oscillating cantilevers has been revolutionized by the Craighead Group and Cornell University (USA) [38]. Their devices, and others of similar size, are referred to as NEMS (nanoelectromechanical systems). In a paper published in 2004 [39], Craighead and his colleagues reported a cantilever system with a resolution of better than 1 ag. Each cantilever is typically 4 μm long, 500 nm wide and 160 nm thick and fabricated from a silicon/silicon nitride wafer. There is a square, paddle-shaped "pan" 1 μm on a side at the free end. The resonant frequency of such an oscillator ranges from 1 MHz to 15 MHz.

How does one determine the sensitivity of such a device? Determining the resonant frequency based on dimensions and materials properties is problematic because, in particular, the thickness of the cantilever is difficult to determine from imaging techniques. The procedure actually adopted was to attach a gold dot to the paddle at the free end of the cantilever. First, the resonant frequency of the cantilever system was measured. Then the cantilever was exposed to thiolate, a sulphur-based organic compound. Thiolate is known to form a self-organized monolayer on gold surfaces but does not adhere to the cantilever itself. A calculation showed that, in one instance, the thiolate layer increased the mass at the end of the cantilever by only 6 ag, which was easily detected by experiment. Cantilevers having gold dots with diameters from 50 nm to 400 nm were studied. A careful calculation estimated the resolution of the device to be about 0.4 ag. It was also possible to measure the effect of the gold dots themselves on the resonant frequency of the cantilevers.

The Craighead Group intends to extend their work to the zeptogram range (1 zg = 10^{-21} g), in which case the devices could be configured to identify DNA and other biological molecules [38,39].

A recent paper estimates the ultimate limits to NEMS devices [49]. Other review papers provide a useful introduction to this subject [50,51]. It is instructive to contrast the fundamental limits of these devices, which may have attogram sensitivities, to those of a 1 kg balance, as derived in ref. [27].

6˙5. *Gravitational mass of a neutron.* – As with most generalizations, the remark that masses of molecular size and smaller are too small to be weighed [48] has at least one exception. Although the exception has no great practical application, it is of considerable

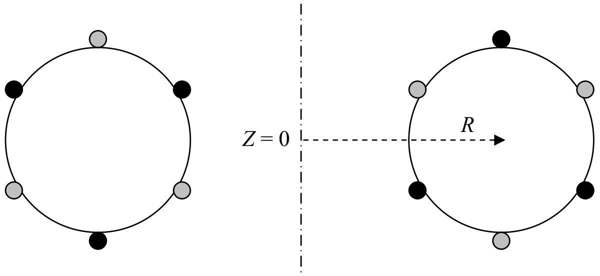


Fig. 9. – The neutron storage ring is a toroid of radius R. A sextupole magnetic field is produced when a current i flows in the six toroidal wires shown in the figure. In the cross-section shown here, gray represents current flowing into the page and black represents the current flowing out of the page. The vertical line is the axis of cylindrical symmetry.

scientific interest. At the end of his Nobel lecture, Wolfgang Paul showed how the mass of a neutron may be determined by weighing [40]. The preceding section of the lecture described a storage ring for confining neutrons.

The force confining the neutrons in the vertical (Z) direction is created by the sextupole magnetic field produced an electrical current i flowing in six toroidal rings (fig. 9). The vertical force on any neutron in the beam is given by

$$F = \mu \frac{\partial B}{\partial Z}\,, \tag{27}$$

where μ is the magnetic moment of the neutron and B is the magnetic induction in the vertical direction. In the storage ring,

$$\frac{\partial B}{\partial Z} \propto iZ,$$

so that the restoring force looks like eq. (22), where the "spring" constant is proportional to i. Thus the product $i\Delta Z$ is proportional to the restoring force. It is remarkable that the spring constant is so weak that the weight of the neutron lengthens the "spring" by a measurable amount. Setting $F = m_{\mathrm{N}}g$ and taking the accepted value for μ [6], we can use eq. (27) to predict that a gradient of 1.7 T/m is required to balance the weight, $m_{\mathrm{N}}g$, of a single neutron.

In the absence of the Earth's gravitational field, the neutron beam would be centred about the plane $Z = 0$. For the apparatus described by Paul, the neutron beam was displaced downward by 4.8 mm when $i = 50\,\mathrm{A}$. We would therefore expect that the downward displacement would be only 1.2 mm at $i = 200\,\mathrm{A}$ and that was indeed the case. Based on a fit to the data of displacement versus current, Paul and his colleagues inferred that the gravitational mass of the neutron is

$$m_{\mathrm{N}} = (1.63 \pm 0.06) \times 10^{-27}\,\mathrm{kg}.$$

This agrees with the inertial mass of the neutron to within 3%. The inertial mass is now known to a relative uncertainty of 1.7×10^{-7} [6] but it is remarkable that the gravitational mass can be measured at all. As Paul noted, the measurement of the gravitational mass in this way is only possible because electrical forces on the neutron are believed to be zero. At any rate, they are completely negligible, even compared to the feeble magnetic trapping force.

7. – Conclusion

We have presented the current state of mass metrology. The near future may bring a new definition of the kilogram, further improvements in analytical balances and mass comparators, and widespread use of NEMS devices in science and medicine. It is hoped that the overview presented here has encouraged the reader to participate in these endeavours or, at least, has provided a context in which to understand future developments.

REFERENCES

[1] Wilczek F., *Phys. Today*, **52** (Nov. 1999) 11.
[2] Wilczek F., *Phys. Today*, **53** (Jan. 2000) 13.
[3] Wilczek F., *Phys. Today*, **57** (Oct. 2004) 11.
[4] Wilczek F., *Phys. Today*, **57** (Dec. 2004) 10.
[5] Wilczek F., *Phys. Today*, **52** (Jul. 2005) 10.
[6] Mohr P. J. and Taylor B. N., *Rev. Mod. Phys.*, **77** (2005) 1.
[7] Quinn T. J., this volume, p. 59.
[8] Faller J. E., *Metrologia*, **39** (2002) 425.
[9] Misner C. W., Thorne K. S. and Wheeler J. A., *Gravitation* (Freeman, San Francisco) 1973.
[10] Kochsiek M. and Gläser M. (Editors), *Comprehensive Mass Metrology* (Wiley-VCH, Weinheim) 2000.
[11] Jones F. E. and Schoonover R. M., *Handbook of Mass MeasurementI* (CRC, Boca Raton) 2002.
[12] International Organization of Legal Metrology, *OIML R 111-1* (OIML, Paris) 2004. May be downloaded from `www.oiml.org`.
[13] Picard A., Fang H. and Gläser M., *Metrologia*, **41** (2004) 396.
[14] Wagner W. and Kleinrahm R., *Metrologia*, **41** (2004) S24.
[15] Picard A., *Metrologia*, **43** (2006) 46.
[16] Tanaka M., Girard G., Davis R., Peuto A. and Bignell N., *Metrologia*, **38** (2001) 301.
[17] Bureau International des Poids et Mesures, *The International System of Units*, 8th Edition (BIPM, Sèvres) 2006. May be downloaded from `www.bipm.org`.
[18] Davis R., *Metrologia*, **40** (2003) 299.
[19] Bich W., *Metrologia*, **40** (2003) 306.
[20] Gläser M., *Meas. Sci. Technol.*, **14** (2003) 433.
[21] Richard P., this volume, p. 499.
[22] Mana G., this volume, p. 519.
[23] Davis R. S., *Philos. Trans. R. Soc. London, Ser. A*, **363** (2005) 2249.

[24] Mills I. M., Mohr P. J., Quinn T. J., Taylor B. N. and Williams E. R., *Metrologia*, **43** (2006) 227.

[25] Davis R. S., *Metrologia*, **26** (1989) 75.

[26] Cumpson P. J. and Seah M. P., *Metrologia*, **31** (1994/95) 375.

[27] Speake C. C., *Proc. R. Soc. London, Ser. A*, **414** (1987) 333.

[28] Speake C. C., Quinn T. J., Davis R. S. and Richman S. J, *Meas. Sci. Technol.*, **10** (1999) 460.

[29] Quinn T. J., *Meas. Sci. Technol.*, **3** (1992) 141.

[30] Usher M. J., Buckner I. M. and Burch R. F., *J. Phys. E*, **10** (1977) 1253.

[31] Davis R. and Gläser M., *Metrologia*, **40** (2003) 339.

[32] Gläser M., *Metrologia*, **36** (1999) 183.

[33] Mana G., Palmisano C., Perosino A., Pettorruso S., Peuto A. and Zosi G., *Meas. Sci. Technol.*, **13** (2002) 13.

[34] International Organization of Legal Metrology, *OIML D 28* (OIML, Paris) 2004. May be downloaded from `www.oiml.org`.

[35] Dadson R. S., Lewis S. L. and Peggs G. N., *The Pressure Balance—Theory and Practice* (HMSO, London) 1982.

[36] Feynman R. P., "There's plenty of room at the bottom", 1959. Available on internet at `www.its.caltech.edu/~feynman/`.

[37] Roukes M., *Sci. Am.*, **285** (2001) 48.

[38] The Craighead Research Group: `www.hgc.cornell.edu`.

[39] Ilic R., Craighead H. G., Krylov S., Senartne W., Ober C. and Neuzil P, *J. Appl. Phys.*, **95** (2004) 3694.

[40] Paul W., *Rev. Mod. Phys.*, **62** (1990) 531.

[41] Sader J. E., *Calibration of Atomic Force Microscopopy: Cantilever Calibration*, in *Encyclopedia of Surface and Colloid Science*, edited by Somassundaran P. (Dekker Encyclopedias, online) 2006.

[42] Burnham N. A., Chen X., Hodges C. S., Matei G. A., Thoreson E. J., Roberts C. J., Davies M. C. and Tendler S. J. B., *Nanotechnology*, **14** (2003) 1.

[43] Pratt J. R., Smith D. T., Kramar J. A., Newell D. B. and Smith D. T., *Meas. Sci. Technol.*, **16** (2005) 2129.

[44] Sienknecht V. and Funck T., *Metrologia*, **22** (1986) 209.

[45] Choi I.-M, Kim M.-S., Woo S.-Y. and Kim S. H., *Meas. Sci. Technol.*, **15** (2004) 237.

[46] Cumpson P. J., Clifford C. A. and Hedley J., *Meas. Sci. Technol.*, **15** (2004) 1337.

[47] Lavrik N. V. and Datskos P. G., *Appl. Phys. Lett.*, **82** (2003) 2697.

[48] Lavrik N. V. and Datskos P. G., *Phys. World*, **17** (April 2004) 19.

[49] Ekinci K. L, YANG Y. T. and Roukes M. L., *J. Appl. Phys.*, **95** (2004) 2682.

[50] Ekinci K. L. and Roukes M. L., *Rev. Sci. Instrum.*, **76** (2005) 061101.

[51] Arlett J. L., Maloney J. R., Gudlewski B., Muluneh M. and Roukes M. L., *Nano Lett.*, **6** (2006) 1000.

Redefinition of the kilogram based on a fundamental constant

P. Richard

Federal Office of Metrology METAS - Lindenweg 50, 3003 Bern-Wabern, Switzerland

1. – Introduction

Mass and the corresponding definition and realisation of units has been of importance throughout human history. Like length or time, mass is a very familiar quantity in daily life as well as in science, technology and industry. The measurement of mass or weighing was and will always be a dominant activity in manufacturing processes and in trade of goods. As can be seen from the many related regulations, the unit of mass and its applications were always of political and economic relevance. The unit of mass *kilogram* was originally derived from one cubic decimetre of water at its maximum density and was realised for the first time in the 1790s in France [1] as a platinum cylinder standard, known as the *Kilogramme des Archives* [2]. Under the Metre Convention of 1875 it was decided that the new kilogram definition had to be consistent with the existing one. After a long development, fabrication and evaluation process for an adequately stable platinum-iridium standard, the *International Prototype Kilogram* ($\mathfrak{K}$) was deposited in 1889 at the Bureau International des Poids et Mesures (BIPM) in Sèvres near Paris. Unlike the other units in the Système International d'unités (SI) [3], the first international artefact definition for the unit of mass is still in use nowadays, but has been quite intensively debated and questioned during the past two decades. There are good reasons for this debate, mainly the annoying and not fully understood drifts [4] between $\mathfrak{K}$, its six official copies and the national prototype copies, and the dependency on the only remaining base unit defined by an artefact. Clearly, a material object like a mass standard does unavoidably have an exchange with the environment across its surface, and therefore, its mass is subject

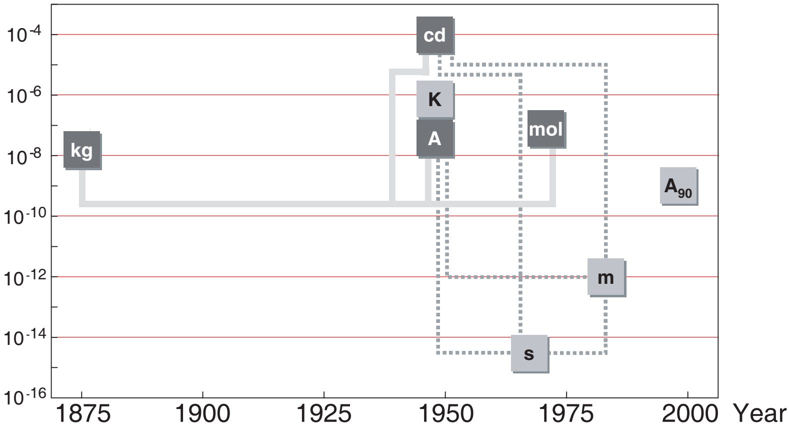


Fig. 1. – The kilogram and its relation to other base units with the year of adoption by the General Conference of Weights and Measures and today's accuracies of realisation. A_{90} is the representation of the ampere realised in terms of the Josephson and quantum Hall effect and the conventional values for $K_{\text{J-90}}$ and $R_{\text{K-90}}$ (see subsect. **2**˙1). For further development of the SI it is crucial to eliminate the dependency (gray lines) on the *International Prototype Kilogram* $\mathfrak{K}$, to define the unit of mass based on a fundamental constant and to allow kg realisations with relative uncertainties $\leq 10^{-8}$.

to small and not easily predictable changes. In view of the accuracies and consistencies within the SI needed these days, it is obvious that the *International Prototype Kilogram* of 1889 may not hold much longer as the definition for the mass unit. As shown in fig. 1, the possible drift of $\mathfrak{K}$ not only affects the mass unit but three other base units as well.

In the present situation we may well recall James Clerk Maxwell's visionary statement of 1870 [5]: *"If, then, we wish to obtain standards of length, time, and mass which shall be absolutely permanent, we must seek them not in the dimensions, or the motion, or the mass of our planet, but in the wavelength, the period of vibration, and the absolute mass of these imperishable and unalterable and perfectly similar molecules"*. During the past decades, base and other important units have been related to atomic or fundamental constants, in the very spirit of Maxwell's vision, with one exception, the mass unit. In the present state of research and technology, the mass unit clearly needs a new definition, based on a fundamental constant, preferably in a way as to allow various realisation methods [6,7]. Furthermore, consistency is required with the present definition and value of the mass unit within the SI, that is with $\mathfrak{K}$. Therefore, $\mathfrak{K}$ has to be measured and monitored by such new realisation methods with a relative accuracy better than the suspected drifts, *i.e.* at the level of 1 part in 10^8. In the following review we summarise the actual status of the unit of mass kilogram with regard to its role in the SI and the problems arisen from the possible drifts of the $\mathfrak{K}$. We discuss the two groups of methods potentially able to monitor $\mathfrak{K}$, namely counting atoms of known mass and electromechanical methods based on equivalence of electrical and mechanical power. We discuss the ion accumulation, watt balance as well as the voltage balance and superconducting levitation methods. A discussion of possible new kilogram definitions based either on a fixed value of the Planck constant or on a fixed value of the Avogadro constant and an outlook on their implementation and realisation concludes the paper.

2. – Present status of the kilogram

The present definition of the unit of mass in the SI is: *"The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram"* [3] is as old as the *International Prototype Kilogram* $\mathfrak{K}$ itself (see fig. 1). This artefact is made of a Pt/Ir alloy (90% Pt and 10% Ir) and has a cylindrical shape (height $\approx$ diameter $\approx$ 39 mm). A large majority of the member states of the Metre Convention possesses a copy of $\mathfrak{K}$, often referred to as the national prototype of the kilogram. National prototypes are directly used for the dissemination of the mass unit in each country. The initial mass determination of the first 40 prototypes using $\mathfrak{K}$ as reference was finished in 1889 with a standard uncertainty of 3 μg([1]).

2˙1. *Role of the kilogram in the SI.* – One of the major disadvantages of the definition of the unit of mass is that the amount of material constituting $\mathfrak{K}$ is subject to changes in time. Long-term observation of the relative mass drift between the international prototype and its copies [4,8] indicate that the long-term variation of the kilogram could be as much as 5 parts in 10^9 per year (see fig. 2).

A drifting mass unit also influences the electrical units, since they are linked to the kilogram through the ampere definition (see fig. 1). In 1990, international electrical reference standards based on the Josephson and the quantum Hall effects were introduced [9] by defining conventional values $K_{\text{J-90}}$ and $R_{\text{K-90}}$ for the Josephson constant and the von Klitzing constant, respectively. All electrical quantities can be measured in terms of these two conventional values. As a consequence, the worldwide uniformity and consistency of electrical measurements has improved by almost two orders of magnitude. However, due to the uncertainty of the values of the constants $K_{\text{J-90}}$ and $R_{\text{K-90}}$ of several parts in 10^7, results expressed in the 1990 "practical system" of electrical units may differ from the results in the SI by this amount. Moreover, the difference may change with time because of the possible drift of $\mathfrak{K}$.

This inconsistency is not yet a problem in most practical applications. However, to prepare the SI for the future needs of science and technology, a replacement of the kilogram based on a fundamental constant is needed. There is general agreement among metrologists that a replacement of the present kilogram definition should be considered when the experiments relating mass and fundamental constants reach a relative uncertainty of $\leq 1.0 \times 10^{-8}$.

[1] It should be noted that throughout this paper, the term uncertainty refers either to standard uncertainty or relative standard uncertainty.

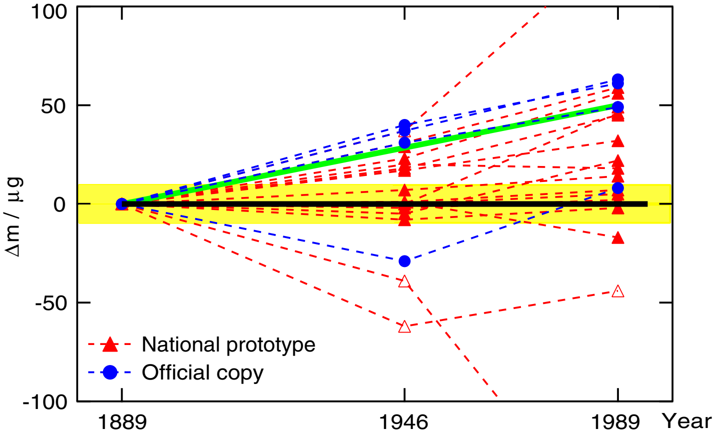


Fig. 2. – Relative change in mass of four out of the six official copies and seventeen national prototypes with respect to the mass of the *International Prototype Kilogram* 𝔎 [4]. The black horizontal line is for the assumed constant value of 𝔎 according to the definition. The solid gray line is the average mass drift (50 μg in 100 years) of the national prototypes. The gray band of ±10 μg represents the uncertainty to be reached to consider a new definition of the kilogram. Finally the open triangles are for national prototypes with low mass stability.

2˙2. *Results of periodic verifications of national prototypes.* – Since the official sanction of 𝔎 in 1889 and the distribution of the national prototypes, only three comparisons were undertaken altogether.

First Periodic Verification (1899-1911). An initial stability check was performed already ten years after the distribution of the national prototypes. Two out of the 25 verified prototypes had changed by as much as 50 μg. The changes of the others were found insignificant in comparison to the uncertainty of measurement of 10 μg.

Second Periodic Verification (1947-1954). Some comparisons between 𝔎 and its six official copies demonstrated the necessity to develop a procedure to reproducibly clean the prototypes. The BIPM cleaning and washing procedure for the Pt/Ir prototypes was developed during the Second World War [10]. Before the second periodic verification, all prototypes were cleaned and washed. Four prototypes out of the first 40 gained more than 30 μg since the first verification. Unfortunately the effect of cleaning and washing was not studied and the uncertainty of measurement was not reported.

Third Periodic Verification (1988-1992). During this verification, the effect of cleaning and washing was studied in detail. It was possible to determine the short-term and long-term re-contamination rate of the prototypes. Based on these studies, the definition of the kilogram was completed the following way: *"The kilogram is equal to the mass of the international prototype kilogram immediately after cleaning and washing using the BIPM method"*. The detailed results of the third periodic verification, which included a total of 52 Pt/Ir prototypes, were described by Girard [4]. The standard uncertainty for the mass of each national prototype was 2.3 μg. The third periodic verification confirmed that the mass of the national prototypes and the six official copies tends to increase over the time with respect to the mass of $\mathfrak{K}$. More detail can be found in a recent comprehensive review of the SI unit of mass by Davis [11].

Figure 2 shows the relative mass change of 23 prototypes with respect to the mass of $\mathfrak{K}$. Up to now no explanation for this average relative increase in mass of the official copies and the national prototypes was found. The doubt will remain until a success in the on-going experiments described below is achieved.

3. – Counting atoms using ion accumulation

Among the approaches to define and realise a unit of mass that is no longer based on an artefact, the idea of an atomic mass standard is probably the most intuitive one. This is quite in line with Maxwell's [5] idea more than 130 years ago, where he suggested to seeking molecular quantities in order to get "permanent standards" of measurement. The underlying concept of atomic mass dates back to times even before Maxwell. But only in 1971 the General Conference on Weights and Measures officially defined the atomic mass unit m_{u} as 1/12 of the ^{12}C nuclide's mass. Today's accepted value of m_{u} is the result of the 1998 CODATA [12] least-squares adjustment with a relative uncertainty of 7.9×10^{-8}, which in turn is mainly caused by the uncertainty of the Planck constant measurement with the watt balance at the National Institute of Standards and Technology (NIST) in the USA [13] in 1998 (see also sect. **5**). In this line of thoughts, the kilogram would be defined as a fixed number of "elementary masses". For the practical realisation, the mass of the chosen atom, particle or similar entities has to be known in terms of the kilogram, and in addition, the number of atoms in a macroscopic body has to be measured. Two approaches are pursued today. The first involves the determination of the number of atoms in a large silicon crystal by measuring the unit cell volume, the macroscopic density and the isotopic composition (for more details on the Avogadro project see the paper from G. Mana in this volume [14] or the comprehensive review of the work carried out so far by Becker [15]). In the second approach, a beam of ionised atoms is collected and weighed. The number of ions is derived from the electric current leaving the collector.

Already in the 1960s research was undertaken to accumulate and count ions by an electrochemical method: the electrolysis. The quantity of interest in this case is the Faraday constant, F, which determines the number of moles of a substance X that the passage of the amount of electric charge $Q = I \cdot t$ will deposit on an electrode or dissolve from it during electrolysis. The Faraday constant may be expressed as (see, *e.g.*, [12])

$$F = \frac{M(X)}{z \cdot m(X)} \cdot I \cdot t, \tag{1}$$

where $M(X)$ is the molar mass of the entity X, z denotes the charge number, and $m(X)$ is the mass change of the electrode. The most accurate measurement of F by an electrochemical method was carried out at NIST using a silver dissolution coulometer.

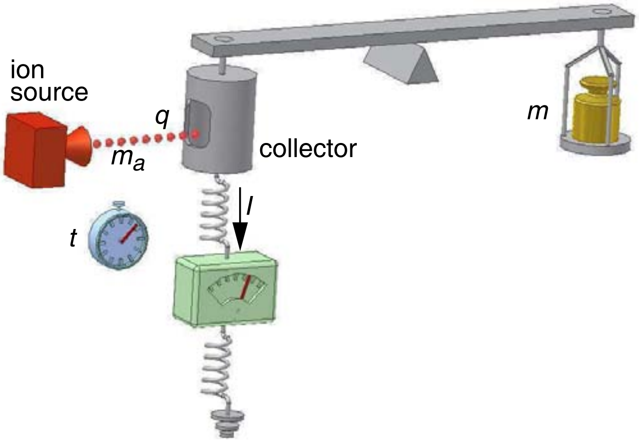


Fig. 3. – Experimental concept of determining a mass based on atomic mass unit by ion accumulation. m_a/q and m/Q are the mass-to-charge ratio of the single ion and the total of accumulated ions, respectively, I is the electric current measured over the time t (see [16]).

In 1980, the experiment achieved a relative uncertainty of 1.3×10^{-6} [17]. By that time, however, the method had reached basic limitations and was, therefore, abandoned.

The idea of counting atoms of known or defined mass by means of ion accumulation in vacuum was proposed in the early 1990s by Gläser (Physikalisch-Technische Bundesanstalt (PTB)) [18]. Because the classic way of counting atoms up to a weighable mass with an electronic counter would need millions of years, the author proposed to use an ion beam and integrate its electric current I over the accumulation time, that is, the total electric charge Q. As the ratio between the accumulated mass and its total charge m/Q is the same as for a single ion m_a/q, a measurement of q/Q immediately leads to m/m_a and therefore to number of accumulated ions N.

As shown in fig. 3, the total electric charge Q carried by the ion beam (typically ^{197}Au or ^{209}Bi) is determined by measuring its electric current I over the accumulation time t. With an ion current of 10 mA, an accumulation of 10 g could be achieved in less than 6 days. In the PTB experiment [16, 19] —the only one of its kind so far— the current I is determined by the voltage drop over a resistance in the current path. The voltage is traceable to the Josephson voltage standard and the resistance to the quantised Hall standard. This leads to a simple ratio between the ion mass m_a and the total collected mass m

$$m_a/m = 2z \Big/ \left(n_1 n_2 \int f \,\mathrm{d}t \right), \tag{2}$$

where z is the charge state q/e of the single ion, n_1 the quantum Hall plateau number and n_2 the Josephson step number. Equation (2) is only valid under ideal conditions.

That is to say for sufficiently small ion energies and for an absolutely pure ion beam. The frequency f and the time t are measured based on the same atomic clock and, therefore, an eventual offset from the SI second is irrelevant. The practical realisation involves a quite big and complicated experimental setup. The ion source (oven, plasma discharge chamber and high voltage extractor) generates an ion beam in vacuum. A magnetic quadrupole focuses the ion beam and a dipole magnet is used as mass to charge separator. At the end, the ions are accumulated in a collector with an optimised shape in order to minimise the loss of atoms by sputtering. The total mass and total charge determination are taking place in the collector. Despite the simplicity of the idea, the PTB experiment is quite demanding according to their recent report [19]. The major difficulties are the loss of particles due to sputtering and reflection, the implantation of foreign particles like electrons and residual gas molecules in the collector, the focusing of the decelerated ion beam, the accumulation of sufficient mass, and finally the accuracy of current and mass determinations. First results with a $^{197}\mathrm{Au}^{1+}$ beam were obtained with a relative uncertainty of 1.5% for the mass of the gold atom and the deduced value for m_u is in agreement with the one published in [12]. The experiment has proved to be working in principle, but is still at an early stage with regard to the projected relative uncertainty of $\leq 1 \times 10^{-8}$. A significant reduction of the uncertainty has been envisaged by improving the ion beam current and optics, the mass comparator, angular distribution of backscattered particles, detection of foreign particles, etc.

4. – Electromechanical methods

The second approach towards a new definition of the kilogram is to use electrical or electromechanical methods to relate the unit of mass to fundamental constants at a macroscopic scale. In this case, the link is established by comparison of electrical and mechanical power. The gravitational force acting on a test mass is compensated by a Lorenz force in the watt balance experiment, by the force acting on a diamagnetic body in the superconducting levitation experiment and by an electrostatic force in the voltage balance experiment. In all three experiments, the electrical parameters are directly measured in terms of the quantised Hall resistance and the Josephson voltage standard, and are, thus, related to the Planck constant.

4˙1. *The watt balance experiment.* – The concept of the moving coil watt balance was proposed almost 30 years ago by Kibble [20]. A comprehensive review on the existing watt balance experiments was published in 2003 by Eichenberger *et al.* [21]. The experiment is performed in two parts within the same experimental set-up (see fig. 4). The first part is a static force compensation where a coil is suspended from one arm of a balance. The coil is immersed in a stationary horizontal magnetic field of flux density B. The current I in the coil exerts a force on the conductor given by

$$\vec{F} = I \cdot \oint \mathrm{d}\vec{l} \times \vec{B}\,, \tag{3}$$

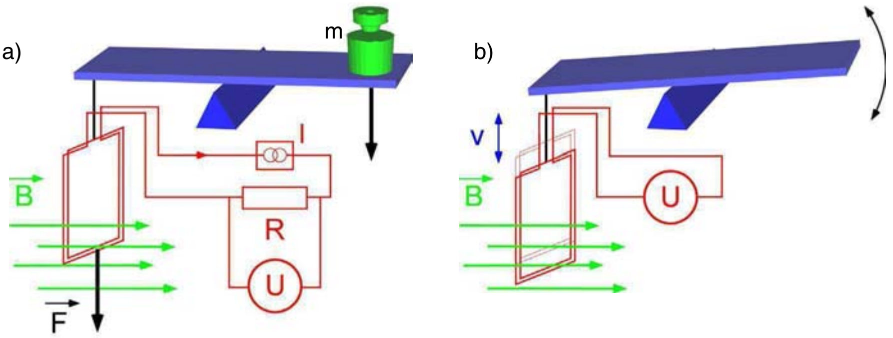


Fig. 4. – Principle of the watt balance experiment. a) Static mode: the electromagnetic force (F) acting on the current-carrying coil (I) is balanced against the weight (mg) of the test mass. b) Dynamic mode: the coil is moved at constant velocity (v) in the vertical direction through the magnetic field (B) and the induced voltage measured in comparison with a Josephson voltage standard.

where l is the conductor length of the coil. The vertical component F_z of this force is balanced against the weight of the test mass m and we have $F_z = mg$, where g is the local acceleration due to gravity. In the second part of the experiment, the coil is moved at a constant velocity v in the vertical direction through the magnetic field and the voltage U induced across the coil is measured, being

$$U = \oint (\vec{v} \times \vec{B}) \cdot \mathrm{d}\vec{l} = -\oint (\mathrm{d}\vec{l} \times \vec{B}) \cdot \vec{v}\,. \tag{4}$$

In case of a strictly vertical movement of the coil, the integrals $\oint(\mathrm{d}\vec{l} \times \vec{B})$ are the same in eqs. (3) and (4) at the location of the weighing. The elimination of these terms then leads to

$$U \cdot I = m \cdot g \cdot v. \tag{5}$$

The experiment thus allows the virtual comparison of the watt realised electrically (left-hand side of the equation) to the watt realised mechanically. The voltage U can be measured against a Josephson voltage standard. This is most conveniently done using a programmable Josephson voltage standard [22].

Using the expressions of the Josephson frequency, the quantum number of the voltage step, the current, the voltage drop across a resistance calibrated against a quantum Hall resistance standard, and the quantised Hall resistance, eq. (5) can be rewritten as

$$m = C\,\frac{f_{\mathrm{J}} \cdot f'_{\mathrm{J}}}{g \cdot v} \cdot h, \tag{6}$$

where C represents different calibration constants, f_{J} the Josephson frequency measured during the dynamic phase and f'_{J} the Josephson frequency measured during the static phase and h the Planck constant. The watt balance, thus, allows us to express the test mass in terms of the metre, the second and the Planck constant. One of the major advantages of the experiment is that neither the geometry of the coil nor the flux density produced by the source magnet have to be known. Moreover, only virtual electrical and mechanical energy are related. This means that in contrast to the superconducting magnetic levitation described below, real energy dissipation does not enter into the basic equation of the experiment.

The velocity signal comes from a carefully designed interferometer [23] and either this signal or the induced voltage can be used in a regulation loop to control the motion. Since the sign of the induced voltage is reversed when the direction of the motion is inverted, voltage offsets in the electrical circuit can be removed when up and down measurement results are averaged.

In the static mode of the experiment, a current I flowing in the coil generates a Lorentz force to balance the mechanical force F produced by the test mass in the gravitational field. In practice, the balance is underloaded by half of the value of the test mass. A first weighing with a current producing the force needed to balance the system without test mass is followed by a second one, where the sign of the current is reversed and the test mass placed on the balance. These currents are controlled to keep the balance at the equilibrium position and the values are measured with the help of a standard resistor, periodically calibrated against the quantum Hall resistance standard, and a voltage reference.

The mass m of the test mass is determined in air by the classical methods of mass metrology using a mass comparator. The test mass is then directly traceable to national prototypes of the kilogram. As watt balances are operated under vacuum, the mass of the test mass should be also known under vacuum. The immediate advantage of the vacuum measurement is the suppression of the air buoyancy correction. The main disadvantage is the discontinuity of the mass scale between air and vacuum due to the physical and chemical adsorbed layers at the surface of the test mass. Different methods to experimentally overcome the discontinuity of the mass scale between air and vacuum were proposed by Kibble and Robinson [24,25].

Finally, an accurate determination of the absolute value of the gravitational acceleration g next to the experiment and synchronously to the static mode measurement is required to get the expression of the mechanical force $F = m \cdot g$.

The very first moving coil apparatus was developed at the National Physical Laboratory (NPL) in the UK, based on Kibble's proposal of 1976 [20]. The final result of the initial set-up with a relative standard uncertainty of 2×10^{-7} was published in 1990 [26]. An improved apparatus, presented the same year reached a short-term reproducibility in the order of 1 part in 10^8 [27], and a new result for the Planck constant may be expected in the near future. The second watt balance was built at the NIST in USA. The first results published in 1989 [28,29] had a relative standard uncertainty of 1.3×10^{-6}. Further improvements [30-32] led to the result reported in 1998 with a relative standard uncertainty of 9×10^{-8} [13]. Finally the last improvements led to the result reported in 2005 with a relative standard uncertainty of 5.2×10^{-8} [33,34]. The NIST group is expecting a reduction of this value down to 3.0×10^{-8} or 2.0×10^{-8} during the next years. The NIST watt balance is currently the biggest and the only one using a superconducting magnet.

More than 25 years after the first proposal, the Federal Office of Metrology (METAS), initiated the third existing watt balance project in 1997 [35]. The new design features implemented in the METAS watt balance consist mainly in a clear separation between the static and dynamic phase of the experiment and a drastic size reduction using a 100 g mass. The Swiss watt balance is the smallest one. The METAS apparatus is fully assembled and after having passed the testing and evaluation phases the experiment is providing experimental data on a daily basis [36]. First results are expected soon. The fourth project was initiated in 2000 at the Laboratoire National de Métrologie et d'Essais (LNE) in France [37]. The main idea of LNE is to move the mass comparator and the coil during the dynamic mode. The French group will have an operational moving beam watt balance in a new laboratory by the end of 2007. Finally, in 2002 the BIPM decided to join the club with the proposal of a single mode cryogenic watt balance. One of the main original ideas of BIPM is to perform both modes (static and dynamic) simultaneously.

Each of the five existing watt balances is based on the same principle but each experiment is totally different in its practical realisation. This is the best way to produce independent results and solve all possible sources of systematic errors. In addition to that, it is not excluded that some new projects will be initiated in a near future.

4·2. *Other electrical methods*

4·2.1. The superconducting magnetic levitation. This experiment makes use of the force acting on a body with diamagnetic properties in a non-uniform magnetic field. The idea was first brought up by Sullivan and Frederich [38] as a possibility to realise the ampere. When a superconductor in the Meissner state is introduced into the field of a coil with decreasing magnetic flux Φ in the vertical direction, a stable levitation of the body can be obtained (see fig. 5(a)). The energy equation of the system can then be written as

$$I \cdot U \cdot \mathrm{d}t = I \cdot \mathrm{d}\Phi = \mathrm{d}A + \mathrm{d}W, \tag{7}$$

where A is the work done by the field to increase the gravitational energy of the body and W the magnetic-field energy. If the levitated body has ideal diamagnetic properties and the coil circuit is superconducting as well, the energy terms are given by

$$W = \frac{1}{2}\Phi \cdot I, \qquad A = m \cdot g \cdot z. \tag{8}$$

Considering two equilibrium positions with heights z_l and z_h, where the subscript h and l denote high and low position, respectively, the energy difference takes the form

$$\int_{\Phi_l}^{\Phi_h} I \cdot \mathrm{d}\Phi = \frac{1}{2}(\Phi_h I_h - \Phi_l I_l) + mg(z_h - z_l). \tag{9}$$

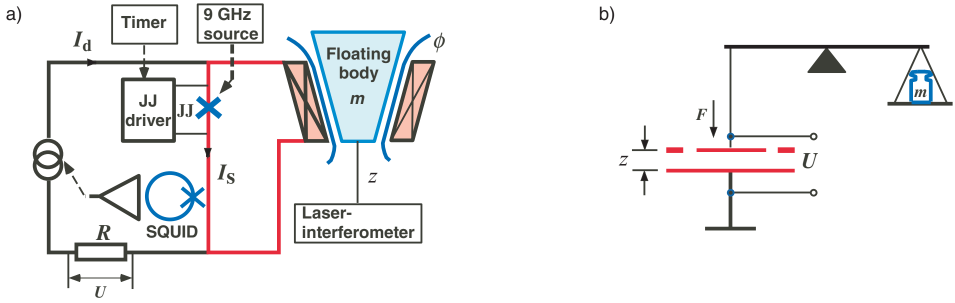


Fig. 5. – a) Principle diagram of the superconducting magnetic levitation. A superconducting body is floating in a magnetic flux, produced by a superconducting coil. b) Principle diagram of the voltage balance.

In practice the experiment can be realised as shown in fig. 5(a). The superconducting coil is driven by a supply circuit which is controlled by a SQUID ammeter so that the drive current I_d corresponds to the coil current I_s. The drive current is determined by the voltage drop across a resistance standard calibrated in terms of the quantised Hall resistance. When the Josephson junction in the coil circuit is biased on the first step for a time interval t, the flux in the coil is increased by $\Delta\Phi = f_J t \Phi_0$, where f_J is the Josephson frequency, and $\Phi_0 = h/2e$ the magnetic flux quantum. If $\Delta\Phi$ is large enough, the superconducting body of mass m is levitated and reaches the equilibrium position z_l which is measured by laser interferometry. By repeating the process, a series of equilibrium positions can be obtained, where eq. (9) describes the energy between any two positions. Since the flux change can be expressed in units of Φ_0 and the coil current can be measured using the Josephson and the quantum Hall effect, the experiment relates the mass of the floating body to the Planck constant.

The superconducting magnetic levitation has some major metrological difficulties to overcome. The most important problems are: All unwanted energy expenditure, *e.g.* due to horizontal force components on the trajectory of the levitated object or distortion of the object under the force of levitation, have to be avoided. The floating body has to be a perfect diamagnet and its mass has to be known in low-temperature environment.

The approach of the superconducting magnetic levitation has been pursued at the National Research Laboratory of Metrology (NRLM, now NMIJ) in Japan [39, 40] and the D. I. Mendeleyev Research Institute of Metrology (VNIIM) in Russia [41]. The NRLM group has reached a resolution of 1 part in 10^6 in its experiment [42]. A new set-up which should reduce some of the systematic errors was proposed in 2001 [43]. In the same year, a design study for a magnetic levitation system was presented by the Centre for Metrology and Accreditation (MIKES) in Finland [44]. The MIKES group is developing a cryogenic calorimeter to measure the energy losses due to the non-ideal diamagnetic properties of the levitated body.

4.2.2. The voltage balance. The principle of this approach is illustrated in fig. 5(b). The electrostatic force acting between the plates of capacitance C is compared with the weight mg of the test mass m, where g is the gravitational acceleration. The movable plate of the capacitor is suspended from the balance. In the equilibrium position of the balance, the forces are connected by the relationship

$$m \cdot g = \frac{1}{2} U^2 \frac{\partial C}{\partial z}\,, \tag{10}$$

where U is the voltage across the capacitor and $\partial C / \partial z$ the capacitance gradient in the vertical direction. The measurement of the voltage is performed against a Josephson voltage standard and the capacitance change is expressed in terms of the quantised Hall resistance. In this way, a link between the test mass and the Planck constant is established. In a typical set-up (see [45] for a review), the voltage needed is around 10 kV and the test mass is a few grams.

Using this approach, Funck and Sienknecht from the PTB [46] reached a relative standard uncertainty of 6.3×10^{-7} in the determination of h [12]. The experiment was also carried out at the University of Zagreb [47]. A relative uncertainty of 3.5×10^{-7} in the determination of the volt was obtained in 1987-1988 [9]. Subsequently, several systematic errors were found by Bego *et al.* in the set-up which led to improvements [47] and the proposition for a new 100 kV voltage balance [48]. To our knowledge, however, this work is not carried on, at least at the moment.

With the present techniques, the voltage balance approach does not promise to reach an uncertainty below 1 part in 10^7. The main problems are the high voltage required in the experiment, the voltage and frequency dependence of the capacitance and its mechanical imperfections.

5. – Possible new definitions of the kilogram and conclusions

Mass is an important physical property and the selection of the mass as a base quantity in the SI is quite logical. But when it comes to a definition and realisation of the base unit of mass, present and future requirements on the stability and reproducibility in the SI are clearly beyond the capabilities of an artefact like the *International Prototype Kilogram* $\mathfrak{K}$. Therefore, a new definition based on a fundamental constant is urgently needed. A quite controversial proposal for an immediate redefinition of the kilogram was recently made by Mills [49]. This proposal was followed by a lot of reactions, not only from the mass community, and by numerous other proposals [50-53].

According to Becker [53] the following general criteria are widely accepted and have to be fulfilled before the proposal of a new definition of a SI unit:

- continuity between the realisations of the old and new definition
- realisation of the unit with a smaller uncertainty of measurement (or at least the same)

- better stability of the quantity considered
- consistency and coherence with the other SI base units
- "continuous" realization of the unit anywhere and at anytime
- based on commonly accepted laws in physics
- the concept of the proposal has to be clear and easy to understand.

The different recent proposals for a possible new definition of the kilogram are summarised below. Each proposal only partly takes into account the stated criteria.

5˙1. *Definitions that fix the value of the Avogadro constant.* – The approach based on counting atoms could lead to a definition of the form [49, 54]: *"The kilogram is the mass of exactly* $5.018\,451\,272\,5 \times 10^{25}$ *unbound* ^{12}C *atoms at rest and in their ground state".* In this definition, the numerical value is derived from $m_\mathrm{u} = M_0/N_\mathrm{A}$, where m_u is the atomic mass unit and M_0 denotes the molar mass constant which has the value $10^{-3}\,\mathrm{kg\,mol^{-1}}$ by definition. The numerical value is, thus, $(N_\mathrm{A}/12) \cdot 10^3\,\mathrm{mol}$. This formulation leads to a new definition of the mole. Using the relation between the mass of the carbon-12 atom and the Avogadro constant N_A through the definition of the mole: "*The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in* 0.012 kg *of carbon 12*", another possible wording of the definition could read as follows: *"The kilogram is the mass of a body at rest such that the value of the Avogadro constant is exactly* $6.022\,141\,527 \times 10^{23}$ *inverse mole".* Both wordings fix the value of the Avogadro constant explicitly.

In his controversial proposal, Mills put forward some advantages of the above possible definitions. The first advantage is the simultaneous redefinition of the mole in a simpler manner. With such a definition, the value of the unified atomic mass unit (Dalton, Da) is fixed. With this unified atomic mass unit the uncertainty of energy equivalence (with h, e and c) is eliminated. Together with the redefinition of the ampere (with a fixed value of the elementary charge e) the uncertainty of energy equivalence relations (with the atomic mass constant, e and c) is eliminated and the value of the Faraday constant F would be exactly known ($F = N_\mathrm{A} \cdot e$). Similarly with the redefinition of the kelvin (fixed value of the Boltzmann constant k) the uncertainty of energy equivalence relations (with the atomic mass constant, k and c) is eliminated and the molar gas constant R would have an exact value ($R = N_\mathrm{A} \cdot k$). Finally, with the definition based on counting atoms (definitions that fix the value of the Avogadro constant), it is not trivial to imagine independent realizations anywhere and at anytime.

5˙2. *Definitions that fix the value of the Planck constant.* – In the second line of experiments comparing electrical and mechanical power, a link between the kilogram and the Planck constant is established. As the Planck constant plays a unique role among the fundamental constants, both as quantum of action and as a factor of proportionality in many equations, it would be a natural choice to fix the value of h and to link the kilogram to this value using experiments like the watt balance.

According to the propositions of Taylor and Mohr [54] or more recently of Mills [52], the first possible wording for a new definition of the kilogram could read as follows: *"The kilogram is the mass of a body whose equivalent energy is equal to that of a number of photons whose frequencies sum to exactly* $[(299\,792\,458)^2/662\,606\,93] \times 10^{41}$ *hertz".* This definition is based on the well-known Einstein relation $E = mc^2$, where c is the speed of light fixed with the definition of the metre, and the relation $E = h\nu$ valid for the energy of photons.

The second wording could read: *"The kilogram is the mass of a body whose de Broglie-Compton frequency is equal to exactly* $[(299\,792\,458)^2/(6.626\,069\,3 \times 10^{-34})]$ *hertz".* This second definition is based on assigning a specific value to the Compton frequency of a body with a mass of 1 kg. The third possible definition simply states that the unit is defined by assigning to the Planck constant an exact specified value as follows: *"The kilogram, unit of mass, is such that the Planck constant is exactly* $6.626\,069\,3 \times 10^{-34}$ *joule second".*

Together with the initial proposal, Mills developed the main reasons for preferring a definition of the kilogram based on a fixed value of the Planck constant h. The first one is that the Planck is the fundamental constant of quantum mechanics. With such a definition, the uncertainty of energy equivalence relations (with h and the speed of light c) is eliminated. Together with the redefinition of the ampere (with a fixed value of the elementary charge e) the uncertainty of energy equivalence relations (with h, e and c) is eliminated. Similarly with the redefinition of the kelvin (fixed value of the Boltzmann constant k) the uncertainty of energy equivalence relations (with k and h, or with k, h and c) is eliminated. If h and e are both fixed, this will also fix the value of the Josephson constant $K_\mathrm{J} = 2e/h$ and the von Klitzing constant $R_\mathrm{K} = h/e^2$. This would lead to a simplification and an increase of accuracy for all electrical units and lead to the elimination of the conventional electrical units. Finally, with such a definition, watt balances could be used to realize the unit of mass "continuously" and simultaneously in different laboratories.

Note that it is also possible to use a definition based on a fixed value of the Planck constant for the Avogadro route using eq. (11). According to Becker [53] such a combined definition could read as follows: "*The kilogram is* $(6.022\,141\,5 \times 10^{23}/0.012)$ *times the rest mass of a particle whose creation energy equals that of a photon whose frequency is* $[(0.012/6.022\,141\,5 \times 10^{23}) \times (299\,792\,458)^2/(6.626\,0693 \times 10^{-34})]$ *hertz*".

5·3. *Planck and Avogadro constants.* – The Planck and Avogadro constants are related by

$$h = \frac{cA_\mathrm{r}(e)M_0\alpha^2}{2R_\infty N_\mathrm{A}}\,, \tag{11}$$

where $A_\mathrm{r}(e)$ is the relative atomic mass of the electron, α is the fine-structure constant, and R_∞ is the Rydberg constant. The combined uncertainty of this group of constants is below 1 part in 10^8. The value of the molar mass constant is $M_0 = 10^{-3}\,\mathrm{kg\,mol^{-1}}$ exactly.

An overview on the present status of the experiments aiming at a new definition of the kilogram can be gained by looking at the published results for the Planck constant. In fig. 6, all results with a relative standard uncertainty below 1×10^{-6} are shown. With

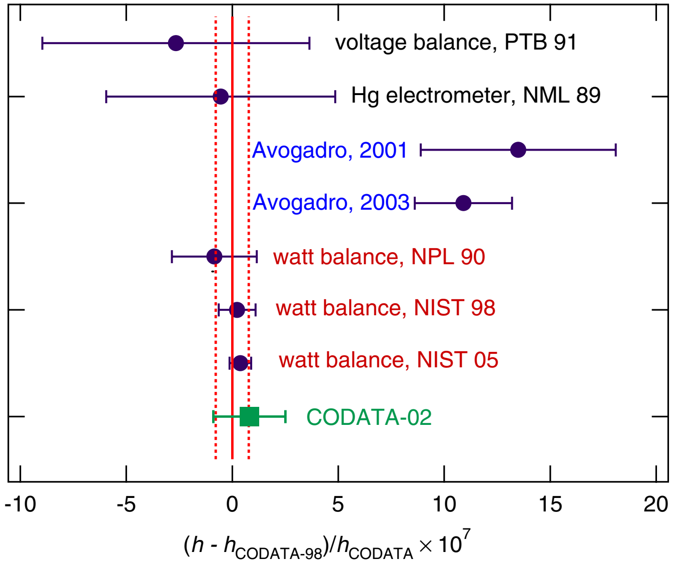


Fig. 6. – Experimental values for the Planck constant with CODATA-98 and CODATA-02 values. The values labeled "Avogadro 2001" and "Avogadro 2003" are determined from the latest published values of the Avogadro constant [55-57].

the exception of the results deduced from the Avogadro experiments, the values are taken from [12]. Due to improvements in the analysis, they may, in some cases, differ from the data first published by the experimenters.

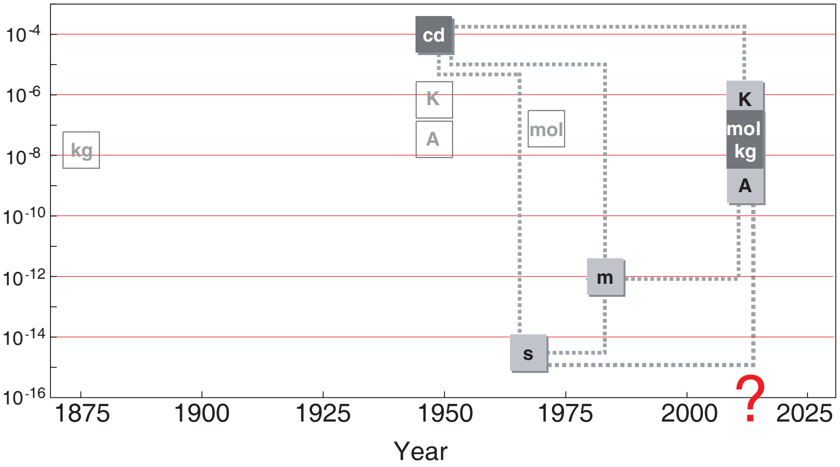


Fig. 7. – The kilogram and its relation to other base units with the possible accuracies of realisation towards 2011 or later.

The relative differences between h calculated using eq. (11) and the CODATA value are $(1.3\pm0.5)\times10^{-6}$ and $(1.1\pm0.3)\times10^{-6}$, respectively. This may point to an unresolved systematic error in one of the experiments.

5˙4. *A decision whose time will come.* – The research work outlined in this paper, undertaken during the past two decades to replace the artefact definition of the base unit kilogram, is impressive, both in extent and in variety. While the options for a new definition based on a fundamental constant are already clear today, the various experiments are at quite different stage of development and none has achieved the necessary accuracy level so far. From this point of view the Avogadro and the watt balance experiments are more advanced than the other routes and important results may be expected within the next few years. But a new definition may only be considered when experimental agreement is reached at the suspected relative uncertainty level of the *International Prototype Kilogram* $\mathfrak{K}$, that is at about 2.0×10^{-8}. This may be the case towards 2011 or later (see fig. 7). Furthermore, the new definition should preferably not refer to a particular atom, but to a fundamental constant only, like the mass of an elementary particle or the Planck constant h. Considering the six proposed new definitions reported here, there is still quite a big challenge to chose the one which is the clearest and easiest to understand by the general public.

* * *

The author thanks particularly W. Schwitz and B. Jeckelmann from METAS for their approval to publish this paper as an update and an extension of the review "Towards a new kilogram definition based on a fundamental constant" [58].

REFERENCES

[1] Smeaton W., The foundation of the metric system in France in the 1790s, *Platinum Metals Rev.*, **44** (2000) 125.

[2] Moreau H., *Le système métrique* (Chiron, Paris).

[3] *The International System of Units* (Bureau international des poids et mesures, BIPM) 2006, 8th Edition.

[4] Girard G., The third periodic verification of national prototypes of the kilogram (1988-1992), *Metrologia*, **31** (1994) 317.

[5] Maxwell J., *Report of the British Association for the Advancement of Science*, **40** (1870) 215.

[6] Taylor B. N., The possible role of the fundamental constants in replacing the kilogram, *IEEE Trans. Instrum. Meas.*, **40** (1991) 86.

[7] Kose V., Siebert B. R. L. and Wöger W., General principles for the definition of the base units in the SI, *Metrologia*, **40** (2003) 146.

[8] Quinn T. J., The kilogram: the present state of our knowledge, *IEEE Trans. Instrum. Meas.*, **40** (1991) 81.

[9] Taylor B. N. and Witt T. J., New international electric reference standards based on the Josephson and quantum Hall effects, *Metrologia*, **26** (1989) 47.

[10] Girard G., *The washing and cleaning of kilogram prototypes at the BIPM*, BIPM report.

[11] Davis R., The SI unit of mass, *Metrologia*, **40** (2003) 299.

[12] Mohr P. J. and Taylor B. N., CODATA recommended values of the fundamental physical constants: 1998, *Rev. Mod. Phys.*, **72** (2000) 351.

[13] Williams E. R., Steiner R. L., Newell D. B. and Olsen P. T., Accurate measurement of the Planck constant, *Phys. Rev. Lett.*, **81** (1998) 2404.

[14] Mana G., Towards an atomic realization of the kilogram, this volume, p. 519.

[15] Becker P., History and progress in the accurate determination of the Avogadro constant, *Rep. Prog. Phys.*, **64** (2001) 1945.

[16] Gläser M., Tracing the atomic mass unit to the kilogram by ion accumulation, *Metrologia*, **40** (2003) 376.

[17] Bower V. E. and Davis R. S., The electrochemical equivalent of pure silver — a value of the Faraday constant, *J. Res. Natl. Bur. Stand.*, **85** (1980) 175.

[18] Gläser M., Proposal for a novel method of precisely determining the atom mass unit by accumulation of ions, *Rev. Sci. Instrum.*, **62** (1991) 2493.

[19] Becker P. and Gläser M., Avogadro constant and ion accumulation: steps towards a redefinition of the SI unit of mass, *Meas. Sci. Technol.*, **14** (2003) 1249.

[20] Kibble B. P., *A measurement of the gyromagnetic ratio of the proton by the strong field method* in *Atomic Masses and Fundamental Constants*, edited by Sanders J. H. and Wapstra A. H., Vol. **5** (Plenum Press, New York) 1976, pp. 545-551.

[21] Eichenberger A., Jeckelmann B. and Richard P., Tracing Planck's constant to the kilogram by electromechanical methods, *Metrologia*, **40** (2003) 356.

[22] Burroughs C., Benz S. P., Harvey T. E. and Hamilton C. A., 1 Volt DC programmable Josephson voltage standard, *IEEE Trans. Appl. Supercond.*, **9** (1999) 4145.

[23] Courteville A., Salvadé Y. and Dändliker R., High-precision velocimetry: optimization of a Fabry-Perot interferometer, *Appl. Opt.*, **39** (2000) 1521.

[24] Kibble B. P., Comparing a mass in vacuum with another in air by conventional weighing, *Metrologia*, **27** (1990) 157.

[25] Robinson I. A., Comparing in-air and in-vacuum mass standards without buoyancy corrections via in-vacuum weighing, *Metrologia*, **27** (1990) 159.

[26] Kibble B. P., Robinson I. A. and Belliss J. H., A realization of the SI watt by the NPL moving-coil balance, *Metrologia*, **27** (1990) 173.

[27] Robinson I. A. and Kibble B. P., *Progress in relating the kilogram to Planck's constant with the NPL watt balance* in *Conference on Precision Electromagnetic Measurements* (CPEM, Conference Digest) 2002, pp. 574-575.

[28] Olsen P. T., Elmquist R. E., Philips W. D., Williams E. R., Jones G. R. and Bower V. E., A measurement of the NBS electrical watt in SI units, *IEEE Trans. Instrum. Meas.*, **38** (1989) 238.

[29] Cage M. E., Dziuba R. F., Elmquist R. E., Field B. F., Jones G. R., Olsen P. T., Phillips W. D., Shields J. Q., Steiner R. L., Taylor B. N. and Williams E. R., NBS determination of the fine-structure constant, and of the quantized Hall resistance and Josephson frequency-to-voltage quotient in SI units, *IEEE Trans. Instrum. Meas.*, **38** (1989) 284.

[30] Olsen P. T., Tew W. L., Williams E. R., Elmquist R. E. and Sasaki H., Monitoring the mass standard via the comparison of mechanical to electrical power, *IEEE Trans. Instrum. Meas.*, **40** (1991) 115.

[31] Steiner R. L., Gillespie A. G., Fujii K., Williams E. R., Newell D. B., Picard A., Stenbakken G. N. and Olsen P. T., The NIST watt balance: progress toward the monitoring of the kilogram, *IEEE Trans. Instrum. Meas.*, **46** (1997) 601.

[32] Steiner R., Newel D. N. and Williams E., Details of the 1998 Watt Balance Experiment Determining the Planck Constant, *J. Res. Natl. Inst. Stand. Technol.*, **110** (2005) 1.

[33] Steiner R. L., Williams E. R., Newel D. B. and Liu R., Towards an electronic kilogram: an improved measurement of the Planck constant and electron mass, *Metrologia*, **42** (2005) 431.

[34] Steiner R. L., Newel D. B., Williams E. R., Liu R. and Gournay P., The NIST Project fort he Electronic Realization of the Kilogram, *IEEE Trans. Instrum. Meas.*, **54** (2005) 846.

[35] Beer W., Jeanneret B., Jeckelmann B., Richard P., Courteville A., Salvadé Y. and Dändliker R., A proposal for a new moving-coil experiment, *IEEE Trans. Instrum. Meas.*, **48** (1999) 192.

[36] Beer W., Eichenberger A. L., Jeanneret B., Jeckelmann B., Pourzand A. R., Richard P. and Schwarz J. P., Status of the METAS watt balance experiment, *IEEE Trans. Instrum. Meas.*, **52** (2003) 626.

[37] Genevès G., Gournay P., Gosset A., Lecollinet M., Villar F., Pinot P., Juncar P., Clairon A., Landragin A., Holleville D., Pereira Dos Santos F., David J., Besbes M., Alves F., Chassagne L. and Topçu S., The BNM Watt Balance Project, *IEEE Trans. Instrum. Meas.*, **54** (2005) 850.

[38] Sullivan D. B. and Frederich N. V., Can superconductivity contribute to the determination of the absolute ampere, *IEEE Trans. Magnetics*, **13** (1977) 396.

[39] Shiota F. and Hara K., A study of a superconducting magnetic levitation system for an absolute determination of the magnetic flux quantum, *IEEE Trans. Instrum. Meas.*, **IM-36** (1987) 271.

[40] Shiota F., Miki Y., Namba A., Nezu Y., Sakamoto Y., Morokuma T. and Hara K., Absolute determination of the magnetic flux quantum using superconducting magnetic levitation, *IEEE Trans. Instrum. Meas.*, **44** (1995) 583.

[41] Frantsuz E. T., Gorchakov Y. D. and Khavinson V. M., Measurements of the magnetic flux quantum, Planck constant, and elementary charge at VNIIM, *IEEE Trans. Instrum. Meas.*, **41** (1992) 482.

[42] Shiota F., Miki Y., Fujii Y., Morokuma T. and Nezu Y., Evaluation of equilibrium trajectory of superconducting magnetic levitation system for the future kg unit of mass, *IEEE Trans. Instrum. Meas.*, **49** (2000) 1117.

[43] Fujii Y., Miki Y., Shiota F. and Morokuma T., Mechanism for levitated superconductor experiment, *IEEE Trans. Instrum. Meas.*, **50** (2001) 580.

[44] Riski K., Heikkinen P., Kajastie H., Manninen J., Rossi H., Nummila K., Frantsuz E. and Khavinson V., *Design of a superconducting magnetic levitation system* in *Proceedings IMEKO TC3 2001* (2001) pp. 239-246.

[45] Bego V., Determination of the volt by means of voltage balances, *Metrologia*, **25** (1988) 127.

[46] Funck T. and Sienknecht V., Determination of the volt with the improved PTB voltage balance, *IEEE Trans. Instrum. Meas.*, **40** (1991) 158.

[47] Bego V., Butorac J. and Poljančić K., Voltage balance for replacing the kilogram, *IEEE Trans. Instrum. Meas.*, **44** (1995) 579.

[48] Bego V., Butorac J. and Llić D., Realization of the kilogram by measuring at 100 kV with the voltage balance ETF, *IEEE Trans. Instrum. Meas.*, **48** (1999) 212.

[49] Mills I. M., Mohr P. J., Quinn T. J., Taylor B. N. and Williams E. R., Redefinition of the kilogram: a decision whose time has come, *Metrologia*, **42** (2005) 71.

[50] Davis R. S., Possible new definitions of the kilogram, *Philos. Trans. R. Soc., London, Ser. A*, **363** (2005) 2249.

[51] Flowers J.-L. and Petley B. W., The kilogram redefinition an interim solution, *Metrologia*, **42** (2005) L31.

[52] Mills I. M., Mohr P. J., Quinn T. J., Taylor B. N. and Williams E. R., Redefinition of the kilogram, ampere, kelvin and mole: a proposed approach to implementing CIPM recommendation 1 (CI-2005), *Metrologia*, **43** (2006) 227.

[53] Becker P., de Bièvre P., Fujii K., Gläser M., Inglis B., Luebbig H. and Mana G., Considerations on future redefinitions of the kilogram, the mole and of other units, *Metrologia*, **44** (2007) 1-14.

[54] Taylor B. N. and Mohr P. J., On the redefinition of the kilogram, *Metrologia*, **36** (1999) 63.

[55] de Bièvre P., Valkiers S., Kessel R., Taylor P. D. P., Becker P., Bettin H., Peuto A., Pettorruso S., Fujii K., Waseda A., Tanaka M., Deslattes R. D., Peiser H. S. and Kenny M. J., A reassessment of the molar volume of silicon and the Avogadro constant, *IEEE Trans. Instrum. Meas.*, **50** (2001) 593.

[56] Becker P., The molar volume of single-crystal silicon, *Metrologia*, **38** (2001) 85.

[57] Becker P., Bettin H., Danzenbrink H.-U., Glaser M., Kuetgens U., Nicolaus A., Schiel D., de Bievre P., Valkiers S. and Taylor P., Determination of the Avogadro constant via the silicon route, *Metrologia*, **40** (2003) 271.

[58] Schwitz W., Jeckelmann B. and Richard P., Towards a new kilogram definition based on a fundamental constant, *C. R. Physique*, **5** (2004) 881.

This page intentionally left blank

Towards an atomic realization of the kilogram: The measurements of N_{A} and $N_{\mathrm{A}}h$

P. Becker

Physikalisch-Technische Bundesanstalt - Bundesallee 100, 38116 Braunschweig, Germany

M. Jentschel

Institut Laue-Langevin - F-38042 Grenoble, Cedex, France

G. Mana(*)

Istituto Nazionale di Ricerca Metrologica - Strada delle Cacce 91, 10135 Torino Italy

G. Zosi

Università di Torino, Dipartimento di Fisica Generale "A. Avogadro" - via P. Giuria, 1, 10125 Torino Italy

1. – Introduction

The mass of the kilogram prototype, invariable by definition, is suspected drifting by parts in 10^8 per century [1]. Therefore, a definition linking it to invariant quantities or to exactly defined values of fundamental constants is highly desirable [2, 3]. Fundamental constants indicate when we are giving different names to quantities that are basically the same, but measured with different instruments. From this point of view, mass is an alias of frequency, the conversion factor being c^2/h; at the atomic scale, we can display this connection making use of the Compton's frequency $\nu_{\mathrm{C}} = mc^2/h$ [4, 5]. A link must

(*) e-mail address: `g.mana@inrim.it`

next be made with the macroscopic scale: a way is the realization of an object whose number of atom can be counted with the necessary accuracy and whose mass is 1 kg. This corresponds to determine the Avogadro's constant, which is a way to express the ^{12}C mass or its twelfth (the unified atomic mass unit) in terms of the mass of the kilogram prototype.

The present paper will survey a collaboration(1) aimed at accurate measurements of nuclear binding energies to determine the product $N_{\mathrm{A}}h$ — molar Planck's constant [6]. The measurement is based on the fact that, in a neutron capture reaction, the daughter isotope is slightly lighter then the ensemble formed by mother isotope and the neutron. This mass difference is equal to the binding energy and can be measured by determining the frequencies of the γ-ray cascade in the decay scheme of the capture-state. In principle, nuclear transitions can set time, length, and mass via the frequency and wavelength of the emitted γ-rays and the mass differences between the relevant energy levels. Although, for what concerns time and length [7] this is far beyond today capabilities, nuclear transitions open the way to measurements of atomic masses in terms of frequency measurements [8,9], provided measurements are performed at the required 10^{-8} accuracy level. The way from the atomic to the macroscopic scales is extremely difficult and it is presently limited to 10^{-7} accuracy. We then consider a second international effort(2) to realize a macroscopic mass to within 10^{-8} accuracy from the counting of the individual atomic constituents of a 1 kg ^{28}Si crystal-sphere [10].

2. – Weighing γ-ray photons

In relativistic quantum mechanics, the frequency and mass of field quanta (photons, in the case of the electromagnetic field, and point-like particles, in the case of matter fields) are linked by the Planck-Einstein relation $mc^2 = h\nu$, which is the zero momentum case of the dispersion relation relating the field's wavelength and frequency. In the case of composite particles, such as atoms and nuclei, there is a spectrum of internal energies and, therefore, of masses. As fig. 1 shows, the mass variations of about 1 eV associated with absorption or emission of visible photons by atoms are negligibly smaller than atomic masses, who are up to tens of GeV, but γ-ray photons related to nuclear transitions in the MeV region are not.

[1] Institut Laue Langevin (ILL - France), Istituto Nazionale di Ricerca Metrologica (INRIM - Italy), Physikalisch-Technische Bundesanstalt (PTB - Germany).

[2] Bureau International des Poids et Mesures (BIPM), Institute for Reference Material and Measurements (IRMM - European Commission joint research center), Istituto Nazionale di Ricerca Metrologica (INRIM - Italy), National Institute of Standards and Technology (NIST - USA), National Measurement Institute of Japan (NMIJ), National Measurement Laboratory (NML - Australia), National Physical Laboratory (NPL - UK), Physikalisch-Technische Bundesanstalt (PTB - Germany), Università di Torino, Dipartimento di Fisica Generale "A. Avogadro" (DFGAA - Italy).

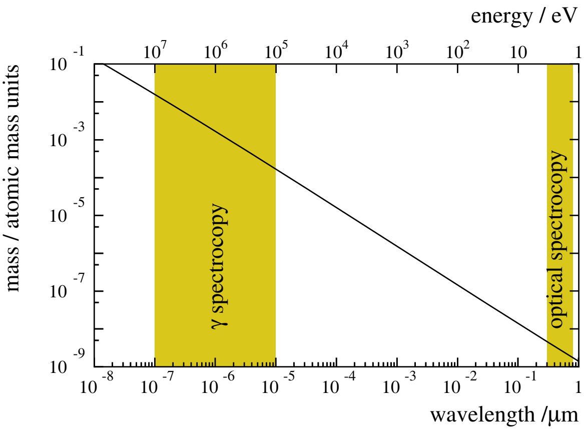


Fig. 1. – Masses of photons in the optical, X, and γ regions of the electromagnetic spectrum.

The measurement of the mass of γ-ray photons is quite different from weighing procedures used in mass metrology. It can be illustrated, for example, by the neutron capture reaction

$$^{35}\mathrm{Cl} + \mathrm{n} \rightarrow {}^{36}\mathrm{Cl}^* \rightarrow {}^{36}\mathrm{Cl} + \sum_i \gamma_i . \tag{1}$$

The decay scheme of ^{36}Cl from the capture state ^{36}Cl* to the ground state is shown in fig. 2. If the molar masses of the neutron, $M(\mathrm{n})$, and of the ^{35}Cl and ^{36}Cl isotopes,

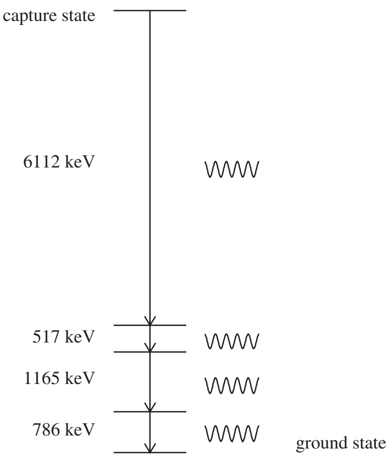


Fig. 2. – Decay scheme of ^{36}Cl*.

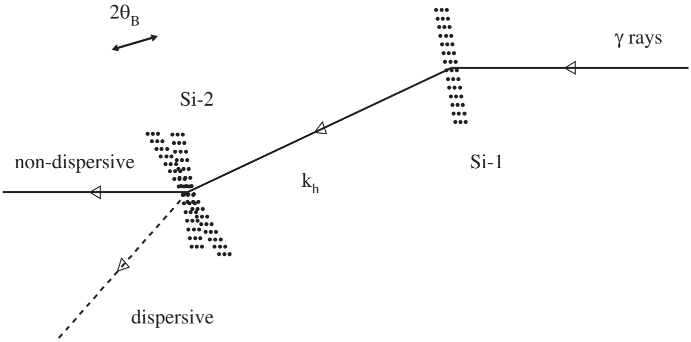


Fig. 3. – Conceptual scheme of a two-crystal γ-ray spectrometer; γ-rays are diffracted by two Si crystals (Si-1 and Si-2). Both dispersive and non-dispersive geometries are shown.

$M(^{35}\mathrm{Cl})$ and $M(^{36}\mathrm{Cl})$, are known, then the mass scale of the $^{36}\mathrm{Cl}^*$ decay is fixed by

$$[M(^{35}\mathrm{Cl}) + M(\mathrm{n}) - M(^{36}\mathrm{Cl})]c^2 = N_\mathrm{A} h\nu_{^{36}\mathrm{Cl}^*\to{}^{36}\mathrm{Cl}}, \tag{2}$$

where c is the speed of light and $\nu_{^{36}\mathrm{Cl}^*\to{}^{36}\mathrm{Cl}}$ is the frequency of the (virtual) γ-ray emitted in the transition from the capture to the ground states. Though the decay frequency $\nu_{^{36}\mathrm{Cl}^*\to{}^{36}\mathrm{Cl}}$ is different from the Compton's frequency ν_u of the atomic mass unit([3]), ν_u can be obtained by scaling $\nu_{^{36}\mathrm{Cl}^*\to{}^{36}\mathrm{Cl}}$ according to the difference between the relative atomic masses of the capture and ground states,

$$\nu_u = \frac{m_u c^2}{h} = \frac{\nu_{^{36}\mathrm{Cl}^*\to{}^{36}\mathrm{Cl}}}{A_r(^{35}\mathrm{Cl}) + A_r(\mathrm{n}) - A_r(^{36}\mathrm{Cl})}, \tag{3}$$

where $A_r(\mathrm{X}) = 12M(\mathrm{X})/M(^{12}\mathrm{C})$ is the relative atomic mass. Since the denominator of (3) is about 2.6×10^{-4}, in order to achieve a Compton's frequency measurement to within a 10^{-8} relative uncertainty, the relative atomic masses must be measured to within 10^{-12} relative uncertainties.

2˙1. *γ-ray spectroscopy.* – The nuclear transition-frequencies $\nu = c/\lambda$ are obtained by wavelength measurements via the Bragg's equation

$$\lambda = 2d\sin(\theta_B), \tag{4}$$

where $2\sin(\theta_B)$ is measured with the aid of a two-crystal spectrometer and the spacing d of the diffracting planes is measured in terms of a primary meter realization by combined X-ray and optical interferometry. The spectrometer operation is described in detail in [11]. As shown in fig. 3, the γ-rays hit the first crystal at the Bragg's angle, θ_B,

([3]) The atomic mass unit is 1/12 the mass of the ^{12}C atom.

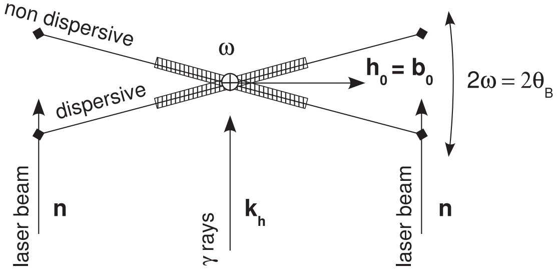


Fig. 4. – Conceptual scheme of an angle interferometer. Two retro-reflection cube corners (here represented by diamonds) are mounted at the ends of a beam that is rigidly fastened to the goniometer spindle. The mid-point of the line joining optical reflection points passes through the axis of rotation.

and are diffracted. By rocking the second crystal about the Bragg's angle in both the non-dispersive and dispersive geometries, diffraction peaks are recorded which identify the two configurations satisfying exactly the Bragg's law (4). The angular interval 2ω between the dispersive and non-dispersive configurations is twice the Bragg's angle. Since the diffraction angle of MeV's γ-rays is less than 5 mrad, in order to achieve 0.01 parts per million measurement-uncertainty, it is necessary to measure angles with a resolution better that 50 prad. With an angle interferometer having 0.3 m baseline (the optical lever of the interferometer), it is necessary to have 15 pm resolution in the measurement of the differential displacements of the lever ends. Additionally, since the interferometer calibration angle is about 250 mrad, utmost care must be placed in linearity.

In practice, the spectrometer operation is three-dimensional; therefore, the relationship between the Bragg's angle and the angular interval from the non-dispersive (goniometer angle equal to $-\omega$) to dispersive (goniometer angle equal to $+\omega$) configurations is

$$\sin(\theta_B) = \hat{\mathbf{k}}_h \cdot \big(\hat{\boldsymbol{\omega}} \times \hat{\mathbf{h}}_0\big) \sin(\omega), \tag{5}$$

where $\hat{\mathbf{k}}_h$ is the propagation direction of γ-rays diffracted by the first diffractometer-crystal, $\hat{\boldsymbol{\omega}}$ is the axis of the rotation, and $\hat{\mathbf{h}}_0$ is the normal to the diffracting planes at the zero angle of the goniometer. When $\hat{\boldsymbol{\omega}}$, $\hat{\mathbf{h}}_0$, and $\hat{\mathbf{k}}_h$ form a right triplet, $\hat{\mathbf{k}}_h \cdot (\hat{\boldsymbol{\omega}} \times \hat{\mathbf{h}}_0) = 1$ and (5) reduces to the trivial relationship $\sin(\theta_B) = \sin(\omega)$.

The rotation angle of the goniometer is measured with an angle interferometer, schematically shown in fig. 4, according to the equation $n\lambda_o = 2\sin(\omega)$, where n is the number of optical fringes, of period λ_o radians, observed during the rotation 2ω. Also in this case, in practice, the interferometer operation is three-dimensional; therefore, the measurement equations is

$$n\lambda_o = 2\hat{\mathbf{o}} \cdot \big(\hat{\boldsymbol{\omega}} \times \hat{\mathbf{b}}_0\big) \sin(\omega), \tag{6}$$

where $\hat{\mathbf{o}}$ is the propagation direction of the laser beam and $\hat{\mathbf{b}}_0$ is a unit vector parallel to the interferometer baseline (the interferometer optical-lever, *i.e.*, the line joining the optical centers of two cube corners) at the zero angle of the goniometer. The equations (5) and (6) express the fact that the γ-ray and optical "encoders" sense the components of the crystal rotation in the reflection planes identified by $\hat{\mathbf{k}}_h$ and $\hat{\mathbf{h}}_0$ (γ-ray) and by $\hat{\mathbf{o}}$ and $\hat{\mathbf{b}}_0$ (optical).

By combining (5) and (6), we obtain the measurement equation

$$2\sin(\theta_B) = \hat{\mathbf{k}}_h \cdot \big(\hat{\boldsymbol{\omega}} \times \hat{\mathbf{h}}_0\big)\, n\lambda_{\rm o}^*, \tag{7}$$

where $\lambda_{\rm o}^* = \lambda_{\rm o}/[\hat{\mathbf{o}}\cdot(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_0)]$. By using $\langle\hat{\mathbf{k}}_h\cdot(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{h}}_0)\rangle = 1-\delta^2/2$, where δ^2 is the variance of the (zero mean) deviation of $\hat{\boldsymbol{\omega}}$, $\hat{\mathbf{h}}_0$, and $\hat{\mathbf{k}}_h$ from a right triplet([4]), (7) reads

$$2\sin(\theta_B) = \big(1-\delta^2/2\big)\, n\lambda_{\rm o}^*. \tag{8}$$

Observing that the standard deviation of $\hat{\mathbf{k}}_h\cdot(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{h}}_0)$ is $\delta^2/\sqrt{2}$, a targeted measurement accuracy equal to $10^{-8}\sin(\theta_B)$ implies that the deviation of $\hat{\boldsymbol{\omega}}$, $\hat{\mathbf{h}}_0$, and $\hat{\mathbf{k}}_h$ form a right triplet must be confined to within 10^{-4} rad, or 20 arcseconds.

In order to convert the number of optical fringes into angles, the interferometer is calibrated against an optical polygon mounted on the rotation axis. An autocollimator senses the polygon faces while it is rotated face by face and the rotation angles are measured. For each rotation step, the calibration equation is

$$n_i\lambda_{\rm o} = 2\hat{\mathbf{o}}\cdot\big(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_{0i}\big)\sin(\alpha_i), \tag{9}$$

where where n_i is the number of optical fringes, of period $\lambda_{\rm o}$ radians, observed during the rotation $2\alpha_i$ equal to the i-th polygon external angle and $\hat{\mathbf{b}}_{0i}$ is the interferometer baseline at the zero angle of the goniometer relative to the adjacent polygon faces. The zero angles of the goniometer relative to the crystal and polygon rotations (*i.e.*, midway between the non-dispersive and dispersive configurations, crystal rotation, and the two autocollimator zeroing, polygon rotation) are different. Let the two relevant baselines (*i.e.*, the lines joining the optical centres of the corner cubes at the zero angles of the goniometer), $\hat{\mathbf{b}}_0$ and $\hat{\mathbf{b}}_{0i}$, deviate by a small random quantity $\boldsymbol{\delta}_i$ with zero mean and ζ^2 variance; hence $\hat{\mathbf{b}}_{0i} = \hat{\mathbf{b}}_0 + \boldsymbol{\delta}_i$, where $2\hat{\mathbf{b}}_0\cdot\boldsymbol{\delta}_i = -\delta_i^2$, since $|\hat{\mathbf{b}}_{0i}| = 1$. Consequently, provided $\hat{\mathbf{o}}\cdot(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_0)\approx 1$ (*i.e.*, that the unit vectors $\hat{\mathbf{o}}$, $\hat{\boldsymbol{\omega}}$, and $\hat{\mathbf{b}}_0$ form a right triplet),

$$\big\langle\hat{\mathbf{o}}\cdot\big(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_{0i}\big)\big\rangle = \big(1-\zeta^2/2\big)\,\hat{\mathbf{o}}\cdot\big(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_0\big). \tag{10}$$

In fact, since the components of $\boldsymbol{\delta}_i$ parallel to $\hat{\boldsymbol{\omega}}$ and $\hat{\mathbf{o}}$ do not contribute to $\hat{\mathbf{o}}\cdot(\hat{\boldsymbol{\omega}}\times\boldsymbol{\delta}_i)$, provided $\hat{\mathbf{b}}_0 = \hat{\boldsymbol{\omega}}\times\hat{\mathbf{o}}$, the only $\boldsymbol{\delta}_i$ contributing component is $\hat{\mathbf{b}}_0\cdot\boldsymbol{\delta}_i$ and $\langle\hat{\mathbf{b}}_0\cdot\boldsymbol{\delta}_i\rangle = -\zeta^2/2$. Therefore, we rewrite (9) as

$$n_i\lambda_{\rm o}^* = 2\big(1-\zeta^2/2\big)\sin(\alpha_i). \tag{11}$$

[4] The expected value of the square of a zero-mean normal variable is equal to its variance.

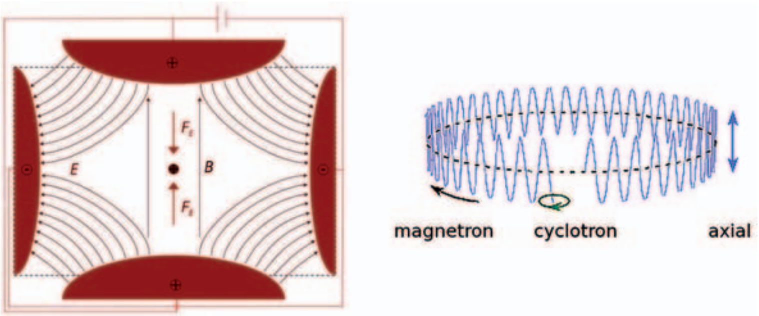


Fig. 5. – Conceptual scheme of a Penning's trap. The electric field is generated by a quadrupole of endcaps and a ring electrode. The magnetic field is generated by an electric toroidal magnet. The ion motions (an axial oscillation with a frequency of about 200 kHz and two circular, magnetron and cyclotron, motions with frequencies of about 5 kHz and 5 MHz, respectively) are illustrated on the left-hand side. (Adapted from Wikipedia.)

After measuring one of the external angles, the polygon is rotated so that the next pair of faces is viewed and the corresponding angle is measured in terms of fringes. The rotation is performed by lifting the polygon and rotating the axis by the desired angle; then, the polygon is lowered again. When all external angles have been measured, using the closure equation $2\sum_i \alpha_i = 2\pi$, since $\arcsin[(1-\zeta^2/2)\sin(\alpha_i)] \approx (1-\zeta^2/2)\alpha_i$, we obtain the calibration equation

$$\sum_i \arcsin\big(n_i\lambda_{\mathrm{o}}^*/2\big) = \big(1-\zeta^2/2\big)\pi. \tag{12}$$

In order to estimate the alignment requirement for the calibration procedure, we can approximate $\arcsin(n_i\lambda_{\mathrm{o}}^*/2)$ by $n_i\lambda_{\mathrm{o}}^*/2$. Therefore, (12) simplifies as

$$N\lambda_{\mathrm{o}}^* \approx 2\big(1-\zeta^2/2\big)\pi, \tag{13}$$

where $N = \sum_i n_i$. Finally, since the standard deviation of $\hat{\mathbf{o}}\cdot(\hat{\boldsymbol{\omega}}\times\hat{\mathbf{b}}_{0i})$ is $\zeta^2/\sqrt{2}$, a targeted calibration accuracy equal to $10^{-8}\lambda_{\mathrm{o}}^*$ implies that the deviation of $\hat{\mathbf{o}}$, $\hat{\boldsymbol{\omega}}$, and $\hat{\mathbf{b}}_{0i}$ from a right triplet must be confined to within 10^{-4} rad, or 20 arcseconds.

2˙2. *Mass spectroscopy.* – The most accurate measurements of atomic-mass ratios are performed by comparing the cyclotron frequencies of ions confined in a Penning's trap [9]. As shown in fig. 5, ions are prevented from escaping radially by means of a uniform magnetic field of about 8.5 T and axially by an electric quadrupole field — a set of rotationally symmetric hyperbolic electrodes biased by about 15 V. Ions display three types of motion, an axial oscillation and two circular — magnetron and cyclotron — motions, and are confined in a small volume, about 1 mm^3, for several weeks. The mass ratio is determined by measuring the ratio of the free space cyclotron frequencies

$$\omega = \frac{eB}{m}\,, \tag{14}$$

where B is the magnetic field and e is the electron's charge. If the $D^{35}Cl^+$ and $H^{36}Cl^+$ ions are simultaneously trapped([5]), since the molar mass of $M(H^{36}Cl)$ is similarly measured by using the ^{12}C reference — after accounting for the missing electron and chemical binding energies — the ion mass ratio gives, without loss of accuracy, the molar mass difference

$$\begin{aligned} M\big(^{35}\mathrm{Cl}\big) + M(\mathrm{D}) - M(\mathrm{H}) - M\big(^{36}\mathrm{Cl}\big) = M\big(^{12}\mathrm{C}\big)\Big[&A\big(\mathrm{D}^{35}\mathrm{Cl};\mathrm{H}^{36}\mathrm{Cl}\big)A\big(\mathrm{H}^{36}\mathrm{Cl};{}^{12}\mathrm{C}\big) - \\ &-A\big(\mathrm{H}^{36}\mathrm{Cl};{}^{12}\mathrm{C}\big)\Big], \end{aligned} \tag{15}$$

where $A(\mathrm{X};\mathrm{Y}) = m(\mathrm{X})/m(\mathrm{Y})$ are atomic mass ratios.

Let now the reaction $\mathrm{H} + \mathrm{n} \rightarrow \mathrm{D} + \gamma$ be considered, so that, by combining (2) and $[M(\mathrm{H}) + M(\mathrm{n}) - M(\mathrm{D})]c^2 = N_\mathrm{A} h\nu(D)$, we obtain

$$\big[M\big(^{35}\mathrm{Cl}\big) + M(\mathrm{D}) - M(\mathrm{H}) - M\big(^{36}\mathrm{Cl}\big)\big]c^2 = \Big[\nu_{^{36}\mathrm{Cl}^* \rightarrow {}^{36}\mathrm{Cl}} - \nu_{\mathrm{D}^* \rightarrow \mathrm{D}}\Big] N_\mathrm{A} h, \tag{16}$$

which can be used to determine both the $N_\mathrm{A} h$ product and the Compton's frequency of ^{12}C

$$\frac{m\big(^{12}\mathrm{C}\big)c^2}{h} = \frac{\nu_{^{36}\mathrm{Cl}^* \rightarrow {}^{36}\mathrm{Cl}} - \nu_{\mathrm{D}^* \rightarrow \mathrm{D}}}{A\big(\mathrm{D}^{35}\mathrm{Cl};\mathrm{H}^{36}\mathrm{Cl}\big)A\big(\mathrm{H}^{36}\mathrm{Cl};{}^{12}\mathrm{C}\big) - A\big(\mathrm{H}^{36}\mathrm{Cl};{}^{12}\mathrm{C}\big)}\,, \tag{17}$$

where we used (15) and $N_\mathrm{A} m(^{12}\mathrm{C}) = M(^{12}\mathrm{C})$.

3. – X-ray crystal density measurement of N_A

If an exactly defined value of Compton's frequency of the ^{12}C atom is provided, the kilogram can be realized from an atomic mass through an artifact whose number of atomic constituents is known. A way for this *mise en pratique* is by Si crystallization which, acting as a low noise amplifier, grows up a macroscopic replica of the unit cell. Crystallization enables the mass of a 1 kg ^{28}Si crystal-sphere (an approximate realization

[5] The ^{36}Cl isotope of Cl does not occur naturally; it is a radioisotope with a half-life of about 300000 years. Therefore, though it is, in principle, correct, the simple strategy here outlined in order to measure Compton's frequency of ^{12}C may be, in practice, not doable.

Table I. – *Uncertainties of amount-of-substance fraction and concentration of* ^{28}Si *contamination delivering* 10^{-8} *contributions to the* N_A*'s error budget.*

Contaminant	$\lvert M_X - M(^{28}\text{Si})\rvert$ (g/mol)	σ_x (μmol/mol)	σ_c (10^{14}/cm^3)
C and O	16	0.018	9
N	14	0.020	10
N_2, V, and I	28	0.010	5

of a mono-isotopic and chemically pure single-crystal) to be related to $\nu_C(^{28}\text{Si})$. This amounts to a determination of Avogadro's number([6]), based on Bragg's equation

$$N_A = \frac{8V_{\text{mol}}}{V_{\text{cell}}} = \frac{8M}{\rho a_0^3}\,, \tag{18}$$

where $V_{\text{mol}} = M/\rho$ and $V_{\text{cell}} = a_0^3$ are the molar and unit-cell volumes, M is the mean molar mass of the enriched crystal, ρ its density, and a_0 its lattice parameter.

The lattice parameter and molar volume are not constants, but, provided the values measured in real crystals are extrapolated to a reference thermodynamical state — by specifying temperature, pressure, and chemical and isotopic compositions — one can provide invariant quantities. The degree to which a particular crystal represents the ideal depends mainly on the amounts of carbon, oxygen, and nitrogen impurities; provided they are sufficiently small, extrapolations have been demonstrated to within 0.01 parts per million relative accuracy [12].

As regards the N_A determination, provided the molar volume and lattice parameter are measured in the same crystal, such an extrapolation is not strictly necessary. In this case M must account not only for the isotopic contamination, but also for the impurity content. To simplify matters, let X indicate a generic impurity. Hence, the mean molar mass is

$$M = \big[1 - x(\text{X})\big]M\big(^{28}\text{Si}\big) + x(\text{X})M_X = M\big(^{28}\text{Si}\big) + x(\text{X})\big[M_X - M\big(^{28}\text{Si}\big)\big], \tag{19}$$

where $x(\text{X})$ is the amount-of-substance fraction of entity X, $M_X = M(\text{X})$ for substitutional impurities (C, N, and vacancy) and $M_X = M(\text{X}) + M(^{28}\text{Si})$ for interstitial impurities (O, N_2, and self-interstitial). Table I gives the maximum admissible uncertainties of the impurity amount-of-substance fraction measurements for a targeted 0.01 parts per million uncertainty of the mean molar mass.

Since the pioneering work by D. Deslattes at NIST in 1974 [13], many N_A determinations have been carried out in the framework of extensive international collaborations by using many different natural-Si crystals [14, 15]. However, a few of the molar volume values so obtained differ unexpectedly by more than one part per million. Additionally,

([6]) The number of atoms in 12 g of ^{12}C; the Avogadro's constant is the same number per mole.

the presently most updated value, $N_\mathrm{A} = 6.0221353(18) \times 10^{23}$ 1/mol [16], is lower by about $10^{-6} N_\mathrm{A}$ than the value, $N_\mathrm{A} = 6.02214180(31) \times 10^{23}$ 1/mol, calculated by

$$N_\mathrm{A} = \frac{cM(\mathrm{e})\alpha^2}{2hR_\infty}\,, \tag{20}$$

where $M(\mathrm{e})$ is the electron's molar mass, α is the fine-structure constant, R_∞ is the Rydberg's constant, and $h = 6.62606901(34) \times 10^{-34}$ Js is the Planck's constant value obtained by the watt-balance experiment [17].

The causes of these inconsistencies were intensively investigated. Si crystals were carefully characterized, both chemically and structurally; lattice parameter and density measurements were repeatedly made again and cross-checked. Although these investigations evidenced systematic effects potentially affecting measurements at the $10^{-7} N_\mathrm{A}$ accuracy level, no faults capable to explain parts per million discrepancies were found. However, molar mass measurements were not so carefully revised, mainly because amount-of-substance measurements with adequate accuracy are performed only at the IRMM; eventually, they became the major accuracy-limiting factor of N_A determinations and no improvement was considered possible with the use of natural-Si crystals. In order to bypass this limitation and to tackle the criticisms raised about the inconsistencies observed, a new N_A determination has been started *ab initio*, which will be based on a highly enriched ^{28}Si crystal [10].

3˙1. *Molar mass.* – In order to estimate the expected uncertainty of the molar mass measurement, let the approximate measurement equation

$$M = x_{28} M\left(^{28}\mathrm{Si}\right) + x_{29} M\left(^{29}\mathrm{Si}\right) + x_{30} M\left(^{30}\mathrm{Si}\right) \approx M\left(^{28}\mathrm{Si}\right) + 3x/2\,\mathrm{g/mol} \tag{21}$$

be considered. In (21), x_i is the amount-of-substance fraction of ^{i}Si, $x_{28} = 1 - x_{29} - x_{30} = 1 - x$, and $x_{29} \approx x_{30} \approx x/2$. The quantities actually measured by mass spectrometry are the ion current ratios, which are calibrated against synthetic mixtures having gravimetrically determined isotopic compositions. In this simplified two-isotope model, the amount-of-substance fraction x can be identified with the amount-of-substance ratio $x \approx x/x_{28} = \kappa I/I_{28}$, where $\kappa \approx 1$ is the calibration factor and I and I_{28} are the relevant ion currents. Hence,

$$\frac{\sigma_M}{M} \approx \frac{3\sqrt{\sigma_\kappa^2 + (\sigma_I/I)^2}\,x}{2M}\,, \tag{22}$$

where the negligible contribution $\sigma_{I_{28}}/I_{28}$ has been omitted. Equation (22) evidences that the greater the isotopic purity the smaller the measurement uncertainty. Since the amount of substance fractions x_{28} and x of natural Si are 0.92 and 0.08, to reduce the present (pessimistically estimated) part-per-million uncertainty by a factor one hundred, the ^{28}Si fraction in the enriched material must be greater than 0.9992.

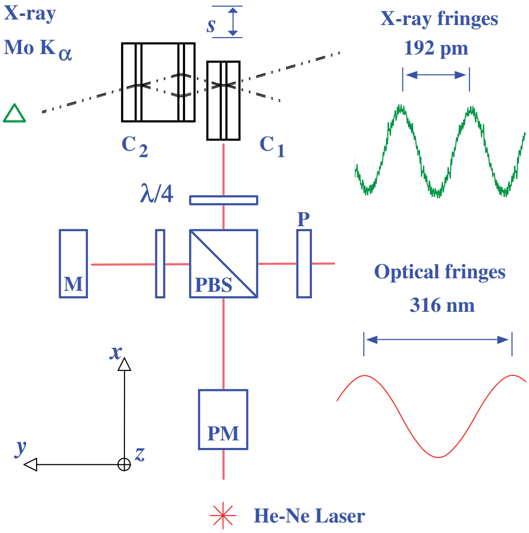


Fig. 6. – Conceptual scheme of a combined X-ray and optical interferometer. C_1 and C_2 movable and fixed crystals of the X-ray interferometer, M fixed mirror of the optical interferometer, PBS polarizing beam-splitter, PM phase modulator, P polarizer, $\lambda/4$ quarter-wave plates.

3·2. *Lattice parameter.* – Silicon is a cubic crystal with eight atoms per face-centered unit cell of edge $a_0 \approx 543\,\text{pm}$, which is related to the measured d_{220} spacing of the planes with Miller indices (2,2,0) by $a_0 = \sqrt{8}d_{220}$. The first absolute measurement of the (220) lattice spacing was performed at NIST in 1970's using a combined X-ray and optical interferometer [18]. The operation of this device is described in details in [19]. As shown in fig. 6, it consists of three thin plane-parallel Si crystals so cut to make the (220) lattice planes perpendicular to the crystal surfaces. An X-ray beam is split by the first crystal and recombined by the third. When this crystal is moved along a direction orthogonal to the (220) planes, a periodic variation of the transmitted and diffracted X-ray intensities is observed, whose period is equal to the spacing between the (220) planes. The X-ray interferometer embeds a mirror ideally parallel to the (220) planes, so that their displacement is measured via laser interferometry. The measurement equation is $d_{220} = (n/m)\lambda/2$, where m is the number of X-ray fringes in n optical fringes of $\lambda/2$ period. The equivalence of d_{220} and $\lambda/2$ and the periods of the X-ray and optical fringes must be examined with care; actually, measurement equations are

$$m\lambda/2 = \hat{\mathbf{k}} \cdot (\mathbf{s} + \mathbf{b} \times \boldsymbol{\omega}) + \varsigma(\boldsymbol{\omega}, \mathbf{s}) \tag{23}$$

and

$$nd_{220} = \hat{\mathbf{h}} \cdot \mathbf{s} + \chi(\boldsymbol{\omega}, \mathbf{s}), \tag{24}$$

where $\hat{\mathbf{k}}$ is the propagation vector of the laser beam, $\mathbf{s}$ is the displacement of the X-ray interferometer, $\mathbf{b}$ is the offset between the centers of the X-ray and laser beams, $\boldsymbol{\omega}$ is the parasitic rotation of the X-ray interferometer, $\hat{\mathbf{h}}$ is the normal to the lattice planes, and $\varsigma(\boldsymbol{\omega},\mathbf{s})$ and $\chi(\boldsymbol{\omega},\mathbf{s})$ are corrections taking into account aberration effects in both the X-ray and optical interferometers.

3·3. *Density.* – In (18), the molar volume is given by the ratio between molar mass and density. As regards density, it is measured on the basis of first principles, that is on mass and volume measurements; however, this approach was not not pursued until the manufacturing of an almost perfect 1 kg sphere was demonstrated [20]. Subsequent investigations showed that an exactly spherical shape is not a strict requirement — but surface smoothness is — as the volume can be precisely calculated from diameter measurements and surface topography [21, 22], which are performed with the aid of optical interferometers [23]. Accordingly, the volume is given by

$$V = \frac{4\pi a_{00}^3}{3}\left(1 + 3\sum_{i=1}^{\infty}\sum_{m=-l}^{l}\left|\frac{a_{lm}}{a_{00}}\right|^2 + \cdots\right), \tag{25}$$

where a_{lm} is the amplitude of the spherical harmonic $\mathrm{Y}_l^m(\theta,\phi)$. There are, however, problems which need careful investigations. Since the sphere diameter is about 93 mm, 10^{-8} accuracy requires diameter measurements with sub-nanometer accuracy and difficulties arise from the way the sphere surface is defined. First, the sphere is continuously deformed by gravity as it is rotated for the diameter measurements. Second, the optical properties of the surface affect the interferometric measurement. As the test beam reflection occurs at the silicon oxide interface, the oxide thickness must be measured and the silica mass must be subtracted to obtain the mass of the sphere alone, without the covering shell. Since several kind of chemico-physical interactions occur during the surface polishing, native oxidation is rather uneven; the surface layer of the sphere must be therefore removed and a thermal oxide layer a few nanometers thick must be grown anew on the sphere surface.

4. – Conclusions

The units of the *Système International* (SI) can be specified by fixing the values of a set of fundamental constants. These constants, together with laws of the physics, fix the entire system without no need to distinguish between base and derived units. The practical realization of any unit is a method defined by an appropriate measurement equation linking the relevant quantity to one or more constants. As regards the kilogram, two different *mise en pratique*, whose relationships and redundancies are illustrated in fig. 7, can be explored.

The first makes use of a watt balance [24] and virtually compares the mechanical power related to the uniform motion with velocity v of a macroscopic mass M in the gravitational field g with the electrical power related to the interaction between the

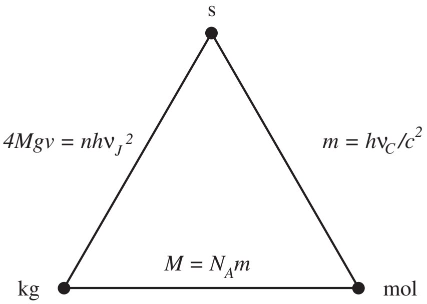


Fig. 7. – Metrological triangle of the mass-related units.

electrical current in the mass support and a magnetic field. Provided the electrical power is measured in terms of the Josephson's and von Klitzing's constants via a Josephson's device irradiated with a microwave of frequency ν_{J}, the measurement equation is given by

$$4Mgv = nh\nu_{\mathrm{J}}^2, \tag{26}$$

where n is the number of power quanta $h\nu_{\mathrm{J}}^2$.

The second *mise en pratique* consists of two steps. One measures the Compton's frequency of an atom, *e.g.*, ^{28}Si, to express an atomic mass in terms of frequency,

$$m = h\nu_{\mathrm{C}}/c^2. \tag{27}$$

The other uses of the Avogadro's constant to relate microscopic and macroscopic mass scales via the constitutive equation

$$M = N_{\mathrm{A}}m = N_{\mathrm{A}}h\nu_{\mathrm{C}}/c^2. \tag{28}$$

It is clear that both physics and technology are essential to bridge the gap between microscopic and macroscopic scales and to allow the kilogram to be defined in terms of fundamental constants. Therefore, the motivations for devoting resources to this effort lie within the wider framework of the synergic interactions between science and technology. Quoting from B. Petley [25], "*high precision measurements push theory and experiment to the very limits of which they are capable. This is an area where theory and experiment are tested to the limit, where the fallibility is often too apparent, and where one's best is only just good enough.*"

* * *

The authors are grateful to financial support from the Centro E. Fermi (Roma) and Compagnia di San Paolo (Torino).

REFERENCES

[1] Davis R., *Metrologia*, **40** (2003) 299.

[2] Mills I. M. *et al.*, *Metrologia*, **42** (2005) 71.

[3] Mills I. M. *et al.*, *Metrologia*, **43** (2006) 227.

[4] Cabiati F. *et al.*, *Il Nuovo Saggiatore*, **9** No. 1 (1993) 51.

[5] Bordé C. J., *Philos. Trans. R. Soc. A*, **363** (2005) 2177.

[6] Rainville S. *et al.*, *Nature*, **438** (2005) 1096.

[7] Peik E. and Tamm C., *Europhys. Lett.*, **61** (2003) 181.

[8] Dewey M. S. *et al.*, *Phys. Rev. C*, **73** (2006) 044303.

[9] Rainville S. *et al.*, *Science*, **303** (2004) 334.

[10] Becker P., *Metrologia*, **40** (2003) 366.

[11] Kessler E. J. *et al.*, *Nucl. Instrum. Methods Phys. Res. A*, **457** (2001) 187.

[12] Becker P. *et al.*, *IEEE Trans. Instrum. Meas.*, **56** (2007) 230.

[13] Deslattes R. D. *et al.*, *Phys. Rev. Lett.*, **33** (1974) 463.

[14] Mana G. and Zosi G., *Rivista del Nuovo Cimento*, **18** No. 3 (1995) 1.

[15] Becker P. *et al.*, *Metrologia*, **40** (2003) 271.

[16] Fujii K. *et al.*, *IEEE Trans. Instrum. Meas.*, **54** (2005) 854.

[17] Steiner L. R., *Metrologia*, **42** (2005) 431.

[18] Deslattes R. D. and Henins A., *Phys. Rev. Lett.*, **31** (1973) 972.

[19] Basile G. *et al.*, *IEEE Trans. Instrum. Meas.*, **44** (1995) 526.

[20] Leistner A. and Zosi G., *Appl. Opt.*, **26** (1987) 600.

[21] Mana G., *Metrologia*, **31** (1984) 289.

[22] Giardini W. J. and Mana G., *Rev. Sci. Instrum.*, **69** (1998) 1383.

[23] Nicolaus A. and Fujii K., *Meas. Sci. Technol.*, **17** (2006) 2527.

[24] Steiner R. *et al.*, *J. Res. Natl. Inst. Stand. Technol.*, **110** (2005) 1.

[25] Petley B. W., *The fundamental physical constants and the frontiers of measurement* (IOP Publishing Ltd, Bristol) 1988.

Metrology to support the development of nanotechnology

K. C. Carneiro

Danish Fundamental Metrology Ltd. - Matematiktorvet 307, DK 2800 Kgs. Lyngby, Denmark

1. – Introduction

"Nano" means "dwarf" in Italian. The term "nano" has also been adopted as the prefix for 10^{-9} in Système International des Unités (SI); and nanotechnology according to the definition by Taniguchi [1] is devoted to technologies, where the "critical dimensions" (CD) are between 0.1 and 100 nanometres (nm). In this lecture we will adhere to this definition, although other "nano"-parameters such as nanonewton (nN) nanovolt (nV) could also be included. Several of these other technologies are dealt with in other lectures of the School. Critical dimensions has to be understood in terms of a dimension that is critical for the performance of a product; it may be a small feature of a large object such as the rugosity of a surface, the width of a conducting strip, which is in itself several mm long, or the curvature of a corner in its departure from an ideally sharp edge.

Nanotechnologies have developed from several classical technologies. This is indicated in fig. 1 for physics, biology, and chemistry. For instance within physics, mechanical engineering has undergone a miniaturisation with typical structural sizes of millimetres (mm) in 1960 over micrometres in 1980 to nanometres at present. Biology has developed similarly; and during the miniaturisation, the improved scientific understanding has allowed a certain "functionalisation" of products, meaning that prescribed functions could be tailor-made by proper production at sufficiently small scale. Approaching the nm scale from the other side, chemistry has grown in complexity from the simple chemistry at ångström scale through complex and supramolecular chemistry into the nano regime. 1 nm equals 10 Å.

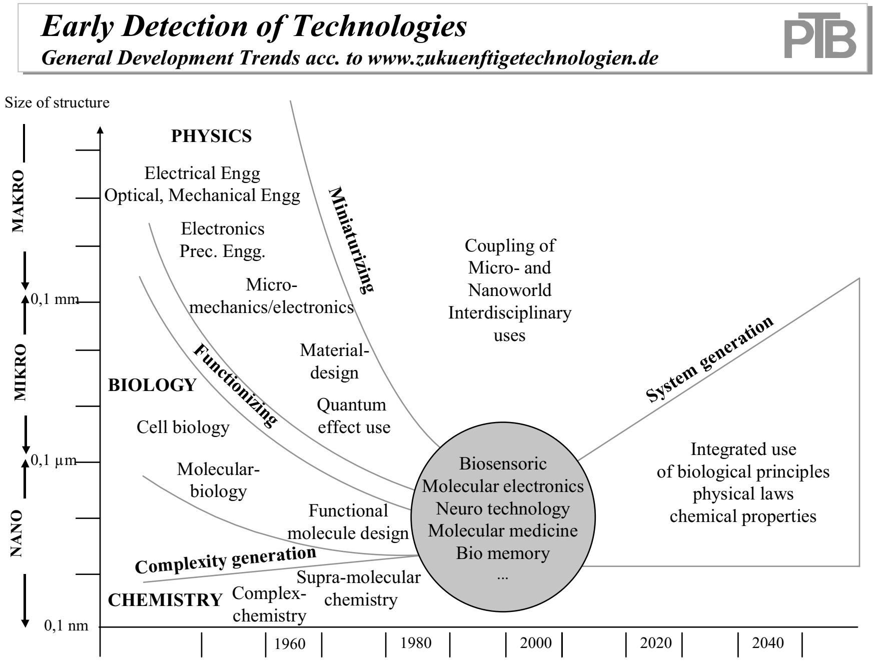


Fig. 1. – The development of nanotechnology from several different —and formerly separate— technologies (courtesy of professor H. Kunzmann, Physikalisch-Technische Bundesanstalt, PTB).

It appears from the above that nanotechnology has developed into a "common pot" from scientific disciplines, which have entirely different origins, concepts and terminologies; and the scientists who today work with nanotechnology have their education from one of these very different schools. This is part of the reason why nanotechnology is such a fertile and exciting area to work in; but it also poses the challenge of formulating concepts and terms that can satisfy the future needs of the —yet unexplored— nano science. And in particular, when taking the nanosciences to an industrially applicable technology, it is of paramount importance that a usable terminology be developed. This is a challenge for the standardisation community, which is currently being taken up by both the international standardisation organisation ISO, as well as its European counterpart CEN.

Similar to standardisation, metrology has to be developed in order to mature nanoscience into an applicable technology. Lord Kelvin is cited for stating: "You cannot produce what you cannot measure", and this of course holds for any new technology, including nanotechnology. It means that traceability has to be established from the realisation of the metre, measurement uncertainties must be developed for the measuring instruments in use, and interlaboratory comparisons must be performed to ensure that global harmonised measurements in the nm regime have been implemented.

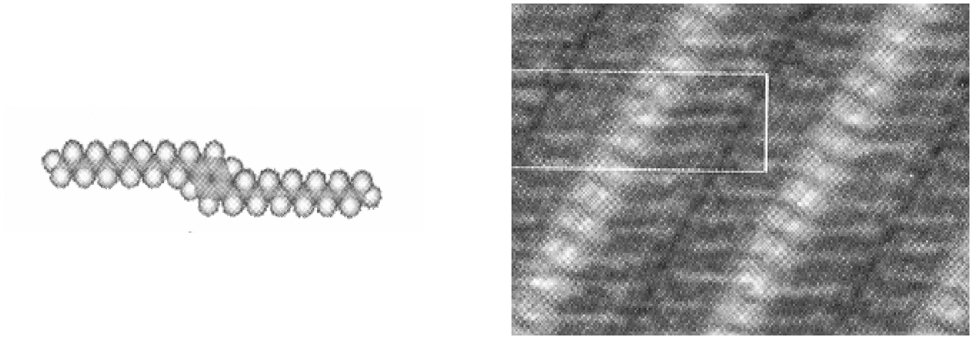


Fig. 2. – Corrected STM image (right) of a two-dimensional crystal of the molecule $C_{28}H_{30}$ (left) formed as a result of physical adsorption on graphite.

Finally, there is a need for instrumentation. Although advanced apparatus has been developed for scientific development and exploration of nanoscience, there are at present few instruments that may be industrially applied. Proper characterisation of manufactured nanosized samples are still reserved for few and advanced laboratories and is mostly too slow to be used in production control. For the time being the scanning probe microscope (SPM) is the only general-purpose measurement instrument; but SPM is not easy to use and is in general far too slow to monitor dynamic processes. Hence, the invention of fast and reliable measuring instruments for practical use is a challenge for the exploitation of nanotechnology.

2. – Early scanning probe microscopes and metrology developments

At the invention of the scanning tunnelling microscope in 1982 [2], the fascination was focussed around its ability to measure with atomic resolution and its potential as a metrology instrument was overlooked for some time. This is despite the fact that early attempts to make an instrument based on electron tunnelling had indeed been made at National Bureau of Standards (NBS, now NIST). This is because the STM and all later scanning probe techniques (SPM) are based on the movement through the ferroelectric effect, which is highly non-linear as well as unrepeatable because of hysteresis. It therefore required substantial corrections to deduce a metrology picture from the original data. During the late 1980s and the early 1990s several national metrology institutes, NMIs, took up scanning probe microscopy as an novel tool to investigate phenomena at the surfaces of matter. Preliminary comparisons were performed but traceability and uncertainty were not established.

An early attempt to make a traceable measurement with an STM is shown in fig. 2. The molecule $C_{28}H_{30}$ was dissolved in the solvent toluene and a droplet was put on a piece of pyrolytic graphite, which was then mounted in a special STM. The STM scan showed the formation of a two-dimensional crystal, but it was heavily distorted. The correction and calibration of the picture was performed by using the signal from the

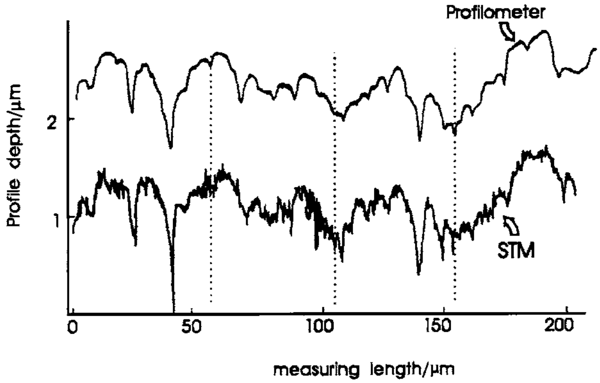


Fig. 3. – Comparison between a combination of STM scans and a scan performed with a classical profilometer on a "Halle" standard, which was used as a standard for roughness measurements (courtesy of PTB).

underlying periodic graphite structure, which is well known to have a C-C bond length of 0.142 nm. This allowed the production of the linearised and traceable picture in the right part of fig. 2.

Another exercise to demonstrate the comparability between a conventional surface measurement and an STM is shown in fig. 3. It shows a trace over 200 μm with a traditional profilometer, where the sensing tip consists of a sphere with a radius of about 1 μm. This was compared with four traces of an STM; and the comparison shows that the results are fully compatible; but it reveals the expected much higher resolution of the STM.

Subsequent to these early measurements, a more systematic approach to the metrology applications of SPM techniques has been pursued. Some developments are:

Traceability to the meter. – A simple way of establishing traceability to the metre in the nm region is to manufacture a grating of appropriate pitch and calibrate this with optical diffraction techniques that have low uncertainties compared to the industrial needs in nanotechnology. Alternatively, one can mount calibrated sensors to monitor the movements of the scanning probe, *i.e.* capacitive sensors; and instruments with this facility are now commercially available as "metrology SPMs". A more accurate sensor is an optical interferometer, and this has been mounted at some NMIs.

Uncertainties. – As SPMs have been better understood, it is now customary for national metrology institutes to supply uncertainty estimates for their measurements based on the universally accepted GUM method [3]. An example of measurement ranges and associated uncertainties are shown in table I. Traceability has been established using transfers gratings, and this is reflected in the measurement uncertainties. Interferometric calibration would result in uncertainties that are typically 10 times smaller.

Table I. – *DFMs measurements ranges and measurement capabilities. LMC: lowest measurement uncertainty; HMC: highest measurement uncertainty. All uncertainties are given for 95% confidence.*

Sample	Quantity	Range	LMC	HMC
2D surface standard	Pitch	500–12 000 nm	0.88 nm	18 nm
Height standard	Step height	20–3 000 nm	1.0 nm	8.0 nm

Interlaboratory comparisons. – The first global comparison on the measurements of a nm height sample is shown in fig. 4. It is an example of a series of comparisons that are being arranged by the Consultative Committee for Length (CCL) under the *Conventions du Mètre.* It has participation from the following 14 countries: Spain, Italy, Korea, United States, Poland, Japan, Holland, Germany, Russia, Taiwan, Denmark, Switzerland, China, and United Kingdom. The sample was made so that several techniques could be used other than SPM in order to demonstrate comparability between SPM and other well-established metrology tools. These are: ST: Stylus instruments (profilometers), IM: Interference microscopy, LHI: Laser heterodyne interferometry, and LMI: Laser Michelson Interferometry. All measurements agree within stated uncertainties.

Calibration software for industrial instruments. – In order to ease the use of SPM for practical use, correction and calibration software has been developed [5]. This is particu-

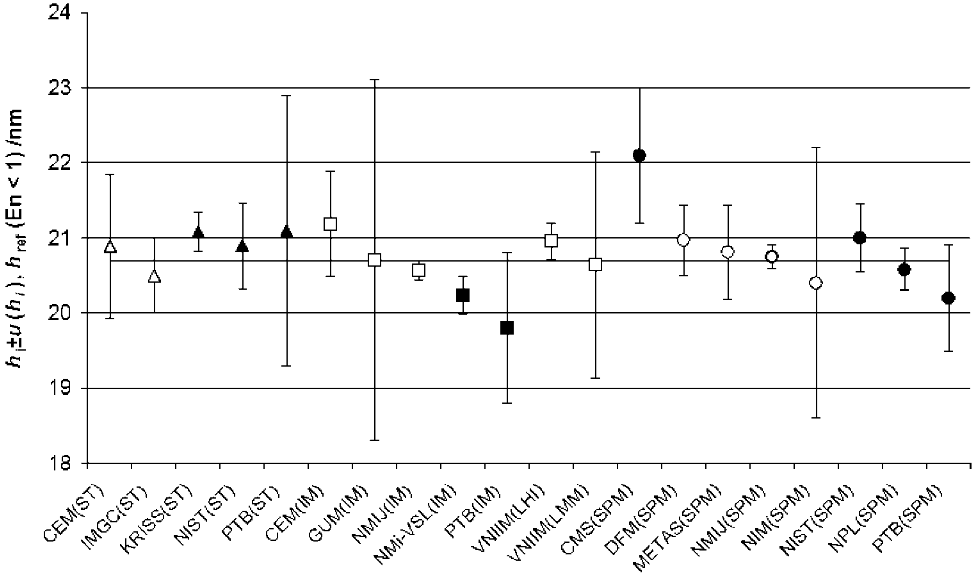


Fig. 4. – Interlaboratory comparison arranged by the CCL. The specimen has a nominal height of 20 nm [4].

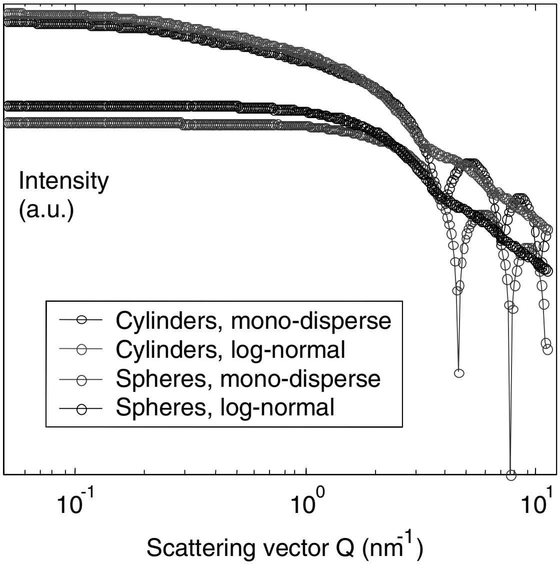


Fig. 5. – Small-angle neutron diffraction from nanoparticles of either cylindrical or spherical shapes (courtesy of Dr. Joachim Kohlbrecher, Paul Scherrer Institute).

larly useful for non-metrology SPMs; but it may also be used in connection with advanced apparatus to diminish the ever remaining measurement errors. And in particular, it may be used by users, who do not have direct access to an SPM, to analyse data.

Setting up nanofacilities for future needs. – Anticipating the big demand for nanometrology services that the massive effort in nanosciences is likely to lead to, a number of NMIs are setting up very comprehensive facilities equipped with traceable measurement facilities.

Surface physics application in vacuum. – A few metrology SPM facilities have been set up with full surface physics facilities including high vacuum possibilities. Such equipment will be used in connection with scientific research; but they are unlike to achieve large industrial attention.

Alternative methods to SPM. – Electron microscopy is a popular method for microscopic analysis of matter; and it is widely used in nanoscience and technology. However, when it comes to metrology, both scanning electron microscopy (SEM) and transmission electron microscopy (TEM) have proven difficult to calibrate, and only very few NMIs have overcome that barrier. An exception is the SEM in the "cross" mode, but the backlash of this method is that is requires that the sample is cut across and thereby destroyed.

Because of the importance of nanoparticles, and the general lack of quantitative methods to characterise them, it is worth mentioning that small-angle neutron scattering offers an interesting opportunity to measure their sizes and shapes. Figure 5 shows the scat-

TABLE II. – *The development with time of typical sizes in selected microelectronic components: Dynamic Random Access Memory (DRAM), Micro Processor Units (MPU), Application Specific Integrated Circuit (ASIC), and Polycrystalline Silicon.*

Typical sizes (nm)	Remark	2005	2010	2015	2020
MPU/ASIC	pitch	180	90	50	28
DRAM	pitch	160	90	50	28
Polycrystalline Silicon	pitch	152	80	46	26
MPU	Physical gate length	32	18	10	5

tered intensity of neutrons from cylinders and spheres of similar size in the nm region. It demonstrates that when the sample consists of close to identical cylinders or spheres ("mono-disperse"), the signal is distinct and can be used do characterise the shapes quantitatively. But if the sizes follow a broad statistical distribution, for instance the so-called "log-normal" distribution, the signal broadens out and scales with the average size of the particles. Hence, the small-angle neutron technique is an interesting alternative to traditional techniques to measure sizes and shapes of nanoparticles.

3. – What nanotechnology needs

The miniaturisation of manufacturing as demonstrated in fig. 1 gives rise to smaller and smaller dimensions to be measured; and this evolution has been quantified by a number of authors (see refs. [6] and [7]. Table II summarises some dimensions in semiconductor components and devices, as well as their foreseen evolution until 2020. It appears that pitches or repetition cycles of random access memory devices, microprocessors, and specific integrated circuits are currently approaching 100 nm and they will shrink a factor of 6 during the period; gate lengths are about one third of the pitch.

As mentioned above, dimensions have to be measured significantly better than their actual sizes. In table III are listed some measurement requirements in semiconductor production up to 2020. It illustrates that the requirements from the semiconductor industry for metrology range from about 1 nm to 15 nm, and these tolerances will diminish by a factor of 6, when we reach 2006.

TABLE III. – *The development with time of typical critical dimensions (CD) in electronic devices.*

Critical dimension (nm)	Remark	2005	2010	2015	2020
Wafer	XY-overlay	15	8	4.5	2.5
Wafer control	Dense lines	8.8	4.7	2.6	1.5
Line width	Roughness	2.6	1.4	0.8	0.5
Wafer precision metrology	Isolated lines	0.67	0.37	0.21	0.12

These requirements may be compared to the results of the comparisons in fig. 4. This figure shows that the most competent laboratories of the world reach an equivalence of about 1 nm in their measurements of a height of 20 nm. Another comparison with a grating with pitch of nominally 290 nm also gave an equivalence of about 1 nm when using SPM, whereas for such a simple structure one can get much better correspondence if one uses optical interferometers or diffractometers. These examples indicate that current best practices for measurements for nanotechnology are satisfactory for the industrial needs. However, there is at present no indication how this satisfactory result at the level of national metrology institutes is reflected in practical measurements at the industrial production level.

4. – Practical examples

The above considerations cannot be taken too literally. There is no universal practice for the calculation of uncertainties, critical dimensions and tolerances; only within the metrology community has the quantitative statement of uncertainties been harmonised according to the GUM method described in ref. [3]. It is therefore of interest to look at the results of measurements on "real" samples. Below we discuss two such examples related the roughness of silicon wafers and to the general characterisation of an optical grating.

Example 1: roughness of a silicon wafer. – Figure 6 shows the roughness of a silicon single crystalline wafer in the three stages of manufacturing from the cast ingot. First the ingot is cut by a diamond saw into wafers, where the wafers each have a surface shown in the upper part of fig. 6. The surface is typical of a cutting process, and the picture has a maximum height difference of about 2 μm. Subsequent etching alters the height span by almost 3 orders of magnitude; and the final polishing takes away the remaining "islands" and gives an apparent flat surface of maximum height differences of 3.4 nm. In line with the last row of table III, this is sufficient for the wafer to be used in the production of chips. The results may be quantified by the roughness parameter $S_q = 0.25\,\text{nm}$.

This example demonstrates that for nanotechnology related to roughness measurements practical measurements can be performed that are fully satisfying the requirements of the industry. The results of the comparisons mentioned above further ensure global equivalence of the measurements. Finally, because of the flatness of the surface after cutting such measurements do not depend very critically on the shape of the scanning tip, and they are fairly easy to perform.

Example 2: Parameters of an optical grating. – Figure 7 shows an SPM picture of an optical grating intended for information processing in telecommunication. This is an example of an optical equivalent of an ASIC listed in table II. Because of the high aspect ratio of the device, the results of the measurements are very dependent on the size and shape of the scanning tip; and it therefore requires special care to correct the measurements for the effects of the tip shape in order to derive the correct dimensions of the measured object.

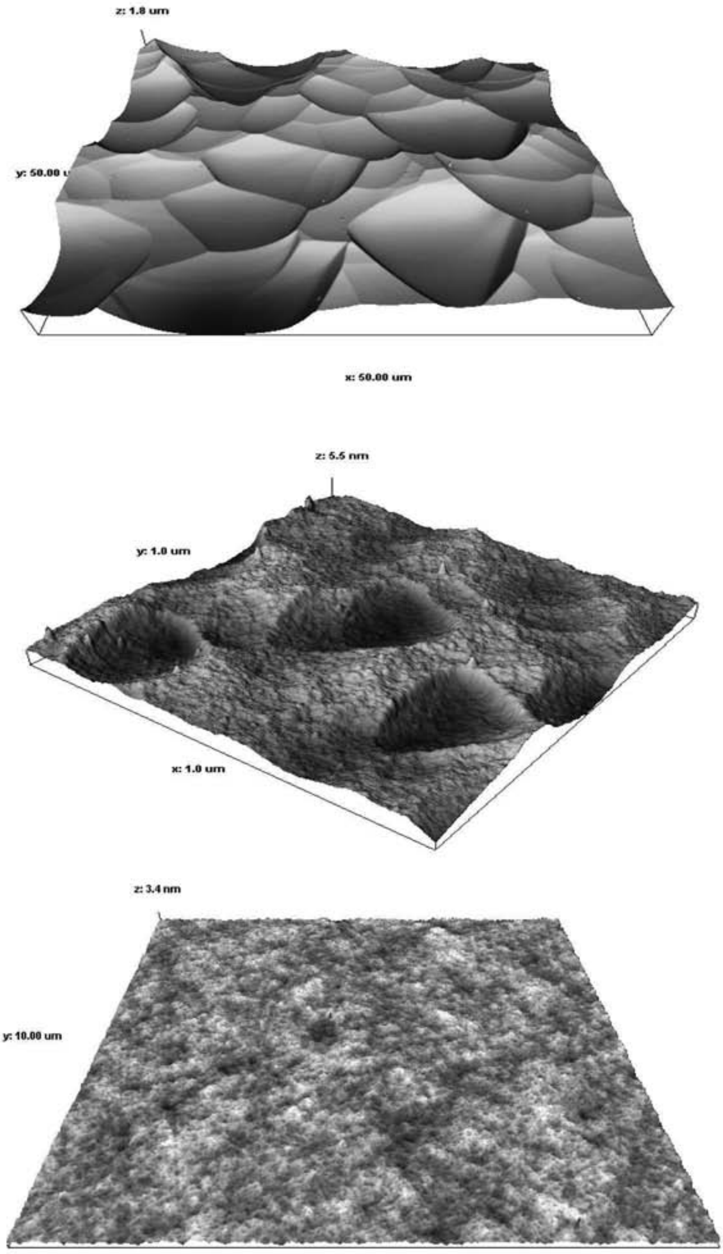


Fig. 6. – SPM measurements of a single crystalline Si wafer at various stages of its preparation. Upper part: The wafer after diamond cutting, XYZ-span: 50 μm × 50 × 1.8 μm. Middle part: The wafer after etching, XYZ-span: 1.0 μm × 1.0 μm × 6 nm. Lower part: The wafer after final polishing, XYZ-span: 1.0 μm × 1.0 μm × 3 nm. Lower part: (courtesy of Topsil Ltd.).

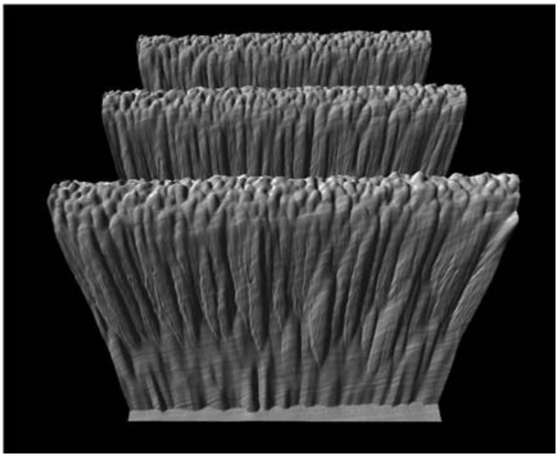


Fig. 7. – SPM picture of an optical grating (courtesy of Ibsen Ltd.). The pitch is about 2 μm.

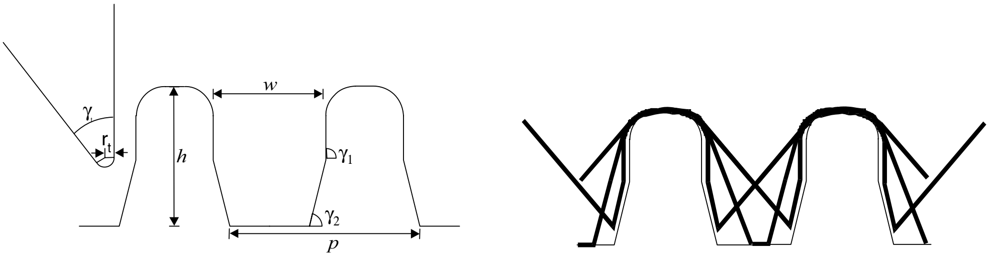


Fig. 8. – Principle of the method to derive at the true sample shape, by deconvolution of the tip shape through a series of tilted scans.

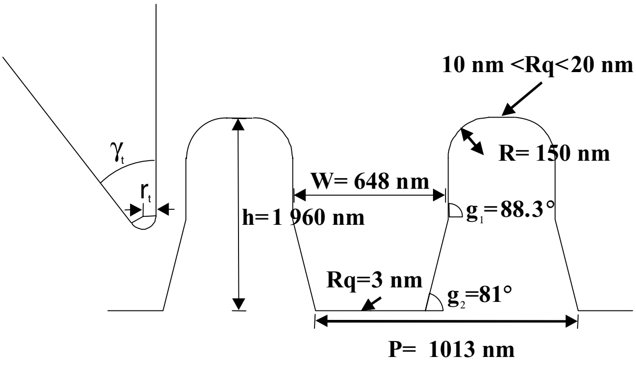


Fig. 9. – Parameters derived through the "tilted scan method" for the optical grating shown in fig. 7.

A method to derive the true shape of specimens with high aspect ratios has been described in refs. [8] and [9], and the principle is described in fig. 8. Assuming that both the sample and the tip have stable shapes, one makes a series of scans with different angles between tip and sample; and from this one is able to deduce the deconvoluted shape of the sample. A result of this procedure is shown in fig. 9. It shows that detailed information about dimension, angle, and roughness can indeed be derived, when proper corrections for the tip shape are done. This example serves as a practical illustration that the requirements of practical nanotechnology are currently satisfied by state-of-the-art metrology.

5. – New developments. Optical diffraction microscopy

From the above it appears that the scanning probe techniques are able to address most of the needs of today's nanotechnology, when properly treated as a metrology tool. This situation is likely to prevail in the coming years with the developments foreseen in tables I and II. However SPM is a versatile technique, it needs special working circumstances and it is a slow technique; therefore it is not suitable for general industrial measurements and measurements in production control.

Optical diffraction microscopy (ODM) is a method that potentially offers very short measuring times and a more robust instrumental set-up than SPM. In ODM one measures the intensity of the diffracted field as function of frequency and uses an inverse algorithm to reconstruct a map of the surface. In order to exploit this, a particular version of ODM, the LuKa Optoscope® has been developed [9], and a side view of the instrument is shown in fig. 10. Light from the white light source is sent through the grating to be characterised, and the diffracted signal is reflected through focussing lenses onto a spectrometer for frequency analysis. Three signals are collected:

- η_0, the zeroth-order diffraction (or transmitted beam)
- η_+, the sum of all orders of diffraction that are scattered to the right, and
- η_-, the sum of all orders of diffraction that are scattered to the left.

However only two signals are retained for processing, namely:

- η_0, the non-diffracted beam, and
- η_+/η_-, the ratio of the two diffracted beams. This signal reflects the asymmetry of the grating.

Figure 11 shows a top view of the reflector. It consists of two mirror-symmetric ellipsoidal parts, which integrate and reflect η_+ and η_-, respectively, onto two focussing lenses. The two signals reach the spectrometer though optical fibres and a switch. The third signal η_0 reaches the spectrometer by passing through the reflector, as shown in fig. 10.

Figure 10 shows LuKa Optoscope® in the transmission mode. Dependent on the refractive index of the materials of the grating, it may be advantageous to change the set-up to reflecting mode in order to avoid long passages through the sample.

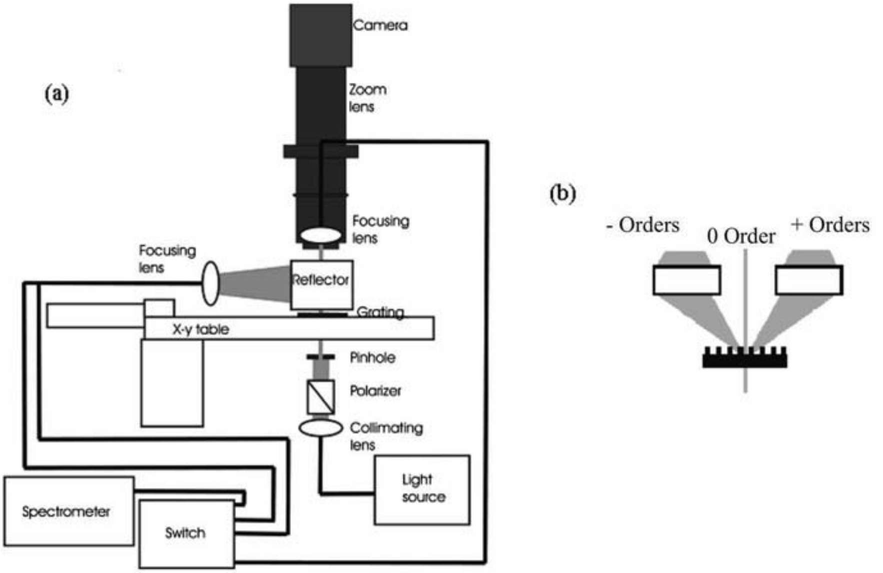


Fig. 10. – Side view of Luka Optoscope® shown in the transmission configuration. Light is sent through the grating from below; the zeroth-order diffracted (transmitted) light passes through the reflector and is sent to the spectrometer for analysis through the switch. The integrated diffracted light is collected in the two pieces of the reflector on its way to the spectrometer. An XY-table and a video camera allow proper positioning of the grating. The right part of the figure shows the separation into transmitted, right scattered, and left scattered light.

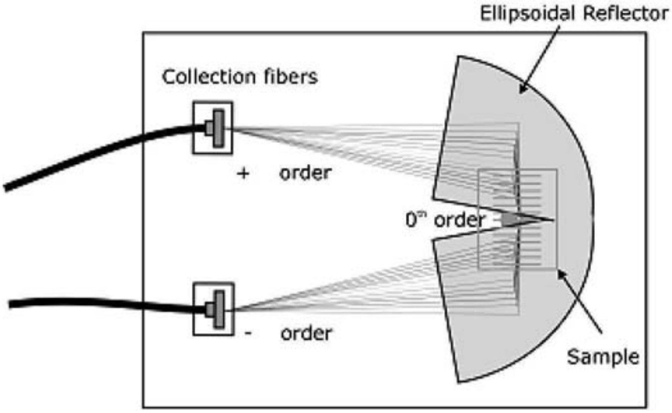


Fig. 11. – Top view of the reflector of Luka Optoscope®. The sinusoidal reflector consists of two mirror-symmetric parts. They integrate the beams that are diffracted to the two sides, respectively, whereas the wedge-shaped separation allows the transmitted beam to reach the collecting fibre above the sample.

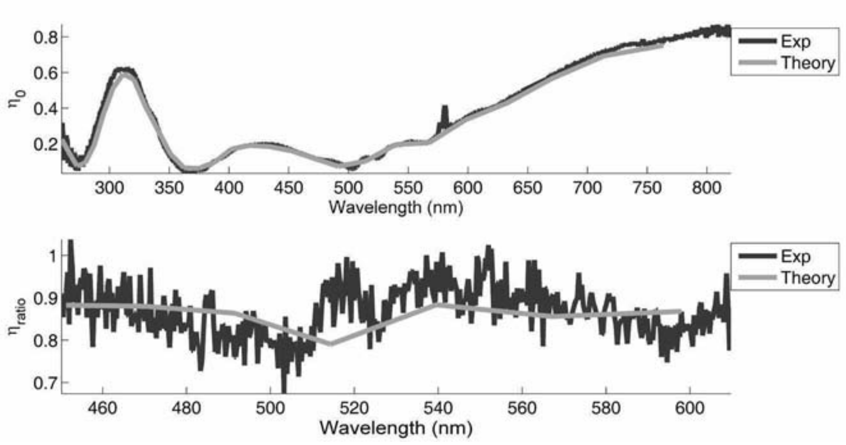


Fig. 12. – Typical experimental output from optical diffraction microscopy (black line). The theoretical grey line is obtained through a least-square fit between the experimental data and the results of a model for the grating with varying parameters.

Figure 12 shows the experimental signals after spectral analysis in the region 250 nm to 850 nm for the transmitted light, and in the region 450 nm to 610 nm for the ratio between the left and right scattered beams from an asymmetric grating. In order to model the results theoretically, one assumes a model for the grating profile. The model is characterised qualitatively by being symmetry or asymmetry, and by having a number of different materials used in the lithographic process of manufacturing the grating. It is characterised quantitatively by magnitudes of parameters such as height, width, pitch, and side angles of the profiles that form the grating. By a specific choice of parameters and using ordinary diffraction theory, a theoretical picture is generated. Using a least-square fit between experimental and theoretical data, one may optimise the theoretical parameters and get a reliable quantitative representation of the grating. The result of such a fit is also shown in figure 12.

In order to verify the measurements obtained by ODM, a comparison was performed between ODM, SPM and scanning electron, microscope, SEM. The SEM was operated in its "cross-cutting" mode, which allows a semi-quantitative comparison with the two other techniques. The grating that was chosen was of the kind shown in fig. 7 to 9. The results are shown in table IV and in fig. 13, and they demonstrate that the techniques are fully compatible. It should be noted that in table IV the results from SPM were derived somewhat differently from the values in fig. 9. In the original SPM measurements of fig. 9 the parameters were given as best representations of the data without assuming a trapezoidal parametric shape of the grating; but in table IV the SPM data were fitted to the parameters of the same model for the grating, which was assumed in the ODM analysis. As the SEM data are difficult to quantify, fig. 13 illustrates the correspondence between the results of the ODM fit and the cross-section SEM picture.

Table IV. – *Comparisons between results from optical diffraction microscopy, scanning probe microscopy, and scanning electron microscopy. The sample was shown in figs. 7 and 9. Uncertainties are shown in brackets. h_0 is the height of the line that separates the two angles of slope in the model structure, and because of the closeness of γ_1 and γ_2 it is difficult to determine.*

Method	γ_1 (degrees)	γ_2 (degrees)	h (nm)	W (nm)
ODM	80	87	1945	652
SPM	80.9	88.3	1950	653
SEM				649

Apart from giving an opportunity to characterise optical gratings much faster than can be done by SPM, when a proper model for the grating can be established, it is interesting that ODM may be extended beyond profiles that can be measured by SPM techniques. Two examples are shown in fig. 14, and they are:

- Gratings using *different materials.* Because of their different (complex) refractive index, over layers of different materials used in the making of a grating, they will give rise to different diffraction profiles in the ODM results. This has successfully been used to characterise gratings with several over layers.

- *Embedded structures.* Because light penetrates matter as opposed to the SPM-tip, "internal" or embedded structures may be characterised by ODM. Figure 14 show a grating that consists of alternating layers of GaAs and AlGaAs (literally $Al_xGa_{1-x}As$).

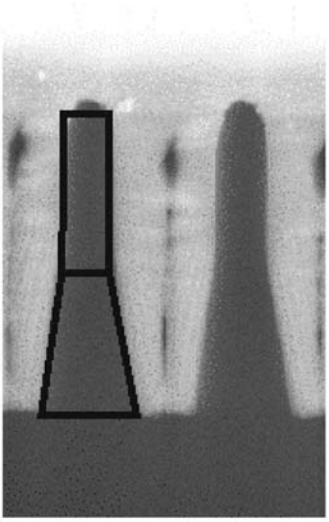


Fig. 13. – Comparison between ODM (solid lines) and SEM from a grating shown in fig. 7.

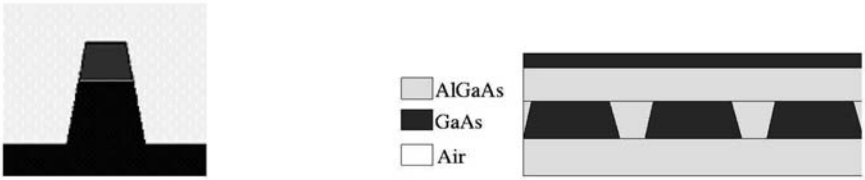


Fig. 14. – Grating profiles that can be characterised by ODM. Left: general two-component grating. Right: two-component embedded structure, where GaAs is present as inclusions of AlGaAs.

Hence, ODM appears to perform satisfactorily as a metrology tool. It is easy to use, performs must faster that scanning instruments, and it is able to measure features that are invisible by SPM. It therefore provides an interesting industrially applicable method for the characterisation of periodic structures in nanometrology.

6. – Summary

During the past decade several national metrology institutes have taken a systematic approach to nanometrology. Traceability to the realisation of the meter has been established, uncertainties have been calculated according to the GUM method, and interlaboratory comparisons have been performed on simple objects in the nm region. The result is that global equivalence is established at the level of national institutes, but only in few cases has these capabilities been demonstrated on practical production objects.

The universal instrumentation is based on the scanning tunnelling microscope and is referred to as scanning probe microscopy. Since this instrument delivers data that require a high degree of correction, calibration software has been developed to ensure properly corrected data, and the comparisons demonstrate that this is indeed well implemented.

As the measurements become validated, there is an increasing need for methods based on well-defined concepts and standardised terminologies to allow comparability of measurements at the applied level. Further, there may be an emerging need for practical temperature measurements of surfaces when large objects are required to perform within nanometer tolerances.

The measurement capabilities that have been demonstrated at national levels appear to meet the current and foreseen needs of industry. For instance, today's semiconductor components typically demand measurement accuracies at the level of 10 nm with exceptional cases of 1 nm and this is indeed possible to achieve with a well-operated SPM. Of course, it will be a challenge follow the development of a six-fold further miniaturisation until 2020, but it does not seem insurmountable.

The greater challenge for metrology in supporting nanotechnology will be to disseminate the measurement techniques that have been developed at national metrology institutes and research institutions to industry, and to invent practical instruments that can perform in an industrial environment at a speed that allows production control. A natural solution to this is to develop optical methods to measure in the nanometer regime, and one example of this is given in this paper.

* * *

This work has been performed during the last 17 years and has largely been funded through performance contracts and innovation contracts with the Danish Ministry of Science, Technology, and Innovation. Further the support form the EU-framwork programs is gratefully acknowledged. Many co-workers have contributed to the work, and I would particular like to thank L. Lithén Madsen, J. Garnæs, A. Kühle, M. Holmberg and P. E. Hansen from DFM, L. de Chiffre and H. Nørgaard Hansen from Technical University of Denmark, G. Wilkening and H. Kunzmann from Physikalisch-Technische Bundesanstalt (Germany), M. Stedman and R. Leach from National Physical Laboratory (United Kingdom), A. Balsamo from INRIM (Italy), as well as J. F. Jørgensen from Image Metrology Ltd.

REFERENCES

[1] Taniguchi N., *Ann. CIRP*, **32/2** (1983) 573.

[2] Binnig G., Rohrer H., Gerber C. and Weibel E., *Phys. Rev. Lett.*, **49** (1982) 57.

[3] *Guide to the expression of uncertainty in measurements* (BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, ISBN 92-67-10188-9) 1995.

[4] The results of several comparisons are available from the CIPM Key comparison database, hosted by Bureau International des Poids et Mesures (`www.bipm.org`). The identifier of the comparison shown in fig. 4 is CCL-S2.

[5] A widely used software is SPIP®.

[6] International technology roadmap for semiconductors - metrology (2005) `http://public.itrs` (January 2006).

[7] Hansen H. N., Carneiro K., Haitjema H. and De Chiffre L., *Ann. CIRP*, **55** (2006) 721.

[8] Garnæs J., Kühle A., Nielsen L. and Borsetto. F. Nanoscale, in *Calibration Standards and Methods* (Wiley-WCH Verlag, Weinheim) 2005, pp. 193-204.

[9] Garnaes J., Hansen P.-E., Agersnap N., Daví I., Petersen J. C., Kühle A., Holm J. and Christensen L. H., *Proceedings of SPIE - The International Society for Optical Engineering*, edited by Duparré A., Singh B. and Gu Z.-H., **5878** (2005) 587803-1.

Metrology for chemical measurements in the environment

M. J. T. Milton

National Physical Laboratory - Hampton Road, Teddington, Middlesex, TW11 0LW, UK

Introduction

The application of the principles of metrology to chemical measurements has been an important activity for many measurement scientists over the last 15 years. The success of this enterprise is in part due to the application of a number of fundamental principles that are well known in metrology that have now been refined to apply specifically to chemical measurements. This lecture gives examples of how the principles of metrology have been applied to measurements made of the environment, which gives some good examples of the benefits of this approach.

1. – Measurement of composition

Chemical analysis is almost exclusively concerned with the determination of composition. Over many centuries, a huge number of methods have been developed that determine the composition of different species in different types of sample. The great variety of these methods is reflected in the large number of different quantities used to characterise composition. Some of these are listed in table I. There is a logical structure behind the definitions of these quantities. For example, they can be grouped into: "fractions" which describe how much of the total property of a sample is contributed by one of its constituent substances; "concentrations" which describe the ratio of one measure of a single substance (for example the mass or amount of substance) to the total volume of the mixture; and "contents" which describe the ratio of one measure of a substance

Table I. – *Quantities used for the measurement of composition (after ref.* [1]*).*

Name	Symbol	Definition	SI unit
Mass fraction	w	$w_i = m_i / \sum m_j$	kg/kg
Volume fraction	φ	$\varphi_i = V_i / \sum V_j$	m^3/m^3
Amount fraction	x	$x_i = n_i / \sum n_j$	mol/mol
Mass concentration	γ	$\gamma_i = m_i / V$	kg/m^3
Volume concentration	σ	$\sigma_i = V_i / V$	m^3/m^3
Amount concentration	c	$c_i = n_i / V$	mol/m^3
Molality	b	$b_i = n_i / m_{solv}$	mol/kg
Volume content	κ	$\kappa_i = V_i / m$	m^3/kg
Amount content	k	$k_i = n_i / m$	mol/kg

(volume or amount of substance) to the total mass of the mixture. The quantity "molality" describes the amount of solute divided by the mass of the solvent and is used when it is necessary to distinguish between the total amount of solution, and the total amount of solvent.

An important property of some of these quantities is that they are "complete" in the sense that they provide enough information to characterise a sample containing two components in full, in amount of substance units, whereas others require additional information, for example the density or relative molecular mass. All of these quantities are different, and are required for use in specific applications.

A common characteristic of all of these quantities that are used to measure composition is that they are ratios and are therefore "intensive quantities". That is, they are the same for sub-samples of the whole as for the whole. This can be considered as being a requirement for their being measures of composition, since, by definition, the composition of any sub-sample should be the same as that of the whole. These intensive quantities are created by the combination of two or more extensive measurements, which may or may not be of the same type of quantity. In the case where the numerator and the denominator are measured by the same type of quantity (for example amount fraction or mass fraction), it is possible to perform a "cancellation" of the units in order to produce an intensive quantity that is "dimensionless". Although this approach is mathematically correct, it is not to be universally encouraged, because it robs such types of quantity of the power to be used in dimensional analysis, which is of real value in the process of checking mathematical manipulations.

This emphasis on the measurement of composition, with its consequent dependence on intensive quantities is one of the strongest reasons why the application of metrology to chemistry has developed differently to the application of metrology in other parts of science. This is because, at a superficial level, it is easier to do metrology with intensive quantities particularly when they are also dimensionless. For example, it is possible to use quantities that are fractions (or ratios) of the same extensive quantities without any reference to internationally recognised references, since the experimental process can be configured as a set of operations that cancel out the unit itself. This principle can be taken a step further when well-known conversion factors, such as the relative molecular mass, are used to convert dimensionless ratios of mass into quantities that are amount fractions, or amount contents. Such quantities can, in principle, be determined without any external reference.

An example of this way of thinking about the measurement of a ratio is in the preparation of a solution with a known mass ratio. This can be carried out such that the solute and solvent are each weighed using equipment that is calibrated with respect to the same laboratory reference mass. When this is the case, the mass of the reference mass piece itself will be cancelled when the mass ratio of the solution is calculated. Consequently, there is no direct requirement for the value of the mass piece to be traceable to any external reference. A major limitation of this approach is evident when considering the preparation of a solution by combining a solute with a solvent of very different masses. In this case, the weighing equipment would be calibrated with more than one laboratory reference mass. Although it might be possible to provide a link between these laboratory reference masses by means of some independent comparisons, it is generally more cost-effective to depend on the masses each being calibrated externally with respect to the same reference, which is the SI.

This discussion seems to support the view that metrology in chemistry ought to be "straightforward", because "it is largely concerned with ratios and these ratios have no direct requirement for traceability". In practice, there have been unfortunate consequences of relying upon this approach. In particular, the omission of any link to external references, even when it is strictly valid, leads to a general weakening of the recognition of the importance of the link with the SI. Consequently, attention has been diverted from other aspects of "good metrological practice" such as the importance of estimating the uncertainty of measurement results and the role of traceability in underpinning measurement results that are stable over time, comparable between laboratories and coherent between different measurement methods. All of these trends serve to weaken efforts to gain a stronger understanding of the measurement process itself, which is fundamental to all types of experimental science.

This discussion of ratios raises the very interesting question: why is the mole needed? I believe that there are very strong reasons why the mole is needed, but they are not specifically because chemistry requires measurement results to be expressed in mol!

2. – The mole

The mole was adopted as a base unit of the SI in 1971 [2] with the following definition:

> "The mole is the amount of substance of a system that contains as many elementary entities as there are atoms in 0.012 kilogramme of carbon-12."

The mole is fundamentally a reference to a specified number of entities, and as the definition states it is applied to a "system" in the same way that the laws of thermodynamics are referred to a "system". The definition of the mole naturally leads to the definition of the Avogadro number as being the number of elementary entities in the mole. Since the mole is so closely related to a number, there is an argument that the mole itself should not be introduced as a base unit of the SI, because a “number” does not constitute another independent “dimension” within the set of base units. This argument has some validity, but is only an extension of the viewpoint articulated above, that chemical measurement “is largely concerned with ratios and these ratios have no direct requirement for traceability”. It also leads to the loss of any opportunity to use dimensional analysis when dealing with the results of chemical measurements. This was, in part, the reason for the adoption of the mole into the SI, together with the need to resolve confusion arising from the use of both g-mol and kg-mol, and both carbon-13, oxygen and oxygen-16 as references for the definition of the atomic mass unit.

This brings us back to the fundamental issue about measurements in chemistry, which is that most chemical measurements are intended to elucidate the composition of a sample and that composition is measured by intensive quantities. Yet, the base unit of the SI associated with chemistry is the mole, which is an extensive quantity. This explains why the mole itself is not of enormous importance in performing analytical chemical measurements. Its importance lies in the opportunity to combine it with other extensive quantities to form quantities that are intensive, but it is rarely used to express the final result of a measurement on its own.

A further “problem” relating to the use of the mole is that the definition of the mole itself does not readily lead to a “realisation”, hence, until the recent effort to determine the Avogadro Constant by the XRCD method [3], the mole has not featured amongst the most important challenges of “fundamental metrology”. Therefore, when the CCQM was founded in 1995, the initial task set for the committee was not to consider how the mole might be realised. The challenge was to consider how a “hierarchical system” could be developed for chemical measurements and whether it would have any practical benefits. This starting point led to what may have been a slightly confused consideration of how “measurements could be made that were expressed in mol”. It was soon realised that the challenge was a great deal more than just “making measurements in mol” and the discussion was broadened to how to produce measurement results that were “traceable to the SI”.

3. – Primary methods of measurement

It was agreed by the CCQM that measurement results that were traceable to the SI could be achieved by use of a “primary method of measurement”. Therefore, a definition was developed from the VIM [4] definition of a “primary standard” which made a direct association with the principle of primary thermometry in temperature metrology [5]. The definition that was eventually agreed was that:

> “A primary method of measurement is a method having the highest metrological properties, whose operation can be completely described and understood, for which a complete uncertainty statement can be written down in terms of SI units.”

The three criteria introduced by this definition highlight that a primary method should: have the highest metrological properties, be completely described and understood, and have a complete uncertainty statement in SI units. The first of these emphasises that a method that is in any sense trivial should not be considered to be primary. The second two properties relate to the unique defining property of a primary method —that it cannot make use of "empirical" or "instrument-dependant" factors in its operation. It can only make use of quantities that are measured in terms of SI units and constants or material properties that have values known in SI units.

4. – The "grande salle" metaphor

Let us summarise the argument so far; measurement in chemistry is largely concerned with the determination of quantities that are ratios. These ratios are often dimensionless and because the link to the mole is almost always made by measurements of mass converted using the RMM, the immediacy of the link to the mole is diminished. However, the application of metrology to chemistry, more so than many other areas of measurement science, makes a lot of use of *ad hoc* local references and standards which, because of the lack of any immediate link to the mole, are overlooked when issues of traceability are considered. Some clarity is brought to this issue through what I refer to as the "grande salle metaphor". Consider a measurement method, and set it up with all of the necessary measurement equipment within, for example, the grande salle at the BIPM. Also within the grande salle is a copy of the SI brochure, and a number of texts referring to the laws of physics. If a sample can be passed into the grande salle and the result of a chemical analysis can be generated, for example, a measurement of the composition of the sample, without reference to any other measurement made outside the grande salle, then the method within the room is operating as a primary method of measurement. This metaphor appears to be rather prosaic, but it is useful to illustrate the principle of the primary method.

The principle illustrated by the grande salle metaphor was reinforced by addition of a note to the CCQM's definition of the primary method of measurement [5], that:

> "A primary direct method can be used to make a measurement that is traceable to the SI without the use of an external reference of the same quantity (for example gravimetry or coulometry)".

This can be considered to be a defining property of a primary method —that it operates without reference to any standard of the same quantity. Whilst this is certainly an important property of some primary methods, it is necessary to recognise that there are also methods that fulfil the central definition of being primary methods, in terms of being completely described and understood, but only fulfil this additional requirement when we consider them to be used to measure ratios. In other words, they measure the ratio of a quantity in a sample to the same type of quantity in some standard or reference. The concept of a "primary ratio method" was introduced in a second note to cover such methods:

> "A primary ratio method: measures the value of a ratio of an unknown to a standard of the same quantity; its operation must be completely described by a measurement equation."

Returning briefly to the "grande salle" metaphor, a further question that may be asked in the case of chemical measurements is whether a copy of the IUPAC "Green Book" [6] which includes a compendious collection of chemical and physico-chemical property data is also permitted within the grande salle. The answer to this question must depend on a detailed consideration of each datum within the Green Book and requires a careful consideration of whether it has been determined in a way that leads to it having a value (and uncertainty) that are traceable to the SI.

Following the agreement of the extended definition of the primary method of measurement discussed above, the CCQM went on to consider a number of methods that might had the potential to fulfil the new definition. They identified three methods that measured extensive quantities: gravimetry, coulometry and the colligative properties, and one that measured ratios (or intensive) quantities: isotope dilution mass spectrometry (IDMS), which is described in greater detail below. In order to clarify the relationship between the direct and ratio methods, a further note was added to the definition:

> "A primary direct method can be combined with a primary ratio method to produce measurements that retain their primary qualities (for example IDMS with a gravimetric assay of the pure spike)."

Figure 1 illustrates a conceptual hierarchy of how primary direct and primary ratio methods may be combined to form a link between the SI and practical measurements. The hierarchy is not a simple description of the stages providing traceability of a measurement result to the SI, but is an indication of the functions that must be brought together to produce a real one.

Figure 1 shows how primary direct methods are used to prepare calibration standards from materials with measured purity. The link between these standards and samples in complex matrices is made by use of a primary ratio method. Subsequent measurements of the ratio of the amount of substance of these standards in complex matrices to a real (unknown) sample is made by what is designated in fig. 1 as a "secondary" method. This is a method that does not meet the definition of a primary method in full. For example, methods such a chromatography (gas, ion or liquid) and mass spectrometry have the capability to distinguish between chemical species and usually include empirical terms in their measurement equation.

We now consider two methods with the potential to be primary, in order to examine how they operate in principle and also to show how they are used. This will show how well the principle of the primary method of measurement works in practice.

5. – Isotope Dilution Mass Spectrometry (IDMS)

One of the most fundamental principles used in chemical analysis is that of "isotope dilution". It relies on the fact that different isotopes of the same element exhibit

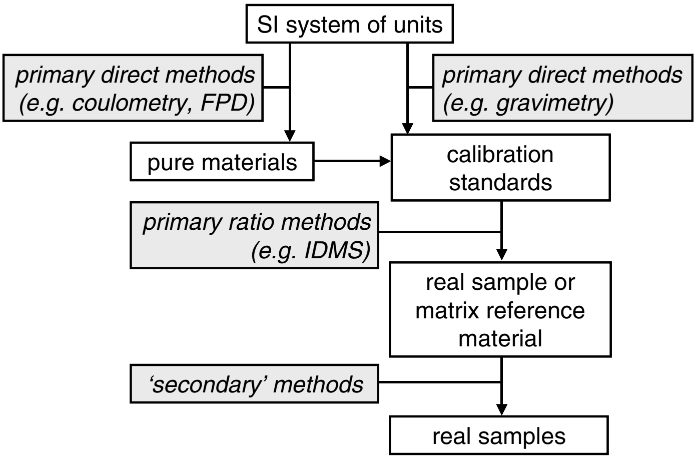

SI system of units
primary direct methods (e.g. coulometry, FPD)
primary direct methods (e.g. gravimetry)
pure materials
calibration standards
primary ratio methods (e.g. IDMS)
real sample or matrix reference material
'secondary' methods
real samples

Fig. 1. – A hierarchical representation of the relationship between primary direct and primary ratio methods.

extremely similar chemical behaviour. When combined with analysis by mass spectrometry, isotope dilution forms the basis of the analytical method known as Isotope Dilution Mass Spectrometry (IDMS).

The principle of isotope dilution is simple and elegant. It relies on mass spectrometers, which are widely available with sufficient stability and resolution and therefore has the potential for application across analytical measurement. Although it has the potential to give results that are independent of both the matrix and the analyte, most IDMS methods require the use of isotopic reference materials to overcome inherent limitations in their accuracy and biases in the operation of mass spectrometers. This raises the question of whether it is practically possible to make use of a method that is recognised as having the potential to be primary in a way that actually meets the definition. Since gas measurements present some of the greatest challenges to isotope dilution, it is illustrative to consider an example in greater detail.

The simplest IDMS method is known as the direct (or "one-step") method [7]. The measurement equation is

$$\frac{N_\mathrm{s}}{N_\mathrm{sp}} = \frac{R_\mathrm{sp} - R_\mathrm{b}}{R_\mathrm{b} - R_\mathrm{s}} \cdot \frac{\sum R_\mathrm{s}}{\sum R_\mathrm{sp}}\,, \tag{1}$$

where N_s = the amount of substance in the unknown, N_sp = the amount of highly enriched spike added to the unknown, R_s = the isotope ratio of the unknown, R_sp = the isotope ratio of the highly enriched spike, R_b = the isotope ratio of the blend formed from the addition of the spike to the unknown.

The summations on the right-hand side of (1) refer to an addition over all isotopes of the species being measured. Writing the measurement equation in this form serves to emphasise that IDMS has the potential to be a primary ratio method since the mea-

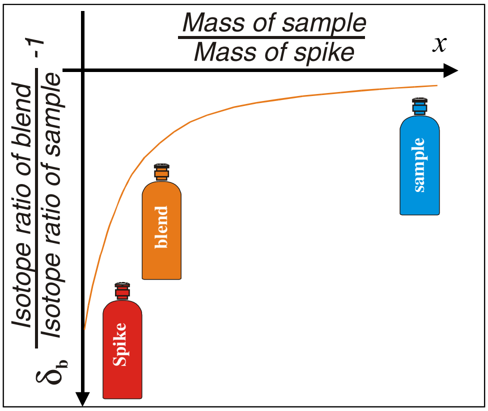


Fig. 2. – The isotope dilution curve. All blends from the same sample and spike have isotope ratios that fall on this curve. The figure shows the isotope dilution curve for a spike that is isotopically deleted with respect to the unknown (natural) sample.

surement of the isotope ratios on the right-hand side gives rise to the evaluation of the amount ratios on the left-hand side. The determination of the amount of substance in the unknown (N_s) requires a measurement of the amount of enriched spike material (N_{sp}) and information about its purity. It then becomes an example of a primary ratio method combined with a primary direct method —as discussed above.

Additionally, reference materials with certified isotope ratios are required for the highest accuracy applications in order to calibrate the isotope ratio scale of the mass spectrometer and every isotope of the unknown must be measured. A development of the "one-step" method is the "two-step" method, which involves an independent certification of the purity of the spike by isotope-dilution of a sample of the pure unknown at natural abundance. This has the advantage that the summations on the right-hand side of (1) are cancelled when the results of the two steps are combined and therefore it is not necessary to measure every isotope present in the unknown. However, it still requires independent certification of the linearity of the scale of the mass spectrometer.

The limitations of one- and two-step IDMS in terms of their needing certified isotopic references can be overcome by the use of the "isotope dilution curve method" (fig. 2) [8]. The principle of the method is that the isotope ratio of any blend made from a given spike and a given unknown must lie on a mathematical curve —known as the isotope dilution curve defined by

$$x = \frac{\delta_{\mathrm{sp}} - \delta_{\mathrm{b}}}{\delta_{\mathrm{b}}} \cdot Q, \tag{2}$$

where Q is a proportionality constant [8] and x is the ratio of the mass of unknown to which the mass of spike was added. The two d terms are defined by

$$\delta_{\mathrm{sp}} = \left[\frac{R_{\mathrm{sp}}}{R_{\mathrm{s}}} - 1\right] \cdot 1000 \tag{3}$$

and

$$\delta_{\mathrm{b}} = \left[\frac{R_{\mathrm{b}}}{R_{\mathrm{s}}} - 1\right] \cdot 1000. \tag{4}$$

It is now clear that by preparing two blends gravimetrically, with known ratios of the spike to a pure un-enriched standard, measurements of the corresponding isotope ratios will define the proportionality constant (Q) of the isotope dilution curve. When the unknown is blended with the spike, a measurement of the isotope ratio of this blend enables the corresponding ratio of material blended to be calculated using the equation that describes the dilution curve. The absolute mass of analyte in the unknown can then be calculated by multiplying by the mass of blend added.

Laboratory work at NPL has validated the performance of the isotope dilution curve method and demonstrated its application to the measurement of Primary Standard Gas Mixtures of carbon dioxide [8]. The work used a custom-built mass spectrometer with an array of Faraday cup detectors configured to measure the isotopes of carbon dioxide. The most important feature of the instrument was that it had four independent gas sample inputs, each of which could be precisely controlled using a bellows valve driven by a stepper-motor. This enabled the different samples to be introduced at the same pressure and at the same flow rate so that there was no change in the inlet conditions to the mass spectrometer.

The experiments used isotopically depleted carbon dioxide diluted in natural carbon dioxide. They showed agreement to better than 1 part in 100 million (relative to the stated value). The validity of this result was further assured by observations that the results of the measurement were largely independent of the drift of the instrument. Subsequent experiments at NPL have substantially increased the scope of application of the method by including a pre-separation stage using gas chromatography (GC) [9]. This enables the measurements to be made in the presence of a matrix, which is separated before the sample is introduced into the mass spectrometer. In these experiments, the matrix was nitrogen and the results of the method were also validated against the value of NPL's internationally recognised Primary Standard Gas Mixtures. The values were shown to be comparable to 0.3% and were largely limited by the repeatability of the GC separation process.

Although the work on the IDMS curve method shows that IDMS can be used as a primary method, it is by no means the case that IDMS is always implemented according to the definition of the primary method. In practice, it is often easier to correct for non-linearity in the response of the mass spectrometer by using selected reference materials certified for their isotope ratios, than to carry out a calibration to determine the dilution curve [10]. In some cases, the use of such reference materials may in turn be carried out in a way that meets the definition of a primary method of measurement hence retaining the traceability of the results. Consequently, it is important to be cautious before concluding that any particular implementation of IDMS fully meets the requirements of being a primary method.

6. – *p*H and electrochemistry

A second illustration of the way that primary methods are implemented in practice can be drawn from the case of measurements of the quantity *p*H.

The quantity *p*H was first defined by Sorenson in 1909 in terms of the concentration of hydrogen ions in solution. Subsequently, its definition has been refined to refer to the "activity" of the hydrogen ion in solution. Whilst this definition has been in widespread use for more than 70 years, it is not widely recognised that it refers to a quantity that is strictly immeasurable. This is because a single ion cannot exist on its own in solution.

In practical terms, *p*H has been defined by the IUPAC [11] as being realised at the "primary level" by an electrochemical cell known as the Harned cell. This consists of a hydrogen electrode and a silver/silver chloride electrode in a cell that uses a "frit" to prevent transfer of material between the two parts of the cell but has no liquid junction. The application of Nernst's Law to this cell leads to

$$\begin{aligned} p(a_{\mathrm{H}}\gamma_{\mathrm{Cl}^-}) &= -\ln(a_{\mathrm{H}}\gamma_{\mathrm{Cl}^-}) \\ &= (E_{\mathrm{I}} - E^0)/\left[(RT/F)\ln(10)\right] + \ln\left(m_{\mathrm{Cl}^-}/m^0\right). \end{aligned} \tag{5}$$

The two terms on the right-hand side can be measured. The first is the difference between the potential across the Harned cell when it is used with the unknown solution (E_{I}) and when it is used with a standard solution of HCl (E^0). The second term involves the concentration of the chloride ions added to the unknown buffer solution (m_{Cl^-}) and can be eliminated by an experimental procedure involving a series of measurements at reducing concentrations of chloride ions. The results of these measurements are extrapolated to the value that would be achieved at zero added chloride concentration. However, we discover that we cannot determine the quantity *p*H itself, since the quantity on the left-hand side of the equation is what is known as the "activity function". In order to determine the *p*H, which is equal to $p(a_{\mathrm{H}})$ in this notation, there is still the remaining difficulty of determining γ_{Cl^-}, the activity coefficient of the chloride ion.

The resolution of this difficulty requires the use of a convention, by which specific agreed values are chosen for two empirical electrochemical parameters used in the calculation of a value for γ_{Cl^-}. This is known as the Bates-Guggenheim convention and is the step in the operation of the Harned cell that might go outside the definition of a primary method of measurement. The status of the Harned cell as a primary method can only be retained if a suitable allowance is made in the uncertainty budget of the final result for the use of the Bates-Guggenheim convention. The IUPAC have agreed that a reasonable value for this uncertainty is 0.01 *p*H [10]. The magnitude of this estimate is put in context by the observation that the estimated uncertainty of a Harned cell measurement (excluding this contribution) is usually around 0.004 pH and the comparability between NMIs participating in key comparisons is typically 0.01 pH. Consequently, it is inevitable that almost all pH measurements are quoted on a "conventional" basis which omits this additional contribution to the uncertainty. However, the work of the IUPAC has clarified the true uncertainty in achieving traceability to the SI for a measurement of pH.

The conclusion we can draw from the example of pH measurement is that even in this field which is known to be very well described by physico-chemical laws, the quantity that we want to determine is intimately coupled to another quantity that we can only estimate. Whilst this may appear to violate the definition of a primary method of measurement, it is possible in this case to meet the definition when an appropriate allowance is made for the uncertainty of this estimate because this uncertainty recognises the extent to which this part of the measurement equation is "known".

7. – Primary methods in practice

The definition of the primary method of measurement has brought a great deal of clarity to the field of chemical measurement. However, it has never been intended to be a designation that is bestowed on measurement methods by some very exalted committee of experts. The implementation of the definition should be "fit for purpose". In the example of IDMS, we saw that it was possible to implement the method in strict accordance with the definition. But the use of IDMS in this very rigid way results in a loss of sensitivity and a loss of flexibility in the application of the method. Consequently, it is usually used —even at the highest levels of the NMIs working within the CCQM, in a way that does not fully and literally meet the definition. Similarly, in the case of pH measurement, which comes from the field of electrochemistry which is generally considered to follow well-understood laws of chemistry very closely; the measurement of pH follows the definition up to a certain point, but to carry out the measurement of the quantity that is needed (pH rather than acidity) requires the use of a "convention" that goes beyond the definition. Since this conclusion appears a little unsatisfactory from the pure metrological point of view, it is worth asking whether it should be an issue of great concern. The answer to this, I believe, is that it is not. The reason for this is that the use of primary methods of measurement to give measurement results that are "traceable to the SI" may bestow three properties on the results. These are that they should be:

– *comparable* with measurements of the same quantity made by different laboratories,

– *stable* with respect to measurements made at other times, and

– *coherent* with the results of measurements made by other measurement methods.

These three criteria form a theme that recurs in the remainder of this lecture.

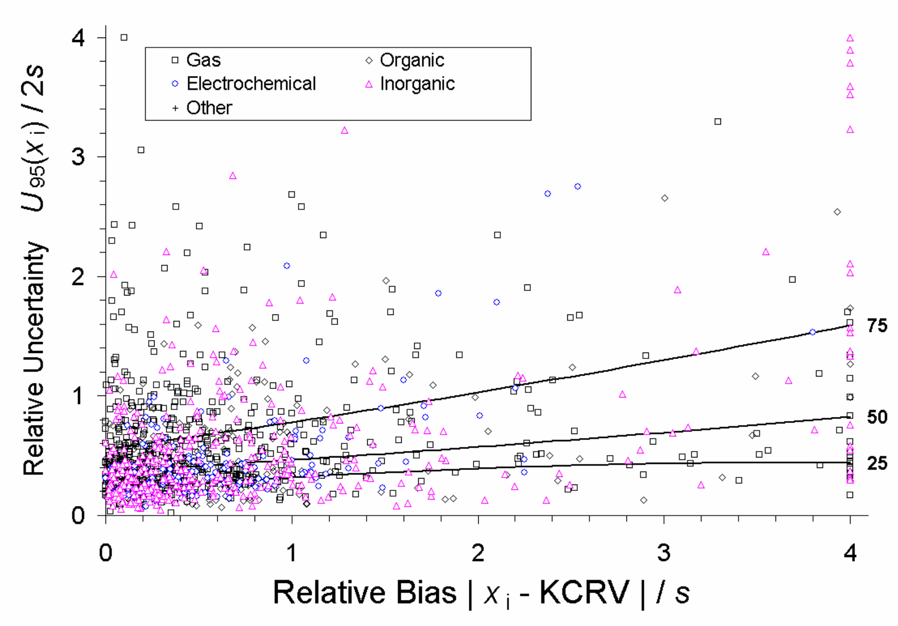


Fig. 3. – Summary of the relative uncertainty of all results submitted to the CCQM *vs.* their relative bias. The data incorporates results from four working groups: gas, organic, inorganic and electrochemical analysis. The uncertainty of each result ($U_{95}(x_i)$) and its bias with respect to the key comparison reference value (x_i-KCRV) have been normalised with respect to the standard deviation (s) of all results in that comparison. The solid lines indicate the locus below which the stated percentage of results fall. (Courtesy D. Duewer, NIST.)

It is useful to illustrate the extent to which the measurements of the NMIs are comparable in this field by reference to fig. 3, which summarises all results submitted in key comparisons organised by the CCQM.

This figure indicates that the vast majority of results are no more than a factor of two worse than the standard deviation of each set of results. Despite this strong emphasis on comparability and with it the associated stability that is implicit with these capabilities being maintained at so many laboratories, it is the final property of "coherence" that has been considered least in specifying whether a true and complete implementation of a primary methods is needed in practice. Whilst the importance of measurement results being stable over time and comparable with those from another laboratory is obvious, the need for them to be coherent with results obtained by another method is not often essential in chemical measurements at their present state of development.

However, as some examples of the use of chemical measurement for the monitoring of the environment will show, the coherence of measurements may become an increasingly important issue, and hence the use of primary methods may also become increasingly important.

8. – Measurement standards for gases

In the second part of this lecture, we will look at the application of metrology to chemical measurements made in the environment on the "local" and "regional" scales. A good starting point for considering chemical measurement on these scales is to consider the measurement of gases. This is because the emission of gases into the atmosphere is closely regulated, and their chemical interactions are studied very closely. Consequently, a substantial infrastructure has been developed to underpin these measurements, much of which is aimed at providing measurement results that are traceable to the SI.

The methods used for providing traceable measurements of gases can be conveniently divided into three groups according to the stability of the gas species over a timescale of one year (table II). Stable gases include those emitted from industrial processes and from automobiles, as well as natural gas. Partially stable gases include those measured as ambient pollutants, and unstable gases include some of the reactive components found in ambient air and emission gases. Unstable gases include the most reactive species emitted as pollutants as well as some of those found in indoor air.

The basic principle used to provide traceability for measurements of gases is to prepare standard mixtures by weighing pure component gases into cylinders. This principle takes advantage of the very good accuracy of commercially available balances and the fact that gases weigh more than might be expected intuitively. For example, 10 litres of nitrogen at 100 atmospheres weigh 0.8 kg. Addition of a mass of 20 g of pure carbon dioxide would form a mixture with an amount fraction of 10 mmol/mol. If the requirement is an accuracy of 0.05% (relative to value) for this amount fraction, then the mass of carbon dioxide must be determined with an uncertainty of 10 mg. If the cylinder containing the mixture has a mass of 10 kg, then the gravimetry must be performed with an accuracy of approximately 1 part per million to achieve the target uncertainty. This is generally achievable with single- or twin-pan balances from commercial sources.

The major source of uncertainty in this preparation process is generally the weighing itself. In some cases, there can be a significant contribution due to the purity of the gases, particularly when low-concentration mixtures are being prepared of species that exist at similar concentrations in the matrix gas. An example of this phenomenon is that argon is present in "pure" nitrogen at the level of at least 500 mmol/mol, which therefore prohibits the preparation of standards of argon in nitrogen close to or below this level. Other possible sources of uncertainty include buoyancy effects caused by changes in atmospheric pressure during the weighing process, and the expansion of the cylinder, which are generally controlled or eliminated by the correct use of tare cylinders [12]. Finally, there is uncertainty in the values of the relative molecular masses of the pure components at the level of several parts per million, which makes no significant contribution to the combined uncertainty.

Having recognised that standard gas mixtures can be prepared gravimetrically with uncertainties of less than 0.05% (relative to value), the question arises as to why the analysis of stable gases is problematic? The reason for this is that the instrumental methods available for analysing gases are highly sensitive to the species and concentration. They

Table II. – *Classification of gas species for which traceability is available according to their stability.*

Classification	Applications	Methods
Stable gases		
CO, CO_2, O_2, propane	Regulation of motor vehicle and industrial emissions	Stable gases can be stored for long periods (> 10 years in high-pressure cylinders). They are generally prepared gravimetrically to very high accuracies (typically $< 0.01\%$ relative to value).
Natural gas	Determination of the energy content of fuel	
Partially stable gases		
NO, NO_2, SO_2 volatile organic compounds (VOCs)	Monitoring the quality of ambient air	Unstable gases may be stored in cylinders for limited periods of time (typically from 3 to 5 years). They are not prepared gravimetrically, since they generally reach a stable value below that of the components added to them. They are usually certified against a "dynamic" method.
H_2O	Purity of raw materials for micro-electronic manufacture	
Unstable gases		
HCl, NH_3	Regulation of industrial emissions	Standards for unstable gases are usually not delivered in high-pressure cylinders. Dissemination of traceable measurements is usually carried out by the distribution of dynamic standards, such as permeation devices (for formaldehyde) or by use of highly stable instruments (for ozone).
Ozone	Monitoring the quality of ambient air	
Formaldehyde	Monitoring the quality of the indoor environment	

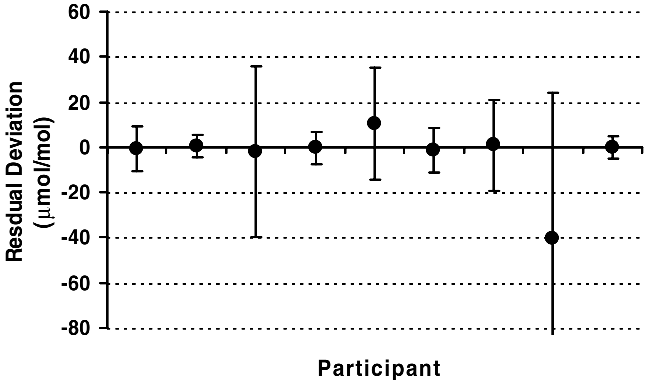


Fig. 4. – Results of a comparison of nine independent standards of carbon monoxide in nitrogen at 50 000 mmol/mol. The data are un-weighted residual deviations between the fitted value and the gravimetric value. The error bars represent expanded uncertainties ($k = 2$). Reproduced from [13].

are also sensitive to other gases that may be present in the mixture. Therefore, the number of possible calibration gases needed to eliminate all of these effects to the level of the gravimetric uncertainty is extremely large and goes beyond what might be practically provided.

9. – Preparative comparisons

One approach that overcomes the relatively poor accuracy of the analytical methods available is to exploit their good precision when measuring near-identical mixtures by carrying out a large number of repeat measurements of a suite of standards prepared independently in order to reduce the standard error on the estimate of the mean value of the set. This is now used as an operational model for CCQM key comparisons where it is known as a "preparative" comparison since it tests the capabilities of the participating laboratories to prepare standards. (The alternative operational model is an "analytical" comparison in which a single laboratory prepares a suite of standards that are analysed by each of the participants).

The results of a preparative comparison are shown in fig. 4. This was led by the NIST and the NPL and involved nine laboratories within the CCQM [13]. The results were calculated from the parameters of a regression relationship that minimised the random contributions from both the analysis and the gravimetric component [14]. The results showed that the standards studied were comparable with a standard deviation of the residuals of seven measurements (after the exclusion of two outliers) of 0.002% (relative to value), or an expanded uncertainty of 0.004% relative. This is a factor of between 2 and 30 less than the estimated uncertainty for the gravimetric preparation of each individual standard and serves to confirm the state of the art in this field.

10. – Complete mixture methods

Another approach to improving the accuracy with which gas mixtures can be analysed is to take advantage of what is called the analysis of a "complete mixture". The most prominent example of this approach is in the analysis of natural gas, which also happens to be an application where the economic value of the mixtures being measured is sufficiently high that the highest achievable accuracies are required. The principle of the complete mixture is that the values for the components expressed as amount fractions within a mixture x_i are subject to an additional constraint —that their sum should be equal to unity. This constraint is written as

$$\sum_{i=1}^{N} x_i^* = 1, \tag{6}$$

where x_i^* represent the corrected values for each component i. The constraint (6) leads to the calculation of adjusted values for each of the measured components (x_i)

$$x_i^* = x_i \Big/ \sum_{i=1}^{N} x_i. \tag{7}$$

The interesting property of these adjusted values is that they have reduced uncertainties. In the simplest case, where there is no correlation in the uncertainties between the x_i, the uncertainties of the adjusted values are given by [15].

$$\frac{u^2(x_i^*)}{(x_i^*)^2} = \frac{u^2(x_i)}{x_i^2}\left[1 - \frac{2x_i}{T}\right] + \frac{1}{T^2}\sum_{w=1}^{N} u^2(x_w), \tag{8}$$

where T is the sum of the unnormalised values. This process can readily produce a reduction in the uncertainty for methane from 0.05% (relative) to 0.002% (relative). Equation (8) applies to the case where there is no correlation in the uncertainties between the x_i. In more realistic cases, there is significant correlation and the reduction in uncertainty is even greater. This consequence of analysing "complete mixtures" may appear to give an unreal reduction in uncertainty until it is recognised that the introduction of the normal constraint (6) introduces additional information to the system which therefore decreases the uncertainty.

11. – Photochemical pollution

One of the recent successes in improving the environment we live in has been by the control of photochemical air pollution in cities. This success has been based on good measurements of the chemical species involved and accurate modelling of their reactions to develop effective strategies for their control. The species that are most important are the oxides of nitrogen (NO and NO_2), the Volatile Organic Compounds (VOCs) and ozone. The processes leading to photochemical pollution can be summarised in a highly simplified form by reference to the reversible chemical reaction

$$NO + O_3 \Leftrightarrow NO_2 + O_2. \tag{9}$$

The formation of photochemical ozone is initiated by oxidation of VOCs in the presence of the oxides of nitrogen and sunlight. The rate at which this oxidation occurs is different for each of the VOCs. In cities, NO emitted by motor vehicles reacts with ozone to form NO_2. Hence ozone levels are often lower in polluted cities than in the countryside. This short discussion illustrates that the species that must be measured in order to monitor the processes responsible for the build-up of photochemical pollution are: NO, NO_2, O_3 and the VOCs.

12. – Monitoring photochemical pollution

The reaction indicated by eq. (9) not only summarises the processes responsible for photochemical pollution but also plays a central role in the operation of the chemiluminescent NO_x analyser which is the most widely used method for measuring the oxides of nitrogen in ambient air. It works by measuring the ultraviolet light emitted when ozone reacts with NO to form NO_2. The same process is used to monitor NO_2 by use of a catalyst that reduces NO_2 to NO with a pre-determined conversion efficiency prior to the measurement. Consequently, it is not necessary to maintain standards of NO_2 since it can be measured with respect to standards of NO.

The importance of standards for NO has been the justification for a key comparison (CCQM-K26a) amongst laboratories of the CCQM. The amount fraction chosen for this comparison was 720 nmol/mol which is typical of the levels used to calibrate monitors used to make routine measurements of ambient air quality. The results are shown in fig 5. Since NO falls in the category of partially stable gases as defined in table II, it is not absolutely stable in high-pressure cylinders. Over the course of CCQM-K26a, the median rate of decay of NO within the cylinders was 0.3% (relative) over six months. In order to avoid this decay in the amount fraction of the standards from influencing the results of the comparisons, the rate of decay of each standard was carefully evaluated by the coordinating laboratory by reference to standards prepared by a series of gravimetric dilutions immediately prior to each measurement. This process introduced additional uncertainty to the results, which contributed to the typical uncertainty in the deviations from the reference values of between 0.5 and 1% (relative). This is a good example of what can be achieved at these very low levels, but it does not approach the accuracies described above that are applicable to stable gases at higher amount fractions.

13. – Volatile organic compounds (VOCs)

Another important class of components in ambient air that is monitored in order to control photochemical pollution is the VOCs. These include: straight chain aliphatics

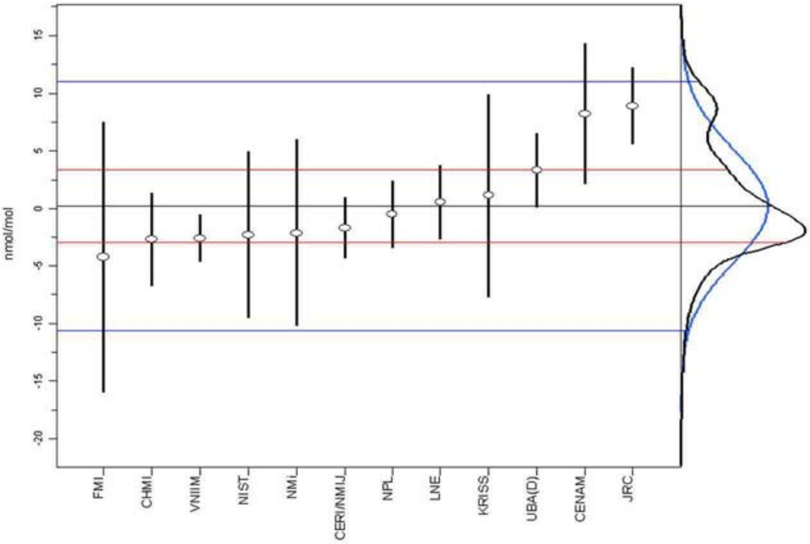


Fig. 5. – Results of CCQM-K26a in which participants measured the amount fraction in standards of NO in nitrogen at a nominal level of 720 nmol/mol. The distributions displayed at the right-hand end are a Gaussian (corresponding to the mean and the standard deviation) and the mixture-model.

(alkanes, alkenes, alkynes), aromatics as well as aldehydes and alcohols. The reactivity, and hence the lifetime of these species in the atmosphere varies significantly so it is necessary to measure the concentration of a suite of around 30 such species. They can be measured by gas chromatography using a cryogenic pre-concentration method at the levels that are found in the ambient atmosphere. These measurements require stable standards to act as calibrants since gas chromatography is dependent on external standards [16].

The amount fractions vary from around 2 nmol/mol down to 10 pmol/mol in an unpolluted atmosphere, and are between a factor of 5 and 10 larger in a polluted environment. Long-term measurements of the trend in the concentrations of VOCs shows decreases of between 2 and 10% per year. This reflects the success of measures introduced across the world to control direct and evaporative emissions.

Standards of these species are prepared by introducing the pure components into cylinder in liquid form with an uncertainty approaching that which can be achieved for gases. However, these species are not stable in either steel or aluminium cylinders because of their high potential for reacting with the walls of the container. The decay of the concentration is reduced by the use of a chemical coating that passivates the wall surface. Figure 6 indicates the effectiveness of this approach. Decays for components present above 1 nmol/mol are less than 5%, and only approach 10% for those at amount fractions much less than 0.5 nmol/mol. The error bars also show that the uncertainties associated with the analysis by GC are between 0.5 and 1%.

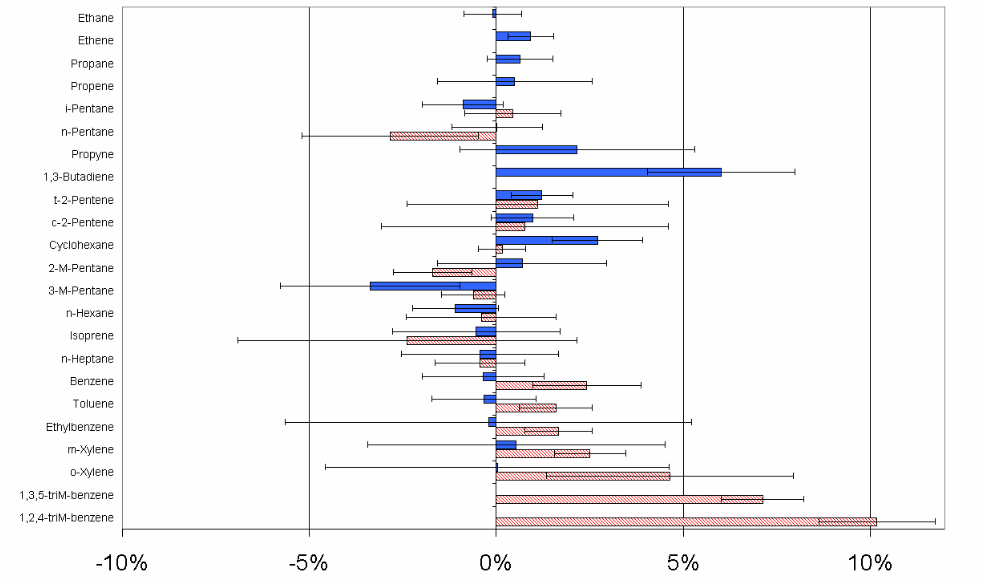


Fig. 6. – Comparison of 23 selected components within standards containing 30 VOCs at levels below 5 nmol/mol prepared in 2005 and 1998 and analysed in 2005. Measurements were made with two different gas chromatograph systems (indicated by the full and shaded bars). The differences are expressed relative to value. (Courtesy C. Plass-Duelmer, DWD, Germany.)

14. – Ozone in the atmosphere

The measurement of ozone presents a complex challenge, because it plays an important role in several different atmospheric chemical processes. Consequently, it is measured in different parts of the atmosphere with different methods. The challenge for metrology is to provide an infrastructure that provides access to traceability for each of these different methods.

The environmental issue that brought ozone strongly into the public awareness at the end of the twentieth century was the discovery of the Antarctic ozone hole. Although many chemists had hypothesised that the build-up of chlorine species in the stratosphere could lead to the depletion of the stratospheric ozone layer, it was not until the observation of the Antarctic ozone hole in 1985 that real evidence of ozone destruction chemistry was found. Within two years of its discovery, a series of international monitoring campaigns based on balloon, aircraft and satellite measurements showed that the reason for the ozone depletion proceeding faster than expected above the Antarctic was because of the very low temperatures found within high altitude clouds known as "polar stratospheric clouds". These act as sources of nuclei for heterogeneous ozone destruction reactions that proceeded more quickly than the expected homogeneous gas phase reactions. The agreement of the Montreal protocol in 1987 to limit the use and emission of chlorofluorocarbon species generated a new requirement for the measurement of these gases at the point of emission. Measurements of the long-term trend in these gases at sites representing the global background now confirms that they are declining in line with the reductions in emissions.

Whilst ozone concentrations in the stratosphere have been observed to decline, concentrations of ozone in the lower atmosphere have been observed to increase. This is largely due to the chemical cycles responsible for photochemical smog described above. Since ozone is known to be an irritant to much of the population, particularly to those suffering from asthma, this rise has led to sustained demands for accurate and stable measurements of ozone at ground level.

15. – Measuring ozone with the NIST SRP

The reference for all measurements of ozone made at ground level is the NIST Standard Reference Photometer (SRP). This instrument has been built according to a standardised design and is in operation at more than 30 locations around the world. It measures ozone by monitoring the absorption of light at 253.7 nm. It uses a pair of absorption cells and a pair of detectors. By alternately flushing the ozone-rich gas and "zero" gas through the two cells, it is possible to cancel effects due to the source and detectors to a high accuracy. The repeatability of the system is approximately 0.25% (relative). Recent work at the BIPM has identified sources of systematic bias in the system due to the heating of the cells by the lamp and the collimation of the radiation from the mercury lamp [17]. Comparisons between different SRP's consistently show agreement to within the 0.2% expected from the repeatability.

The NIST SRP has very successful achieved the original intention of delivering comparable measurements of ozone by constructing a series of instruments according to exactly the same design. However, there is now an increasingly strong requirement for the absolute values produced by the SRP to be comparable with those arising from other measurement methods such as from the NO/NO_2 gas phase reaction (described below) and those from other spectroscopic methods operating in other spectral regions.

16. – Measuring ozone by gas phase titration

It is possible to measure ozone by an entirely different method that does not make any use of spectroscopy. This makes use of the chemical reaction shown in eq. (9) and takes advantage of the availability of standards for NO with uncertainties less than 0.5% and highly linear chemilmuninescent instruments for the analysis of both NO and NO_2.

If the reaction shown in eq. (9) is initiated with a significant excess of NO over ozone, then the amount of NO required to destroy all of the ozone can be measured and will be equal to the amount of ozone when the reaction was initiated. Alternatively, the reaction can be carried out with an excess of ozone, in which case the change in ozone required to destroy all of the NO will be equal to the amount of NO. The use of this type of "gas phase" titration is an elegant approach, but any practical implementation

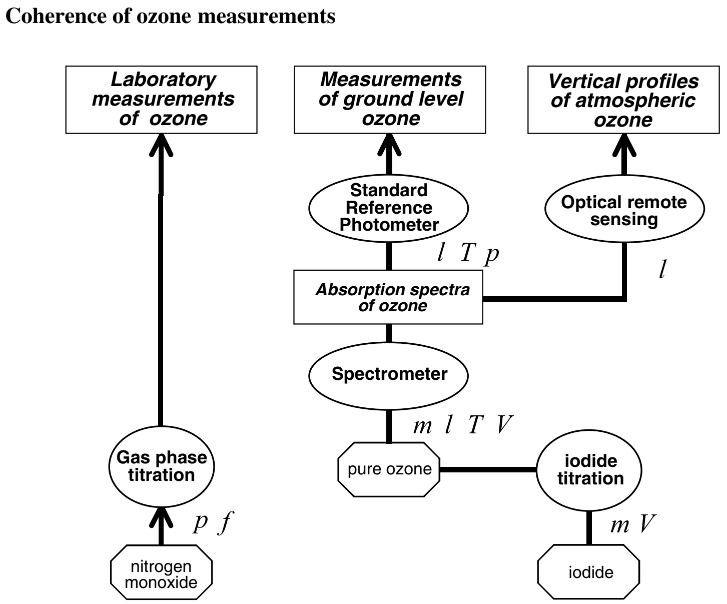

Coherence of ozone measurements
Laboratory measurements of ozone
Measurements of ground level ozone
Vertical profiles of atmospheric ozone
Standard Reference Photometer
Optical remote sensing
l T p
l
Absorption spectra of ozone
Spectrometer
m l T V
pure ozone
iodide titration
Gas phase titration
p f
m V
nitrogen monoxide
iodide

Fig. 7. – Schematic of routes used to provide traceability for measurements of ozone in the laboratory, at ground level and in the atmosphere. The letters correspond to requirements for traceable measurements of mass (m), length (l), temperature (T), pressure (p), volume (V), flow (f).

must take full account of the potential for other chemical reactions to occur at the same time. Examples include the depletion of NO_2 by reaction with ozone to produce NO_3 and then N_2O_5.

An experimental implementation of this method has been implemented at the BIPM and used to measure ozone simultaneously with measurements of the same gas stream with a NIST SRP. The results indicate that the conventionally agreed value for the absorption coefficient of ozone may be in error by up to 2%. This is the subject of further investigation.

17. – Coherence of ozone measurements

The preceding sections describe two methods used to measure ozone. Figure 7 is a simplified illustration of how these two methods depend on traceable measurements of mass (m), length (l), temperature (T), pressure (p), volume (V) as well as the availability of nitrogen dioxide, iodide and pure ozone. Figure 7 also shows the dependence of optical remote sensing methods on absorption coefficient data of the same type as used by the SRP. The complexity of fig. 7 demonstrates how the comparability of laboratory, ground-level and atmospheric profile measurements of ozone relies upon the coherence of the SI.

18. – Particles

This discussion about the gases responsible for the development of photochemical pollution has been simplified by the omission of any discussion of the role particles. They are of particular importance because they can be directly responsible for certain effects on health and they play a role in global climate change. Consequently, they are the subject of regulation that requires a basis for comparable and stable measurements.

Measurements of particles are concerned with two properties: the size, which must be defined in some way that captures information about shape correctly, and the chemical composition, which must take account of any inhomogeneity. There is no adequate metrological approach to solving either of these challenges, except in the case of the most straightforward distributions of particles.

19. – Summary. Why traceability to the SI?

At the beginning of this lecture, I introduced the concept of a primary method of measurement as being a measurement method that was capable of producing measurement results that were "traceable to the SI". However, I did not give any reason why one might want to produce measurement results that are "traceable to the SI".

The SI is a coherent system of measurement units. The reasons for using a measurement system such as the SI as the reference for measurements are the three criteria given above:

- The results of the same measurements against the same references made in a different laboratory should be the same ("comparability").
- The same measurement made against the same references should remain the same over time ("stability").
- The results of the same measurements against different references, usually on the basis of different measurement methods, will be the same ("coherence").

These three properties of comparability, stability and coherence are the fundamental reasons why scientists go to the trouble of using a measurement system such as the SI. If we look at these properties in a little more detail, we can see that the first two of them —stability and comparability— can be provided by mechanisms that are altogether less complex (and expensive) than providing an entire system of units such as the SI. For example, comparability between a group of laboratories is readily achieved by the exchange and maintenance of an artefact standard. When the only objective is to produce comparable results, there is no requirement for the artefact to be labelled with any externally recognised value, it is only necessary for it to retain its value as it moves between laboratories and for it to be used validly in each case. If the artefact is also stable, then it should be possible to use it at a later time to provide a stable reference. This type of approach has no value beyond the group of laboratories who have access to the standards, but is a very efficient means of ensuring the comparability and stability of measurement data.

The third property, coherence, is different, in that measurement results can only be coherent if they are carried out by reference to a system of units that is itself coherent. For example, the measurement of the pressure (P), volume (V) and temperature (T) of a sample of gas leads to a value for the amount

$$n_{\mathrm{GL}} = \frac{PV}{RT}\,, \tag{10}$$

where R is the gas constant. The measurement of the mass of the same sample of gas together with a value for its relative molecular mass (M) also leads to a value for the amount of gas

$$n_{\mathrm{RMM}} = \frac{m}{M}\,. \tag{11}$$

As a result of the coherence of the SI, we expect that these two values will be the same ($n_{\mathrm{GL}} = n_{\mathrm{RMM}}$). This can only be achieved if the base units to which measurements of pressure, volume and temperature are traceable are themselves coherent. Similarly, it is expected that the coherence of the SI system of units enables measurements of the heating effect of optical radiation (measured in optical units) to be equivalent to the heating effect of electricity (measured in electrical units). It is the coherence of the unit system that is the major justification for its complexity and the most compelling reason for its use. Therefore, any justification articulated for the use of traceability to the SI must be principally based on the importance of the coherence of measurement results. The example discussed here of ozone measurements provides a good example of the need for coherent measurement results. Measurements of ozone are carried out by different communities of scientists using quite different measurement principles; and there is an expectation that, since they are all measuring the same species, their results should be on the same basis. Similar examples could be given for the need for a coherent basis for the measurement of particulate matter, which is not discussed at length here. The argument for coherence can also be made for any other species, such as mercury, that is measured in entirely different chemical forms as it interacts with the ecosystem in complex cycles.

In conclusion, our understanding of many of the environmental issues that are of interest to researchers and policy makers depends on the availability of stable and comparable measurement results. Additionally, since it is essential to bring together measurement data of different species from different methods, it is also essential that they are produced by reference to a measurement system that is coherent. All of these can be achieved through a focus on developing a hierarchical measurement system that provides traceability to the SI.

REFERENCES

[1] Cvitas T., *Metrologia*, **33** (1996) 35.
[2] McGlashan M. L., *Metrologia*, **31** (1994) 447.

[3] Fujii K. *et al.*, *IEEE Trans. Inst. Meas.*, **54** (2005) 854.

[4] *International Vocabulary of Basic and General Terms in Metrology* (the VIM) 2nd edition (International Organization for Standardization, Geneva) 1993.

[5] Milton M. J. T. and Quinn T. J., *Metrologia*, **38** (2001) 289.

[6] *Quantities, Units and Symbols in Physical Chemistry* (the IUPAC Green Book) 2nd edition (Blackwell Science) 1993.

[7] Milton M. J. T. and Weilgosz R. I., *Metrologia*, **37** (2000) 199.

[8] Milton M. J. T. and Wang J., *Int. J. Mass Spectrom.*, **218** (2002) 63.

[9] Milton M. J. T. and Wang J., *Rapid Commun. Mass Spectrom.*, **17** (2003) 2621.

[10] Watters R. L., Eberhardt K. R. *et al.*, *Metrologia*, **34** (1997) 87.

[11] Buck R. P. *et al.*, *Pure Appl. Chem.*, **74** (2002) 2169.

[12] Milton M. J. T., Woods P. T. and Holland P. E., *Metrologia*, **39** (2002) 97.

[13] Milton M. J. T., Guenther F. *et al.*, *Metrologia*, **43** (2006) L7.

[14] Milton M. J. T., Harris P. M., Smith I. M., Brown A. S. and Goody B. A., *Metrologia*, **43** (2006) S291.

[15] Brown A. S. *et al.*, *J. Chromatogr. A*, **1040** (2004) 215.

[16] Slemr J. *et al.*, *J. Geophys. Res.*, **107** (2002) .

[17] Viallon J. *et al.*, *Metrologia*, **43** (2006) 441.

Developments in optical radiometry

N. P. Fox

National Physical Laboratory, Quality of Life Division - Hampton Rd, Teddington, Middlesex TW11 0LW, UK

1. – Introduction

This paper reviews some of the developments in optical radiometry made over the last 25 years that have led to a ten-fold improvement in the accuracy for many optical radiation measurements. Throughout this paper optical radiometry is defined as the measurement of electromagnetic radiation in the spectral region 200 nm to 30 μm, although many of the principles apply outside of this spectral region, a particular case being that of Vacuum UV radiation.

The paper also aims to show how the wide variety of technical disciplines that have an interest in optical radiation measurement, *i.e.* Radiometry, Photometry, Spectroradiometry, Pyrometry, Solar Physics, Environmental Science, Earth Observation, etc., can have measurements performed against a common physical scale based on the International System of units (SI) and also a common realisation of that scale with the cryogenic radiometer. Figure 1 is a schematic representation of the scope and interaction of the various quantities being discussed in this paper.

It should be noted that for brevity, reference will be made where possible to review articles which will serve as a source of further references rather than to give a complete bibliography within this paper. The author also recognises that the breadth of the subject matter being covered means that a significant number of examples and references will be missed and the reader should take no inference from their absence.

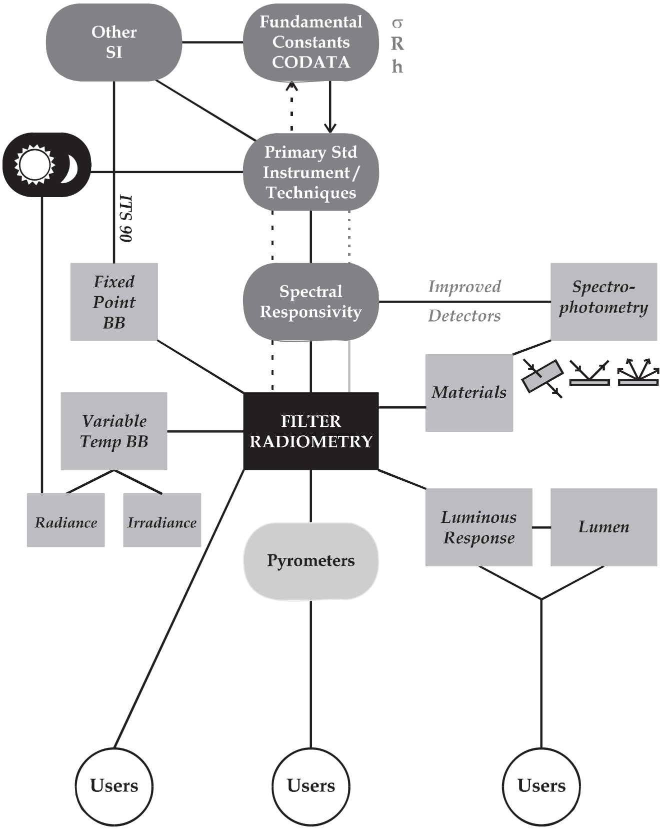


Fig. 1. – Schematic representation, showing the inter-linking and interdependence of primary radiometric quantities and the traceability route to the user.

2. – History

Optical radiometry can be considered to have started with the theoretical confirmation of fourth-power-of-temperature law of thermal emission by Boltzman in 1883 building on the empirical proposal of Stefan in 1879. This was rapidly followed by the first "absolute" radiometers of Ångström, 1893 [1], and Kurlbaum, 1894 [2]. Ångström's radiometer was designed to measure solar radiation for meteorological purposes, and Kurlbaum's radiometer was the first radiometer designed as an absolute standard for light measurement (the first attempt to replace a source as the primary standard).

Kurlbaum of the German standards laboratory, Physikalisch-Technische Reichsanstalt (PTR) in Berlin, now known as Physikalisch-Technische Bundesanstalt (PTB), developed his radiometer to replace the Heffner amyl acetate lamp which was the basis for the German unit of luminous intensity and optical radiation measurements at that time.

The operating principle of these absolute radiometers is the substitution of an equivalent and measurable amount of electrical Joule heating, for the optical radiation, which is to be measured. The optical radiation is absorbed on a black element and causes a temperature gradient across a heat link to a heat sink. The electrical power is adjusted to give a similar temperature rise so that the radiant flux absorbed by the radiometer can be calculated. This type of radiometer, which has been refined over the last century, is commonly known as an electrical substitution radiometer (ESR).

The most significant event this century for optical radiometry was probably the adoption by the 16th General Conference of Weights and Measures in 1979, of a new definition for the SI base unit for luminous intensity, the candela [3]. The new definition "The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hertz and that has a radiant intensity in that direction of (1/683) watt per steradian" effectively linked photometric units to those of optical radiometry. This allowed the SI unit to be realised by an absolute detector rather than an absolute source, as had previously been the case. It gave a new stimulus to improve and link optical radiometric based units and has invigorated the work of many national standards laboratories.

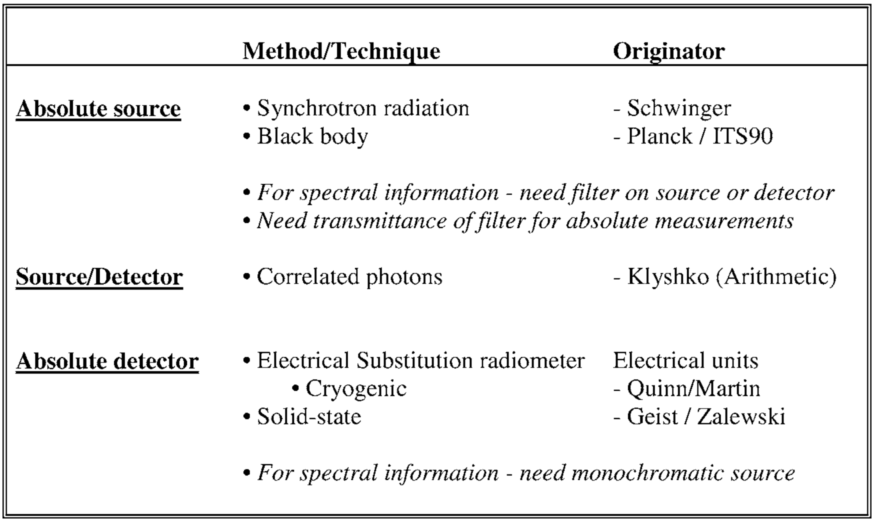
It is not completely coincidental that the new definition was formulated at a time when a renewed technological interest had arisen for the measurement of optical radiation from an increasing number of industrial applications facilitated by the availability of new devices, and techniques.

It is also perhaps interesting to note that this new definition effectively separated the formal interdependence of thermal metrology and radiometry. These two disciplines have always had and always will have, close links because of the black-body laws. This is particularly apparent at relatively high thermodynamic temperatures ($>$ few thousand K) when the peak of the spectral emission curve of a black-body radiator is tending towards visible wavelengths. For this reason prior to the above redefinition the SI candela had been defined with respect to predicted radiation from a black body operating at the temperature of the freezing point of platinum. However, as will become apparent this link will in practise be rapidly replaced and may in the longer term lead to even closer links between these two SI quantities.

3. – Base radiometric units or methods

In optical radiometry there are a number of primary quantities each with its own chain of secondary quantities, transfer standards and methodologies which have developed to serve a specific metrological sector, *e.g.*, spectral irradiance, photometry, spectral responsivity, etc. However, the philosophy of the last 20 y or so has been to try to link all of these to one common base unit, the Watt, which surprisingly to those not involved in this field is not the SI base unit, which is of course the candela.

In fig. 1 the base radiometric unit or method could in principle be based on any of a number of different methodologies (see fig. 2). A brief description of the underlying principles and reference to a recent review of each is given below.



	Method/Technique	Originator
Absolute source	• Synchrotron radiation	- Schwinger
	• Black body	- Planck / ITS90
	• *For spectral information - need filter on source or detector*	
	• *Need transmittance of filter for absolute measurements*	
Source/Detector	• Correlated photons	- Klyshko (Arithmetic)
Absolute detector	• Electrical Substitution radiometer	Electrical units
	• Cryogenic	- Quinn/Martin
	• Solid-state	- Geist / Zalewski
	• *For spectral information - need monochromatic source*	

Fig. 2. – Options for a primary standard to establish a base unit for optical radiometry.

3˙1. *Sources.* – Black bodies and synchrotrons are two types of source that can be considered as absolute or primary; the term primary meaning that the equation of state can be written down explicitly so that the quantity of emitted radiant flux can be calculated using a physical law, requiring only the measurement of parameters independent of the radiant output. Primary sources do not, therefore, require calibration against any other radiometric standard. Detailed review of both types of sources is given [4, 5] and only a brief summary will be given here. These sources normally require some form of spectral filtering before they can be used as primary standards and this can often limit the ultimate accuracy that can be achieved.

3˙1.1. Black bodies. The radiant flux from a black body at a known thermodynamic temperature can be calculated from Planck's equation. Therefore, black bodies can be used to realise absolute or primary radiometric scales for total radiant flux, or with characterised spectral filtering, spectral radiant flux. However, the requirement to know accurately the thermodynamic temperature tends to limit the application of these as primary radiometric standards to temperatures below about 1000 K (due to the availability of independent thermometry with sufficient accuracy). This in practise means wavelengths longer than around 3 μm. However, they are finding wide use as transfer standards, with their absolute radiance or irradiance being determined radiometrically [6-8]. This approach is gaining in popularity with the development of high-quality, high-temperature black bodies from Russia [9], which is resulting in uncertainties around the 0.1% level. Of course with the development of the metal carbon Eutectics [10] and the prospect of assigning a thermodynamic temperature to them, brings again the prospect of "absolute sources" for shorter wavelengths, even into the UV region.

3˙1.2. Synchrotrons. Synchrotron radiation is emitted by accelerating electrons to near relativistic speeds. The radiance of a machine built to accelerate electrons in a controlled way can be calculated from the operating parameters of the machine (*e.g.*, magnetic field, electron energy, electron current) using the Schwinger equation [11]. Such machines tend to be of the storage ring variety since this allows greater control of the parameters, examples being BESSY I now replaced by BESSY II in Germany and the recently upgraded SURF III of NIST. This type of machine can realise a primary scale for spectral radiance with an accuracy of around 0.1% [12]. However, its use as an absolute source is tending to be confined to the shortest wavelengths, in particular those in the vacuum UV and shorter. Instead it is more commonly being used as an intense relatively stable source of continuum radiation, which at wavelengths shorter than 400 nm is not readily available from other sources. After spectrally filtering it is then used with an absolute detector to determine the radiant power, from which other measurements and calibrations can be derived.

3˙2. *Techniques and methods*. – Before considering the most common approach for establishing the base unit, absolute detectors, and the hierarchy of quantities linked to them it is necessary to mention a relatively new concept, which is discussed in more detail in this book by Rastello [13] and elsewhere [5, 14]. In terms of its applications to radiometry these follow from the original ideas of Klyshko [15]. The technique relies upon the fact that a photon incident into a non-linear material will create pairs of photons. The presence and properties of one, being predicted by those of the other. In this way it is possible to consider a whole series of measurement systems that allow all the normal radiometric quantities to be derived or established, in many cases directly. The method can in this way be considered to act like an absolute detector or source and thus in reality is perhaps best called a technique. Its attraction comes from the fact that it relies simply on the counting of events and is in principle independent of other artefacts or standards. In this way it can be considered that dissemination of a traditional radiometric quantity becomes the teaching of a technique rather than the provision of a transfer standard. The present uncertainty estimates for the technique are at best at the few tenths of a percent level. They are dominated by the requirement and available techniques to characterise the efficiency of the system in particular the non-linear crystals spectro-photometric properties. However, even at this level of uncertainty the technique has its attractions and advantages. For example, it gives the ability, through the use of parametric down-conversion, to transfer the measurement regime from a more difficult spectral region, *e.g.*, IR to the more convenient visible.

Recent work at NPL on the use of laser-based spectrophotometry indicate that it may be possible to reduce the uncertainties of this technique down to the 0.01% level or lower when it becomes comparable to other techniques [16]. The remaining task of scaling the operating regime from the photon counting level to that of more normal levels can then be easily achieved using well-characterised attenuators.

The prospect of a primary standard radiometric technique operating through the principle of photon counting leads to the inevitable conclusion that a new definition of the candela may be timely. Since the current definition has already effectively become a measure of monochromatic spectral radiance weighted by a constant (1/683) to represent the efficacy of the human eye, it is simple to convert this to a measure of photon number. Of course there should be good reasons to do this, one of which should be that there is a practical way of realising the definition in this way. Another might be that this provides an opportunity to bring the photon into the SI in a more direct and transparent way, with the added advantage of making the candela more quantum based [16].

It should also be noted that many of the challenges utilising optical radiation (photons) relate to situations where the flux levels and photon number is small, for example in bio medical applications and where they are carrying "information", *e.g.*, communications, Quantum Information Processing (QIP), etc. In these situations often the development of standards and techniques which are designed to operate at these low flux levels has many advantages over developing a hierarchical power chain.

3˙3. *Detectors.* – Arguably the most significant advances in optical radiometry over the last 100 y have come about through developments and applications of absolute detectors. This is particularly true of the last couple of decades. This paper therefore will put greater emphasis on these developments than those of sources. There are currently two techniques of significance; one based on thermal detectors and the other on photon detectors. However, before considering these in any detail it is worth noting that in contrast to sources, to establish spectral radiometric quantities these detectors need to obtain their spectral information from a source. The characteristics of this source, *i.e.* its spectral purity, intensity and stability, ultimately limiting the performance of the detector. It was thus timely that laser radiometry was developed at NIST in the 1970s as the renewed interest in radiometry came about with the redefinition of the candela.

3˙3.1. Laser radiometry. The use of lasers for high-accuracy radiometry originated from work at the then National Bureau of Standards (NBS), now NIST (National Institute of Standards and Technology) [17]. Lasers have many obvious advantages over filtered continuum radiation as a source for establishing spectral responsivity scales and measuring the properties of detectors. They are truly monochromatic, have a well-defined and measurable if not known wavelength, highly collimated making them easy to align in optical systems, and have a high radiant output. However, they are not spatially uniform, Gaussian, can be highly polarised and have variations in their output power of typically a few percent. It is the latter, which was solved by Geist *et al.* [17].

A relatively simple technique was developed for stabilising the intensity of laser radiation based on the use of a pockels cell, see fig. 3. Polarised laser radiation is passed through a birefringent crystal, *e.g.*, KDP. On application of an electric field to the crystal, the electric vector of the polarised laser beam passing through the crystal is rotated to a degree dependent on the strength of the electric field. At the exit of the birefringent crystal, there is a polariser with its transmitting axis parallel to that of the laser beam. If no electric field is applied to the crystal the polarisation state of the incoming beam is the same as the exiting beam and is parallel to that of the polariser and thus there

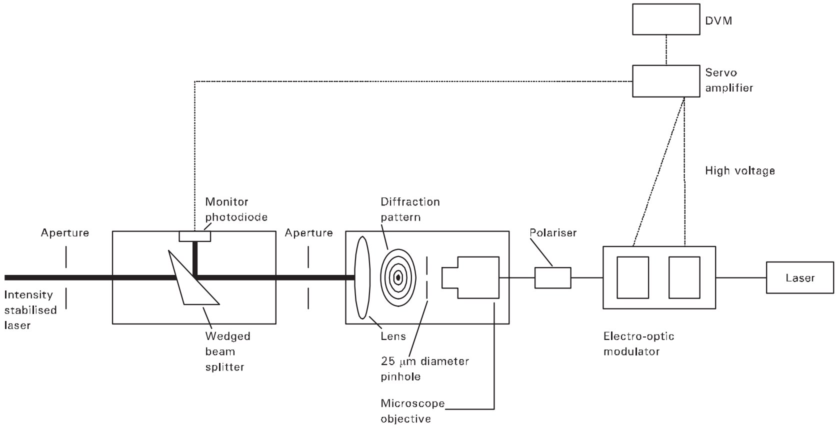


Fig. 3. – Schematic representation of an intensity stabilised laser facility.

are no significant losses. However, if an electric field is applied, then the beam exiting the crystal will have a polarisation state rotated with respect to its original axis and will then be attenuated as it passes through the polariser. Thus by varying the electric field to the crystal, the intensity of the transmitted laser beam can also be varied. This principle is exploited in the laser stabilisation technique. A beam splitter positioned in the laser beam after the crystal as shown in fig. 3 reflects a small fraction of radiation to a silicon photodiode. The photodiode can detect any small changes in intensity of that radiation and through a simple electronic servo system it varies the electric field to the birefringent crystal to try to maintain a constant intensity.

The spatial filter assembly placed between the above two elements of the laser stabilisation system is to provide an optically clean beam and to expand its diameter to around 6 mm. The latter helps to reduce the effect of spatial non-uniformity. The laser stabilisation system described above can maintain an intensity stability of better than ±0.002% over many hours, see fig. 4.

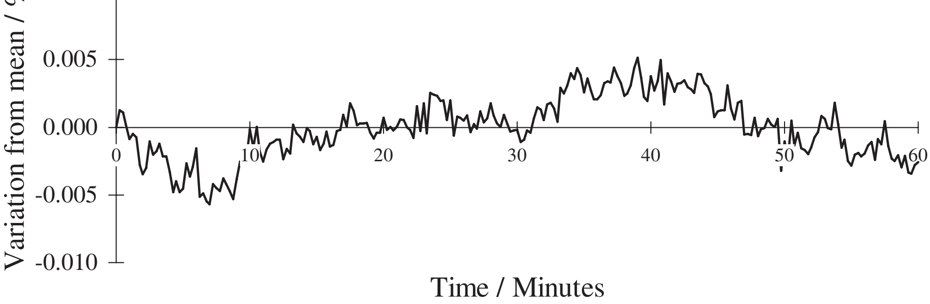


Fig. 4. – Plot showing the variation in intensity of a stabilised laser as a function of time.

The availability of a wide range of lasers, both single line ion lasers and tuneable dye and solid-state lasers, makes it possible to select radiation of any wavelength for any type of optical radiation measurement. The few remaining problems associated with polarisation and spatial non-uniformity are also, to some degree, common with other sources of radiation. However their effect can be minimised through careful selection of spatially uniform, polarisation insensitive detectors, performing measurements on the same optical axis and in some cases the use of novel techniques [18].

3˙3.2. Thermal-detector–based scales. These use the electrical substitution principle described above and vary only in the type of temperature sensing element used to detect the temperature rise, *e.g.*, pyroelectric, thermopile or bolometer. These realisations can be further subdivided as operating at either ambient or cryogenic temperatures. Cryogenic radiometers will be discussed in detail in sect. **4**.

3˙3.3. Photon-detector–based scales. These use solid-state physics modelling techniques to predict the response of photodiodes, commonly called the self-calibration technique. In this technique, a number of measurements can be performed to calculate the quantum efficiency of the device, *i.e.* the number of charge carriers reaching the sensing electrical circuit per incident photon. This technique is best described by the originators of the technique Geist and Zalewski [19, 20]. However, in summary the method relies on two simple concepts and measurements to address them:

- Some of the charge carriers resulting from photons absorbed near the rear of the depletion region of a photodiode are not collected because they are absorbed in traps. This loss can be reduced and thus increasing the collection efficiency (to near unity) by extending the depletion region further toward the rear of the device. This is achieved by applying a "reverse bias" to the photodiode structure.

- Charge carriers absorbed near the front of the photodiode are similarly absorbed in traps. However, in this region the process is unfortunately enhanced through the presence of a small electrical field resulting from $+ve$ charge trapped at the interface of the silicon photodiode and its anti-reflection/passivation coating. By temporarily storing $-ve$ charge on top of this passivation layer the effect of this field can be cancelled and consequently increase the collection efficiency to unity. In this case the $-ve$ charge is applied through the use of a water drop as an optically transparent, removable, conductive electrode.

It soon became apparent that the technique had many advantages over room temperature ESRs, not least because of its simplicity and low cost. However, the technique was initially limited to the relatively narrow, although important, spectral region between around 400 and 900 nm. Its accuracy around 0.1% was comparable with the best ESRs of the time [21]. The technique lost favour as a primary standard in the latter half of the 1980s as cryogenic radiometers were developed which could achieve uncertainties of $< 0.01\%$ [22] whilst uncertainties for the self-calibration technique was limited to at best about 0.05% [23, 24]. However, Geist continued to carry out further work utilising software models developed for solar cell development and demonstrated that subject to more detailed characterisation of key materials input parameters, the technique could be improved significantly, possibly to a level approaching 1 part in 10^8 [25]. This is obviously a significant challenge to the community and if achieved leaves open the possibility of new concepts and approaches. Recent work by Gran in Norway has shown that electrically conductive coatings can be applied, instead of the water drop, to cancel losses due to charge at the surface, reducing some of the difficulties associated with the technique [26]. Although this work identified other issues which were limiting, it showed that the principle was viable and work continues to develop this concept further.

The models have been further developed to extend the application into the UV spectral region [27, 28] and have allowed spectral responsivity scales to be established with uncertainties approaching 0.5% through extrapolation of visible measurements.

In the authors opinion without the concept of "self-calibration" many of the advances and techniques that are today used routinely in optical radiometry, *e.g.*, high-accuracy filter radiometry, would not have developed to their current level.

4. – Cryogenic radiometry

4˙1. *Background.* – The operating principle of a cryogenic radiometer is the same as that of any ESR but with the advantage that operation at low temperatures reduces all of the parameters limiting the accuracy of room temperature ESRs to a very low level. How this is achieved for each parameter will be described by reference to the schematic representation of an ESR shown in fig. 5. Each parameter limits the accuracy by which the equivalence of optical and electrical heating can be determined [29]. A recent review of the subject and its applications can be found in [30].

- *Reflectance of the absorbing coating.* Cryogenic operation makes it possible to construct a very large absorbing copper cavity whilst maintaining a high sensitivity because of the increased thermal diffusivity of copper at low temperatures.

- *Lead heating.* Electrical connecting leads can be made from superconducting materials to avoid Joule heating of the leads and the resultant corrections and uncertainties due to these often difficult-to-quantify losses in room temperature instruments.

- *Heat flow.* Thermal resistance results in different temperature gradients for optical and electrical heating and hence different heat flow to the surrounding environment. Operation at low temperatures reduces radiative heat loss to a negligible level and operation in a vacuum removes convective heat exchange with the surrounding environment. This ensures one and the same heat flow path for both electrical and optical heating.

- *Background radiation (P_{back}).* Operation at low temperatures with suitable cold shields greatly reduces and stabilises the level of background thermal radiation.

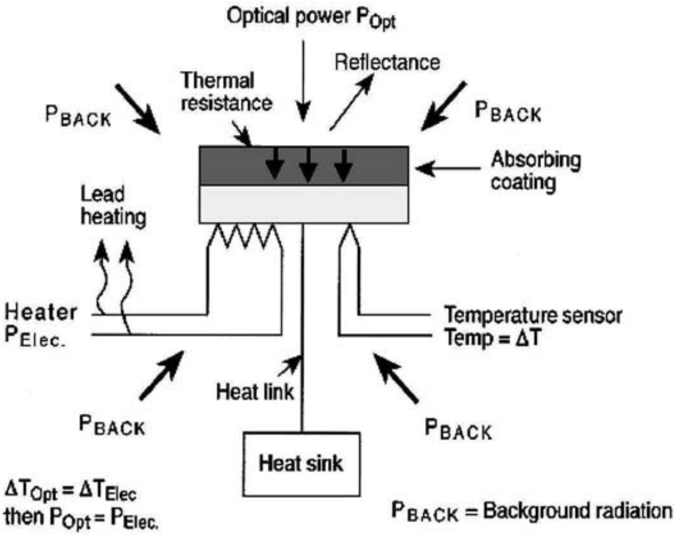

Optical power P_{Opt}
Reflectance
Thermal resistance
P_{BACK}
P_{BACK}
Absorbing coating
Lead heating
Heater P_{Elec.}
Temperature sensor Temp = ΔT
P_{BACK}
Heat link
P_{BACK}
Heat sink
ΔT_{Opt} = ΔT_{Elec} then P_{Opt} = P_{Elec.}
P_{BACK} = Background radiation

Fig. 5. – Schematic representation of the electrical substitution principle. $P_{\rm E}$ is adjusted so that $\Delta T_{\rm E} = \Delta T_{\rm O}$. $P_{\rm B}$ = background radiation input; $P_{\rm E}$ = electrical power input; $P_{\rm O}$ = optical power input; $\Delta T = T(P_{\rm E+B})$ or $P_{\rm (O+B)} - T(P_{\rm B})$.

The first successful cryogenic radiometer was that of Quinn and Martin [29] which followed earlier work by Ginnings and Reilly [31]. The Quinn-Martin radiometer or QM radiometer, was not designed for work on optical radiometric scales but for the determination of the Stefan-Boltzman constant (SB) and thermodynamic temperature by total radiation thermometry in the range from −40 to +100 °C. The instrument consisted of a black-body receiver operating at around 2 K, the cryogenic radiometer, and a variable temperature blackbody radiator. The instrument measured the total radiation emitted from the black body through a precisely known solid angle defined by two apertures. The SB constant could then be determined from eq. (1):

$$M(T_{\rm tp}) = g\sigma {T_{\rm tp}}^4 \tag{1}$$

where $M(T_{\rm tp})$ is the total radiant exitance of the black body at a temperature $T_{\rm tp}$, g is a geometric factor defining the solid angle of collection, σ is the Stefan-Boltzman constant and $T_{\rm tp}$ is the thermodynamic temperature, of the triple point of water, 273.16 K.

Quinn and Martin calculated the uncertainty attributable to their experimental determination and compared their value for σ with the theoretical calculation given by fundamental constants in Codata [32].

$$\sigma = 2\pi^5 k^4 T^4 / 15 h^3 c^2. \tag{2}$$

The experiment was similar to one carried out earlier by Blevin using an ambient temperature ESR and a black body at the freezing point of gold [33], but with a cryogenic radiometer the uncertainties are an order of magnitude smaller.

The experimental value measured for σ was

$$5.66959 + 0.00076 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$$

compared with that from Codata of

$$5.67051 + 0.00019 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}.$$

The agreement of these results within their combined uncertainties confirms that the systematic and experimental uncertainties determined for the QM radiometer are reasonable and that the instrument is truly an absolute radiometer to at least 2 parts in 10^4.

Thermodynamic temperatures, T, were determined by measuring the ratio of the total radiant exitance $M(T)$ to that of the black body at the temperature of the triple point of water. From the relation $M(T)/M(T_{\mathrm{tp}}) = T^4/(T_{\mathrm{tp}})^4$, T can be calculated. The scale could then be disseminated by means of platinum resistance thermometers which were in thermal equilibrium with the blackbody at the temperature T [29].

The application of cryogenic radiometry to optical radiometry was first suggested by Geist in combination with Blevin and Quinn [34], and led to an experiment performed by Zalewski and Martin to compare the self-calibration technique with the QM radiometer. The experiment required the modification of the QM radiometer to allow laser radiation, rather than the black-body radiation to be measured. The results of the experiment were not published but demonstrated the feasibility of using a cryogenic radiometer as the basis for optical radiometric scales [35].

This experiment led NPL to design a new cryogenic radiometer optimised for laser radiation, the so-called primary standard (PS) radiometer [22]. Similar instruments built to the NPL design by Oxford Instruments Ltd, UK, are in current use at NIST and PTB.

Recognising the limitations of size and the need for liquid-helium NPL designed a new smaller radiometer utilising a mechanical cooling engine to cool the radiometer to cryogenic temperatures [36]. A schematic representation of this radiometer is shown in fig. 6.

The operation of this radiometer is similar to that of the PS radiometer and the QM radiometer except that in this case the radiation enters in the horizontal plane rather than vertical, as with the others. Since the mechanical cooler only cools to around 15 K it also needs to use high-temperature superconductors for electrical connectors rather than conventional Niobium wire. The overall uncertainty of this instrument is presented in table I.

4ˈ2. *Applications of cryogenic radiometers.* – In the last 20 years there have been many designs of cryogenic radiometer from different groups to meet an increasing range of applications. The highest-accuracy applications of a cryogenic radiometer require

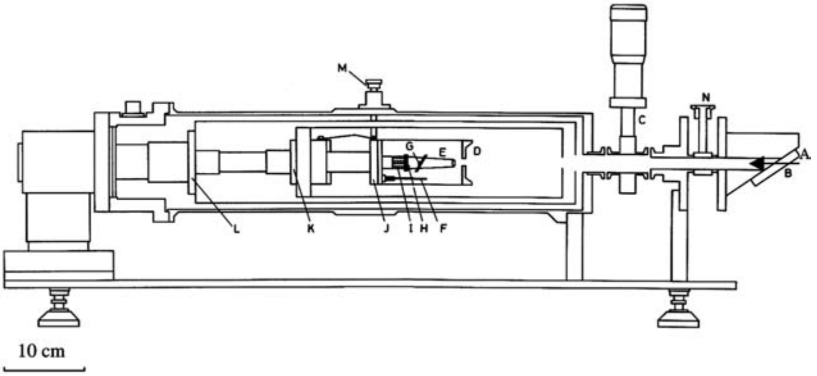


Fig. 6. – Schematic drawing of the mechanically cooled cryogenic radiometer. A: laser beam, B: Brewster angled window, C: gate valve, D: Quadrant detector, E: cavity, F: high-temperature super-conducting heater leads, G: heater, H: RhFe thermometer, I: reference heat link, J: reference temperature heat sink, K: second-stage cooler, L: first-stage cold head, M: thermal shorting mechanism, N: vacuum port for window chamber.

the radiometer to be used directly, measuring either total radiation or monochromatic radiation. For total radiation measurements, the source must be within the same vacuum chamber and at temperatures low enough to prevent significant outgassing. As cryogenic systems are very effective cold traps for emitted contaminants, their application in normal laboratory situations is limited.

The following list of examples of the use of cryogenic radiometers is not necessarily comprehensive but gives an overview of the diversity of applications:

- The current main application for cryogenic radiometers is as a basis for optical radiometric scales in national standards laboratories. Designs include the NPL PS radiometer [22], radiometers from Cambridge Research and Instrumentation [37], Oxford Instruments Ltd [36] and Finland [38].

Table I. – *Corrections and uncertainties of the power of the laser beam expressed as parts in* 10^4 *of the measured power.*

	Correction	Uncertainty
Window transmittance	3.0	0.3
Beam scatter	2.0	0.15
Absorptance of cavity	0.2	0.05
Electrical power measurement		0.05
Sensitivity of radiometer		0.1
Changes in thermal and scattered radiation		0.1
Sum in quadrature.		0.37

- Cryogenic radiometers have been used to measure total radiation from blackbodies as a means of direct calibration [39] and for thermodynamic temperature scale realisations [29]. There is also a design for a space-based instrument capable of measuring the solar constant, or Total Solar Irradiance, TSI, with an uncertainty of more than 10 times the current best achievable [40] and this has been further extended into its use for Earth Observation in general see subsect. **10**˙2 [41].
- The wide dynamic range and spectral non-selectivity of cryogenic radiometers has been used to calibrate space instrumentation such as CERES [42], for Earth radiation budget experiments.
- Cryogenic radiometers have been used to measure spectrally filtered synchrotron radiation from electron storage rings as a basis for calibrating detectors in the deep UV spectral region [43].

The most common use of cryogenic radiometers is to measure monochromatic radiation to establish spectral responsivity scales; the radiation coming from either an intensity stabilised laser [22] or an incandescent lamp dispersed by a monochromator [44, 45]. In each case, the ultimate goal is to determine the response of transfer standard detectors. These detectors can then be used as the basis not only for spectral responsivity measurements (sect. **5**) but also for polychromatic scales (sect. **6**).

4˙3. *Confirmation of accuracy.* – Uncertainties of $< 0.01\%$ are now routinely reported for many radiometric quantities traced to a cryogenic radiometer. However, it should of course be remembered that a cryogenic radiometer on its own is simply a well-characterised instrument and as such, unless linked to a more fundamental concept, has the potential for unknown systematic errors to exist or to develop and that these could then propagate into all other radiometric quantities.

The check mechanism is of course similar for all metrological quantities, intercomparison. For optical radiometric quantities, as is the case for all other SI quantities, such comparisons are organised (at the highest level) by the appropriate Consultative Committee (CC) of the Comité International des Poids et Mesures (CIPM), (in this case CCPR, Consultative Committee for Photometry and Radiometry).

In terms of cryogenic radiometers, a CCPR comparison has recently been completed using trap detectors as a transfer standard. The comparison showed that providing the radiometers were operated correctly, the results from all cryogenic radiometers, irrespective of the mode of operation or manufacture agreed within their combined uncertainties [46]. See fig. 7. It should be noted, however, that in carrying out such a comparison a number of users found that their operational procedures were not adequate, and in some cases requiring fairly large uncertainties to be added for the actual comparison process, demonstrating the importance of such an activity and how easy it can be to introduce significant errors to a measurement even though the basic standard has such a high intrinsic accuracy.

In addition to the above there have also been a series of direct comparisons of cryogenic radiometers performed as bi-laterals. In these, one radiometer was transported to the

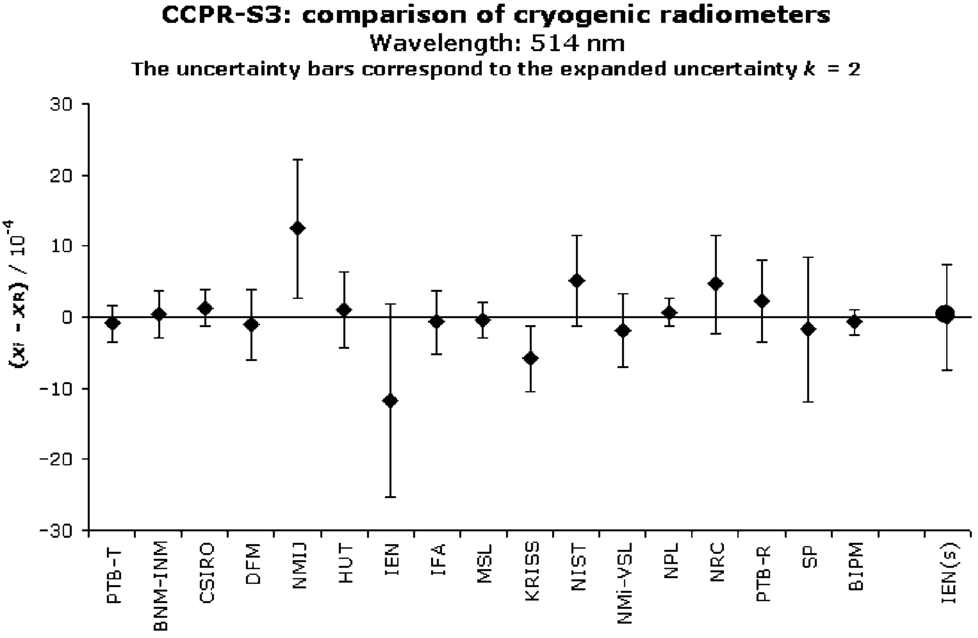


Fig. 7. – Results of the CCPR supplementary comparison of cryogenic radiometers CCPR-S3 at 514 nm taken from the BIPM MRA database [47]. The results are presented as the relative difference of the participant to a weighted mean. The value for IEN(S) refers to a supplementary bilateral comparison performed after the original comparison but following the same protocol.

location of a second and both instruments were then used to view the same source under the same conditions. This resulted in higher accuracy due to the removal of many of the uncertainties associated with the transfer standard, which no longer applied under these circumstances [48-51].

Such comparisons whilst obviously important are time consuming and difficult to organise and only check equivalence of measurement of a particular quantity. In order to be confident of the uncertainty relative to the SI quantity it is essential that a cryogenic radiometer or a quantity it measures can be compared to an independent quantity of similar or lower uncertainty. For the PM radiometer described earlier this was carried out through a comparison to the QM radiometer, effectively linking optical radiometric quantities and in particular the candela to SB constant [52]. However, the design of both these instruments would only allow this experiment to be carried out to an uncertainty of around 2 parts in 10^4 at best. Therefore a new instrument Absolute Radiation Detector (ARD) [53] was built to do this with a target uncertainty of around 0.001%.

The ARD instrument seen schematically in fig. 8 has been designed to measure total radiation from a black body or by substitution of a Brewster-angled window, monochromatic radiation. In measuring total radiation a link to SB can be established confirming the performance of the instrument. Thus when ARD is used to measure monochro-

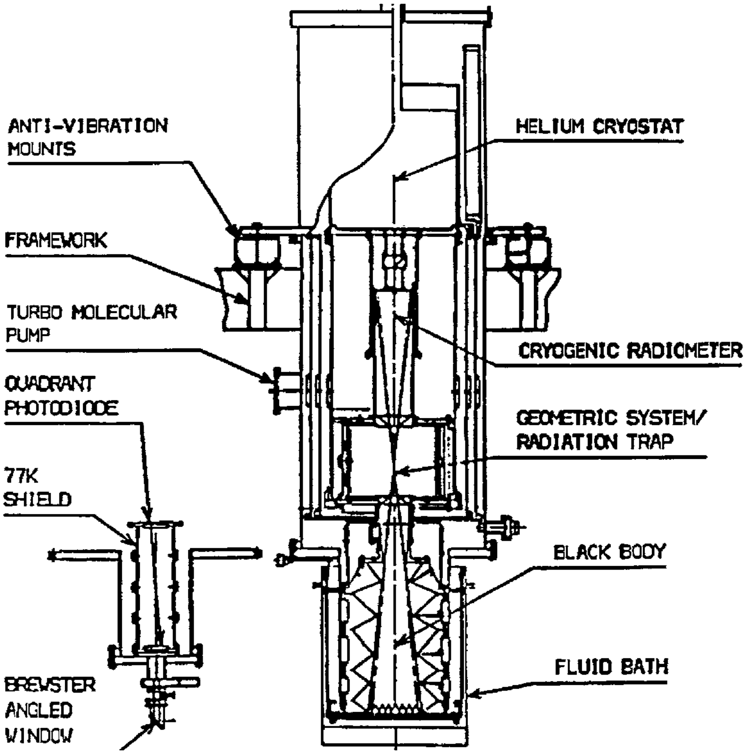


Fig. 8. – Schematic diagram of the Absolute Radiation Detector (ARD).

matic radiation, confidence in this latter measurement can be inferred. This then allows the establishment of a spectral responsivity scale through the calibration of secondary detectors to this monochromatic beam of known power.

The design of this instrument has included many of the developments and ideas of the last decade to maximise its performance including the use of the "Christmas tree" black body [54]. This concept first proposed by Quinn involves the creation of the effect of a large aspect ratio black body whilst only using a relatively simple black plate as the emitting source. The design principle underpinning this approach can be seen in fig. 9. Here a highly reflective hemisphere surrounds the black target. Any detector through geometric considerations can only view radiation originating from the black target. Since the internal walls of the hemisphere are highly reflective, via the reciprocity law they have a very low emissivity. Thus radiation emitted by them and reflected from the black target makes an insignificant contribution to that emitted by the target due to its own temperature. Similarly, radiation emitted by the black target and incident on the walls will be reflected back onto the black target with little change to its spectral properties, *i.e.* it will still be Plankian and representative of the emissivity of the black targets surface temperature. This means that the walls of the black body, whilst contributing through

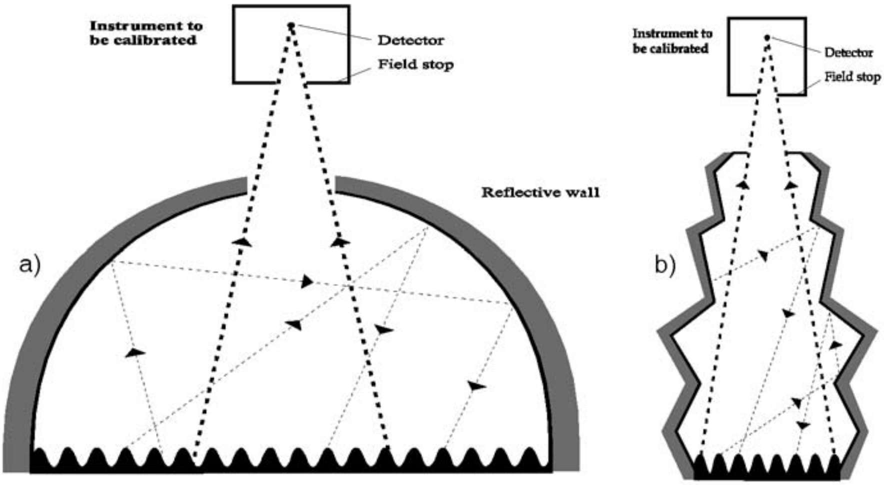


Fig. 9. – a) Reflective wall black body. b) "Christmas tree" black body.

geometric considerations to improving the emissivity of the black body as a whole, can vary in temperature to a significant level without affecting the overall performance. It can be shown that the effect of the reflecting surfaces increases the allowed temperature difference (to cause a specific effect in emissivity of the cavity) by a factor of 1/2 emissivity of the black coating. The application of this concept allows the construction of a black body, which is much less dependent on the needs of any temperature control system to provide uniform isothermal conditions for the whole of the black body.

Although the ideal hemisphere black body described above will give the performance required, its major limitation in practise is that its physical size is still extremely large. This is because it is necessary to move the walls of the hemisphere a significant distance from the black emitting surface in order to get an appreciable gain in emissivity. However, by careful design, a novel shape can be produced which simulates that of a hemisphere but in a more useful and compact size. An example of such a design can be seen in fig. 9b where any ray of light incident on the side-wall undergoes a maximum of two reflections before being incident back on the emitting black surface of the black body. In this way a relatively small black emitting surface can be used whilst gaining all the advantage of a large reflective hemisphere. In a practical realisation of this example at near ambient temperatures using Martin Marietta Infra-black (reflectance say 0.02) and gold reflecting walls (emissivity 0.01), a temperature difference of 1 K can be allowed before changing the total emissivity by as much as 1 part in 10^5. For the same criterion without the gold coating a temperature difference of 50 times smaller would only be allowed, namely 20 mK, clearly much more difficult to achieve.

The dominant source of uncertainty in this experiment is the measurement of the radiative transfer function between the black-body source and the detector. In this context the term includes not only the defining geometry but also the effects of diffraction and scatter in the intervening path. Again a fairly novel approach in the ARD design was to build an absorbing cavity, which is large enough to fully collect all diffracted radiation from the top-defining aperture. However, since to do this would make it so massive that the time constant would be impractical, use was again made of a mirror. In this case a mirror forms the top part of the cavity in terms of its absorption properties but remains mechanically detached and thus not contributing to the mass. Since its task is simply to ensure that all the diffraction is collected, any small losses due to imperfect reflectance are negligible. It is also worth noting that this experiment can also be considered as linking the so-called primary detector and detector-based radiometric quantities to the arguably more fundamental absolute source in terms of the Planckian radiator. This again returns the link between the SI base units, the candela and Kelvin discussed briefly in sect. **2**, although now the link is at a more fundamental level since it is being carried out at the triple point of water and not that of the freezing point of platinum.

However, as in many fundamental metrology projects the realisation of a concept is often extremely hard and can sometimes lead to disappointing and unexpected results.

In the case of ARD, it gave results for the Stefan Boltzman constant significantly different from those anticipated. Following significant analysis and evaluation the error was traced to the model used to design the radiation trap. This required a redesign to one approx twice the size (in terms of diameter). Similarly, the mechanical/engineering realisation of the "Christmas tree" black body was also flawed and so at the time of writing a new experiment is being performed following a reconstruction of the instrument.

4˙4. *Future developments.* – It is likely that future developments in cryogenic radiometry are most likely to concentrate on increasing the range of applications, tailoring designs for specific tasks. NPL for example is currently building a new radiometer with the aim of measuring radiation dispersed by a monochromator at power levels of around $1\,\mu$W. In doing this to a sufficiently low uncertainty requires that the instrument has a NEP of around 1 pW. This in turn puts significant demand on the stability and level of the background radiation. The effect of this can be reduced by using "cold filters", *i.e.* windows which only transmit in the spectral region of interest, and thus can absorb the majority of the radiation from ambient temperature background (IR) and are cryogenically cooled so that their own emission is only characteristic of their thermodynamic temperature.

5. – Basis for optical radiation scales

5˙1. *Spectral responsivity scales.* – The increased use of spectral responsivity scales as the basis for all optical radiation measurements has led to a rapid improvement in accuracy. Traditionally such scales have been based on the combination of two independent scales [55, 56], a primary scale at a single wavelength from a primary radiometer and a relative scale based on the use of a spectrally non-selective thermal detector. The latter requires only knowledge of the variation in reflectance of the detector as a function of wavelength. It is of course possible for the same detector to be used for both scales, *e.g.*, the use of a cryogenic radiometer directly with a monochromator [44, 45] but often this is not possible because the detector holding the relative scale needs to have a very high sensitivity so that it can be used to measure radiation from a monochromator across the complete spectral region of interest. In contrast, the primary reference point can make use of much higher intensity radiation at specific wavelengths.

To establish, maintain and disseminate a spectral responsivity scale requires the calibration of transfer standard detectors at sufficient wavelengths to make interpolation of the spectral response possible at any intermediate wavelength. This may only require one or two wavelengths in the case of a very spectrally flat thermal detector but more often requires measurements at 10 or 20 nm intervals to characterise solid-state photodiodes across the visible part of the spectrum. Therefore the calibration of solid-state detectors usually requires the use of a monochromator as the source of radiation. Obviously tuneable laser sources can replace the monochromator although this is generally a costly and unnecessary process. Similarly the use of Fourier transform instruments are also starting to be investigated for the dissemination of spectral responsivity scales at least when comparing similar detectors [57].

There are two ways to use a cryogenic radiometer for the realisation of spectral responsivity scales, and the relative benefits of each depend very much on the final application. The starting point for both cases is of course identical, *i.e.* the cryogenic radiometer, which is used to measure the power in a beam of monochromatic radiation. A transfer standard detector is then placed in this beam and its response measured, the process being repeated for each wavelength. It is assumed that the power in the beam of radiation is suitably stable and that all the radiation is measured by both detectors and corrections are applied for losses due to scatter, or transmittance of windows, etc.

They differ in how finer spectral detail is determined. In one case a monochromator can be used directly with the cryogenic radiometer, although this generally requires that the reference detectors are in the vacuum enclosure. In the second, tuneable lasers can be used to fully characterise the detector, this is generally in-practical due to cost. The more usual alternative is to use a secondary thermal detector, operating at ambient temperatures, which is spectrally flat and has a high sensitivity. These can take many forms but NPL is currently making use of pyroelectric detectors with gold black coatings. This type of detector is very fast, relatively large and through the properties of the gold black, spectrally flat. Any residual spectral variation in the gold black coating (largely affecting the IR spectral region) is removed through the use of a reflecting hemispherical reflector in a similar way to that described earlier in subsect. **4**˙3.

This spectrally flat detector can simply calibrate the relative spectral responsivity of the unknown detector with the absolute level being set using the cryogenic radiometer. The uncertainties achievable using this technique are such that it leads to a disseminated scale with uncertainties comparable with any in the world today, [55, 56], see fig. 10.

In disseminating spectral responsivity to users the greatest difficulty is the choice of transfer standard. Unlike many metrological fields, there is a wide choice of transfer standards available, each serving a different spectral region. They all have advantages and limitations for particular applications and operational environments. These limitations are particularly true in the spectral regions outside of the visible.

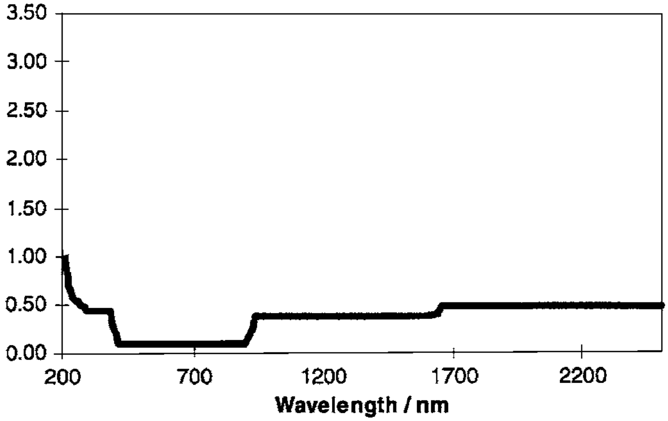


Fig. 10. – Combined uncertainty of the of the NPL spectral responsivity scale in the 0.2 μm to 2.5 μm range.

5˙2. *Transfer standard detectors.* – There are two types of detectors that can be used as transfer standards:

- Thermal devices—These are spectrally non-selective devices which sense the heating effect of optical radiation and convert it to a measurable signal, *e.g.*, a thermopile or pyroelectric detector.
- Photon detectors—These are detectors that normally generate an electrical signal proportional to the number of photons falling on the device, *e.g.*, photomultiplier tubes, photo-emissive detectors and solid-state photodiodes. Their response is usually very wavelength dependent and usually limited to a relatively narrow spectral region. Solid-state photodiodes are most commonly used as transfer standards because of their wider spectral coverage and better overall performance characteristics. Therefore this paper will only consider solid-state photodiode transfer standards.

Solid-state photodiodes absorb photons within a semiconductor, exciting electrons from the valence band to the conduction band and, ideally, detecting an electron for every absorbed incident photon. The ideal transfer standard detector should have the following properties: high sensitivity, wide dynamic range, spatially uniform response, large sensitive area, fast response, robustness, stability with time, and have a spectrally flat, or at least, a smoothly varying, spectral response over the spectral range of interest.

For each of these parameters, solid-state detectors are available with the required performance, with the exception; they have a variable spectral response. Unfortunately, this is the most important parameter for interpolation between calibration points from a cryogenic radiometer.

This limitation has been overcome in the visible and near-infrared spectral region by using photodiodes, which have a spectral response that can be modelled, Geist *et al.* [24].

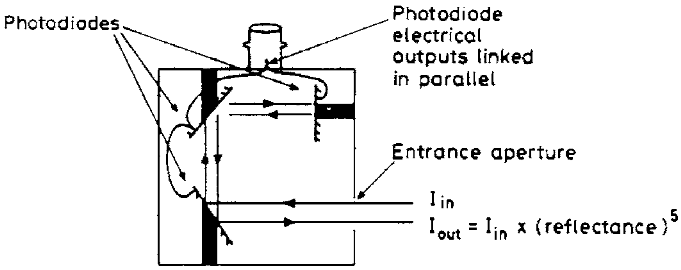

Photodiodes
Photodiode electrical outputs linked in parallel
Entrance aperture
I_{in}
$I_{out} = I_{in} \times (\text{reflectance})^5$

Fig. 11. – Schematic representation of the trap detector.

These models accurately predict the responsivity of a silicon photodiode from a few simple measurements through knowledge of the physical structure of the photodiode.

The use of these models is simplified if the diodes are mounted in a light-trapping arrangement first proposed by Zalewski and Duda [58] in which the incident beam undergoes multiple reflections at a succession of photodiodes. These are now available in many forms, including the commercial QED 100 and 200 series [59], Hamamatsu trap [60], tunnel trap [61], and transmission trap [62]. Each arrangement has the same goal, to reduce the net reflectance of the detector to a small level and hence reduce the spectral variations in the quantum efficiency. Trap devices have then been shown to have a quantum efficiency that is near unity and that varies very little with wavelength; for example, for a trap constructed from Hamamatsu S1337 photodiodes the quantum efficiency variation over the region 550 to 920 nm is $< 0.1\%$, and the device can be assumed "quantum flat" [60]. A schematic representation of such a device is shown in fig. 11.

The predictable spectral dependence of responsivity of trap detectors make them well suited for use as transfer standards and the advantages of such detectors was demonstrated in a CCPR comparison where the light trapping arrangement gave significantly less scatter than the traditional single element devices [63].

The choice of transfer standard detector outside of the visible spectral region is not as easy. In the near infrared, from 1000 to 1640 nm InGaAs, traps [64] have shown promise but manufacturing difficulties have delayed their availability. However, single-element InGaAs photodiodes when selected for spatial uniformity, perform adequately. The best choice of photodiodes for the ultraviolet spectral region is more uncertain. Silicon is a possibility, although structure in its spectral responsivity curve caused by multiple ionisation effects makes it more difficult to model and interpolate however this is now being achieved to some level [27, 28]. But unfortunately silicon photodiodes have also been seen to have some stability problems particularly when exposed to UV radiation [65]. Other candidates are based on wide-band gap materials such as GaP, GaAsP and SiC; a description of some of the properties of these devices can be found elsewhere [66]. A more recent development and potentially the best candidate is that of PtSi [67]. These detectors appear to be highly stable under intense UV radiation are fairly large area

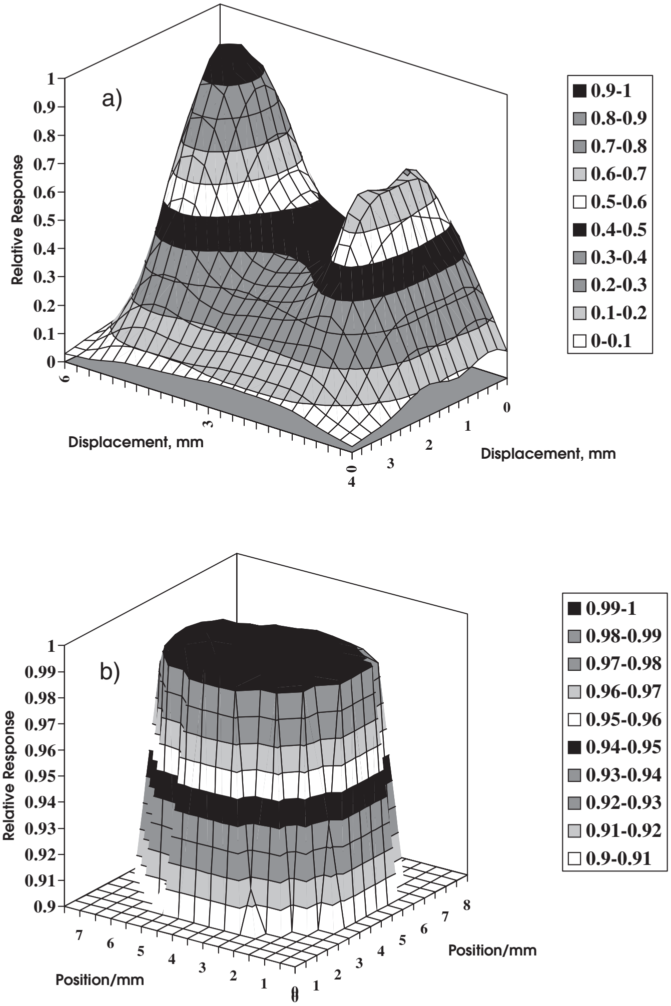


Fig. 12. – Relative spatial uniformity of responsivity of HgCdTe detector at 10.6 μm, a) standard detector, b) same detector as a) with addition of sphere.

and can be selected to be spatially uniform. However, their responsivity is not as high as some other devices and so is not as useful when very low power levels are available.

At present there are no solid-state detectors ideally suited for use as transfer standards in the mid to far infrared spectral region. Detectors available are still of relatively poor quality and have particularly poor spatial uniformity. This problem can be minimised through the incorporation of another reflecting device in front of the detector [68]. In this case a complete sphere is used which has an internal coating of roughened gold so as to ensure that the detector viewing the walls of the sphere always sees the same uniformity of radiation, having resulted from multiple internal reflections. The improvement in uniformity can be seen in figs. 12a and b.

6. – Thermal radiometry

6˙1. *Primary quantities.* – The different methodologies traditionally used in thermometry and optical radiometry are becoming more blurred as many of the techniques being used by the two communities are becoming similar in nature. A more detailed treatise of the thermal aspects highlighted in this section can be found elsewhere in this book [69]. In thermometry, the kelvin is currently defined at a single point, the triple point of water and a series of practical reference points approximating thermodynamic temperature. The embodiment of this and the approved method of interpolating and extrapolating thermal measurements is known as ITS 90. Similarly the candela is based on a definition at a single spectral radiometric point and is then disseminated in a practical sense by means of a defined weighting function approximating the response of the average human eye as defined by CIE ($V\lambda$) [70].

However, in both cases the use of black bodies (or quasi black bodies in the form of incandescent lamps) and Plancks law are widely used to establish or disseminate various associated quantities. Clearly, in terms of thermal measurements below 1000 K the transfer standards most commonly used take the form of contact thermometers but these are linked to ITS 90 through the maintained fixed reference points. For photometric and colorimetric measurements the effective colour temperature of the reference source is always defined. In terms of spectral radiometric measurements, high-temperature black bodies (> 2800 K) are used with measured thermodynamic temperatures to ease interpolation of spectral data.

Higher temperatures, using spectral radiometric measurements linked to a cryogenic radiometer can also be carried out. NPL and other NMIs have already shown that measurements can be made of thermodynamic temperatures at temperatures greater than 933 K with uncertainties of less than 0.05% [71, 72], which is better, in terms of thermodynamic temperature, than ITS 90. These measurements make use of filter radiometers calibrated in terms of radiance or irradiance against a cryogenic radiometer. The basic experimental techniques have been described elsewhere [71] and so will not be repeated here. It should, of course, be noted that whilst to the radiometric community the term filter radiometer is used as in this paper, to the thermal community the same instrument would be called a radiation thermometer or pyrometer. In the case of the pyrometer the spectral response need only be known to estimate an effective wavelength, its absolute calibration originating from a reference black body; the radiometric filter radiometer, in contrast, would be calibrated in spectral radiometric units and would directly measure the absolute emittance of the source in terms of radiance or irradiance.

Perhaps the most exciting development in source radiometry and radiation thermometry in the last 20 years is the relatively recent development of metal carbon Eutectics pioneered by Yamada of NMIJ [10]. The latest review of this subject [73] describes all of this work in more detail and so will not be reproduced here. However, in summary Eutectics can be considered as providing high emissivity, planckian radiators at high temperatures up to around 3000 °C. As the eutectic point is a physically reproducible "transition", dependent only on the purity of its compound, it is effectively similar in

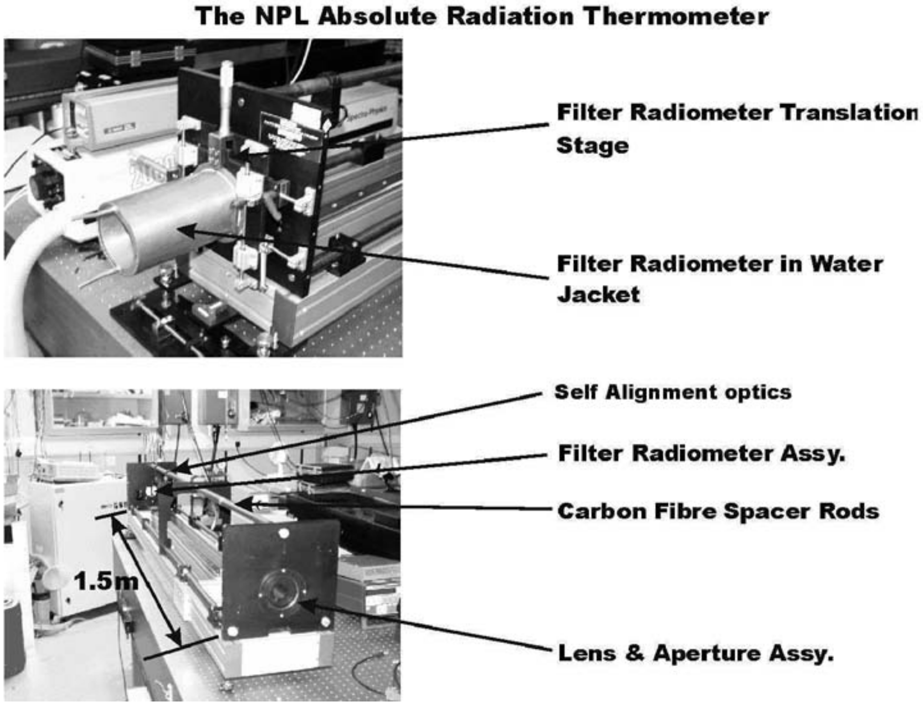

The NPL Absolute Radiation Thermometer
Filter Radiometer Translation Stage
Filter Radiometer in Water Jacket
Self Alignment optics
Filter Radiometer Assy.
Carbon Fibre Spacer Rods
1.5m
Lens & Aperture Assy.

Fig. 13. – Photograph showing the design of the NPL Absolute Radiation Thermometer (ART).

nature to that of a pure metal fixed point and thus in use, offers the prospect of a major improvement in uncertainty at high temperatures. Since the uncertainty in the International Temperature Scale (ITS -90) for temperatures above 1000 °C are relatively large and scale as T^2 referenced, as a radiometric ratio, to a fixed point. At present the fixed points are limited to the pure metals with the highest temperature being that of copper at 1086 °C. Providing the Eutectics temperatures can be determined by an independent means their use in the ITS will reduce uncertainties by around a factor of ten.

The absolute filter radiometric techniques described above offer this capability and so work is now in progress at a number of the worlds NMIs to optimise these techniques, intercompare and assign values to the Eutectics.

The Absolute Radiation Thermometer (ART) constructed by NPL for this activity, fig. 13, is expected to have an overall uncertainty of less than 0.05% in radiance and a precision of around 0.005% or lower [74]. The success of this work and its comparison with others will lead to detailed discussion about new practical temperature scales, not only about the incorporation of eutectic points but also whether it is necessary to maintain a series of fixed point black bodies as reference sources as at present. It should be possible to use absolute radiation thermometers calibrated directly against radiometric standards to derive and disseminate the temperature scale more simply. In this way more accurate, cost effective and convenient measurements could be made.

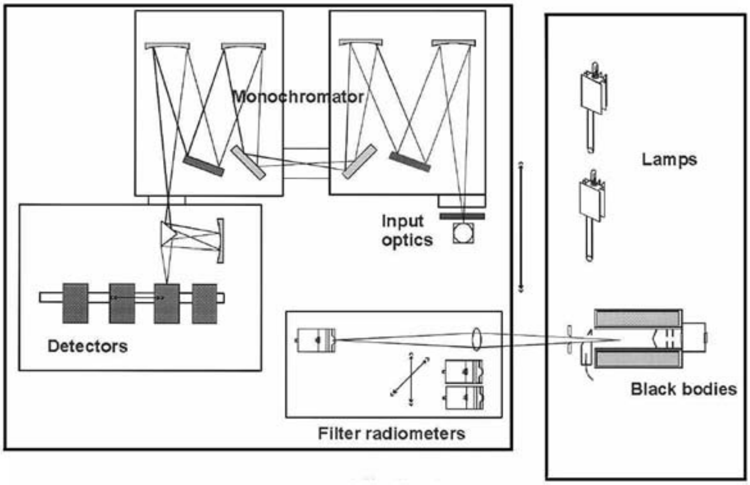


Fig. 14. – A schematic representation of the NPL primary spectral irradiance facility.

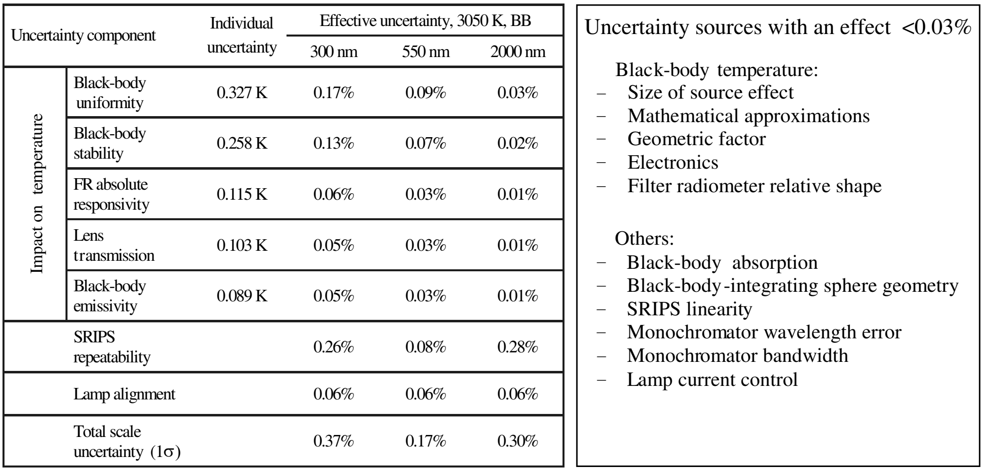
7. – Spectral irradiance and radiance

The primary quantities of irradiance and radiance can be established from primary sources such as black bodies and electron storage rings as described in sect. **2**, but to obtain spectral information requires the characterisation of spectral filters. In practise synchrotron radiation provides a good source for the UV and VUV spectral regions whilst high-temperature black bodies ($> 3000\,\mathrm{K}$) are generally being used for wavelengths longer than 250 nm. The primary source is generally used to provide a good relative spectral shape, which can be transferred to another transfer standard source, (usually a lamp of some form), via a monochromator used as a scanning spectral filter. The absolute level often comes from a single band filter radiometer, although, in the case of synchrotron radiation, the absolute level can be derived directly from the source parameters.

The single-band filter radiometers used to calibrate the absolute spectral irradiance or radiance from the black body are in practice similar to those discussed for radiation thermometry in subsect. **3**˙1, since their application is the same, the measurement of thermodynamic temperature of the black body. The design of the filter radiometers varies between research groups and can be based on interference filters or glasses with narrow or broad spectral bandwidths [75-78].

After determining its temperature, the black body is then viewed by a monochromator (spectroradiometer) to give a short-term calibration of its spectral response before it is in turn used to view a transfer standard lamp, in effect transferring the calibration from the black body. A schematic representation of the NPL Spectral Radiance and Irradiance Primary Scales (SRIPS) facility is shown in fig. 14. NPL currently uses a black body operating at temperatures up to 3500 K to establish its scales down to 200 nm [8].

TABLE II. – *Uncertainty budget for the NPL spectral irradiance scale.*



Uncertainty component		Individual uncertainty	Effective uncertainty, 3050 K, BB		
			300 nm	550 nm	2000 nm
Impact on temperature	Black-body uniformity	0.327 K	0.17%	0.09%	0.03%
	Black-body stability	0.258 K	0.13%	0.07%	0.02%
	FR absolute responsivity	0.115 K	0.06%	0.03%	0.01%
	Lens transmission	0.103 K	0.05%	0.03%	0.01%
	Black-body emissivity	0.089 K	0.05%	0.03%	0.01%
	SRIPS repeatability		0.26%	0.08%	0.28%
	Lamp alignment		0.06%	0.06%	0.06%
	Total scale uncertainty (1σ)		0.37%	0.17%	0.30%

Uncertainty sources with an effect <0.03%

Black-body temperature:
- Size of source effect
- Mathematical approximations
- Geometric factor
- Electronics
- Filter radiometer relative shape

Others:
- Black-body absorption
- Black-body-integrating sphere geometry
- SRIPS linearity
- Monochromator wavelength error
- Monochromator bandwidth
- Lamp current control

These new “detector-based” scales have much lower uncertainty than previous realisations, which will allow significant improvements in the accuracy of derived quantities such as radiance. An example of an uncertainty budget for such detector based scales illustrating the range of factors that need to be considered is given in table II.

These transfer standard lamps are then used to maintain and disseminate the spectral irradiance or spectral radiance to the user community. Although the above procedure appears relatively straightforward there are many sources of uncertainty, some of which can have significant impact in certain parts of the spectrum, *e.g.*, absorption features within the black body [79], etc. The objective of most NMIs is to be able to offer spectral irradiance measurements with uncertainties approaching 0.1% in the visible region of the spectrum.

As in all metrology the only real test of an SI quantity is a comparison. The last international comparison of spectral irradiance was piloted by NPL and published in 2005, where significant improvement in the equivalence between NMIs was demonstrated.

8. – Photometric units

Although the candela is the SI base unit, through its definition, it is really simply a measurement of spectral radiance at a specified monochromatic wavelength. It is therefore no surprise that the artefacts used to establish and disseminate it, photometers, are simply filter radiometers with a spectral filter designed to match that of the CIE definition of Vλ (spectral responsivity of average human eye) [70]. Luminous intensity and illuminance of a source can then be measured by using the photometer with a suitable defined viewing geometry. The transfer standards still most commonly used to disseminate these quantities are generally incandescent lamps, which is perhaps surprising given the relative performance and convenience of well-designed and characterised photometers [80].

Whilst the relative merits of using photometers or lamps as the primary medium for disseminating photometric quantities to the user can be debated, a more important practical point is worth making. Photometers of varying declared quality are widely used in industry but are rarely calibrated adequately. A calibration of a photometer usually takes the form of determining its response at a given distance to a standard lamp operating at a specific colour temperature. If the photometer is used to measure similar sources, then this can be considered a reasonable calibration. However, only too frequently, photometers are used with sources of different colour temperature or even quasi-monochromatic radiation like LEDs. In these cases, an accurate calibration of the spectral responsivity of the photometer is required to allow account to be taken of the imperfect match to Vλ. Few, if any photometers, other than those used to establish the primary quantity at NMIs, are ever calibrated using spectral responsivity measurements in any traceable way and are therefore likely to exhibit errors when used with sources differing in spectral content than that for which they were calibrated.

The lumen or total luminous flux is similarly determined through the use of a photometer, in this case either by scanning the photometer around the full 4π geometry using a gonio-photometer or using the absolute sphere method as developed by Ono [81] and later revised into the AC/DC technique [82].

9. – Transfer standards for source measurements

Transfer standards for source quantities are traditionally incandescent lamps. For photometry a wide range of tungsten lamps exist which are, in principle, fairly stable when used under well-controlled conditions. However, as with all lamps they are prone to shock and damage. Spectral irradiance is perhaps more worrying, since there are still no high colour-temperature ($> 3000\,\mathrm{K}$) lamps available that are reliably stable, or even necessarily indicate when they have changed. Tungsten Halogen lamps of the FEL type, which are routinely used, even after careful selection and ageing [83], have been found to suffer significant changes in their output on even gentle transportation at the rate of up to 1 in 3 lamps [84].

A more revolutionary approach is also being developed by a number of institutes, in which transfer standards are being developed specific to customers. For example a transfer standard filter radiometer or group of radiometers is provided to disseminate spectral radiance or irradiance at a number of fixed wavelengths to the users [85]. In this way the customer utilises these artefacts either to directly view their unknown source or to calibrate an intermediate source on site, which can then provide a reference for other instrumentation. This approach has the potential to significantly improve accuracy, reliability and reduce costs for the user.

A further development of this would be a fully spectrally tuneable transfer standard spectrometer perhaps based on the use of a diode array [86]. This latter type of instrument has, because of its full spectral resolution capabilities, the potential to act as a transfer standard for photometry, colorimetry, spectral radiance and irradiance quantities simultaneously. However, such devices whilst often stable and compact, suffer in terms of their stray light characteristics. The development of these devices limits the requirement for high-performance transfer standard lamps to only a niche set of specific applications.

10. – Climate change

10˙1. *Radiometric requirements.* – Although the full range of customers utilising optical radiation measurements continue to seek improved accuracy and better transfer standards, the community that probably has the greatest immediate need is that looking at climate change. In order to quantify the rate or confirm the existence of climate change and its anthropogenic origin, changes of physical parameters of a few percent per decade need to be measured. As an example it can be shown that this means that the reflected sunlight from the Earth's surface (spectral radiance) needs to be measured to 0.2%, the solar spectral irradiance to 0.1% and Total Solar Irradiance, TSI to 0.01% [87]. In practice this requires an improvement of nearly 2 orders of magnitude on current capability. These radiometric parameters ultimately relate to changes in Earth and Ocean temperature as well as surface UV radiation levels, which are of course the critical output.

Given the perceived difficulty of making measurements fully traceable to SI some researchers argue that an alternative approach exists. They argue that the establishment of well-controlled local site measurements of these parameters using highly stable instrumentation or a very stable reference is adequate to detect change. However, as history has shown many times, such a philosophy is liable to be doomed to failure and that unknown, slow drifts or time-variant systematic errors will be a likely source of error. In addition, there are always likely to be local geographical anomalies, particularly when making measurements over such a long time base, making it essential that truly global measurements are made at as many sites as possible. The need for this global averaging and inter-site comparison requires that each of these sites measures the same thing in the same way, and it is difficult to simply compare results between sites, even if regular intercomparisons are carried out, unless the instrumentation is fully traceable.

In more recent times the use of satellite data has helped to provide real global information, and for future missions this is becoming more focused, reliable and integrated. The cost of such missions and the time taken to develop and launch has led to a growing requirement to coordinate and combine data from different instruments to generate a desired data product. This requirement is also enhanced by the greater synergy that often occurs through the use of multiple data sets, which differ slightly in characteristics, *e.g.*, spectral band. These enhanced data sets can often produce different end data products, which often serve a larger user community than simply the sum of those associated with the individual instruments. However, since each individual mission has a different project team and may even be developed by a different space agency, it is essential that each sensor is calibrated in as accurate a way as possible and that the calibration is demonstrably traceable to SI units. This ensures that any measurements or conclusions drawn from the data can be substantiated and believed.

10˙2. *Total Solar Irradiance (TSI).* – In terms of climate change, a long time base of measurements is required to detect any effect. This time interval is likely to be beyond the active lifetime of any one instrument and so it will be essential to have well calibrated and consistently calibrated instruments to allow this record to be established. It can of course be argued that providing there is always an overlap between the operational lives of two instruments then this is adequate for climate change studies. The argument is probably true to some extent and indeed is the argument that is used in the solar physics community when monitoring the solar constant or Total Solar Irradiance, TSI. This community has been flying high accuracy (around 0.1% is typically claimed) ambient operating temperature ESRs in space for nearly 3 decades and have been monitoring changes in TSI as a function of time/solar cycle. Unfortunately, although 0.1% is often the claimed accuracy of any one instrument, when simultaneous measurements are made using different instruments, they often differ by as much as 0.7% [88] (fig. 15). Differences even exist between, nominally, the same design of instrument. In addition, such radiometers age with exposure to solar radiation.

It is perhaps notable that until a few years ago, it might have been argued that the biases between instruments was largely resolved, however the launch of TIM on SORCE simply emphasised the problem differing by nearly 0.4% from the others. For the purposes of monitoring change in TSI there has always been at least two active instruments in orbit. This has allowed researchers to normalise data and therefore monitor change resulting in a normalised TSI record [88] as shown in the lower curves of fig. 15. This is obviously fortunate, but may not always be the case in the future, *e.g.*, it is relatively easy for disasters like the SOHO incident to occur, when the satellite lost power and control and thus data for many days. Should this happen whilst no other ESR is operationally in-orbit the TSI record is irretrievably lost or at best has an uncertainty of $\sim 0.5\%$.

Does this matter? The Sun is the driving force of the planet and as can be seen from fig. 15 has a variable output over an 11 y timescale of around 0.1%, which similarly correlates with the sun spot cycle. However, whilst apparently regular at present there is clear evidence that in the past this has not always been so and in fact during the late 17th century the so called "little age" period the Northern hemisphere of the Earth was around 2 °C cooler than today and the river Thames in London froze regularly (which of course it does not do today) sufficient to hold fairs upon its surface. This cooler temperature is believed to be the result of the Sun having a reduced output of just 0.3%.

This example shows the potential for disaster, in terms of our understanding of the climate and the establishment of any mitigation policy. It also demonstrates the need for accuracy and demonstrable traceability to SI and also provides an argument to fly instruments of different design, which are less susceptible to change and have a higher inherent accuracy, *e.g.*, a cryogenic radiometer like CSAR [40, 41].

10˙3. *Earth viewing instruments.* – The situation becomes worse as the radiometric measurement becomes more complex. For example, in viewing the Earth to detect vegetation, pollution, etc., instruments are flown in space which must ideally take imaging pictures of the Earth to allow the surface to be characterised, using the Sun as an illumi-

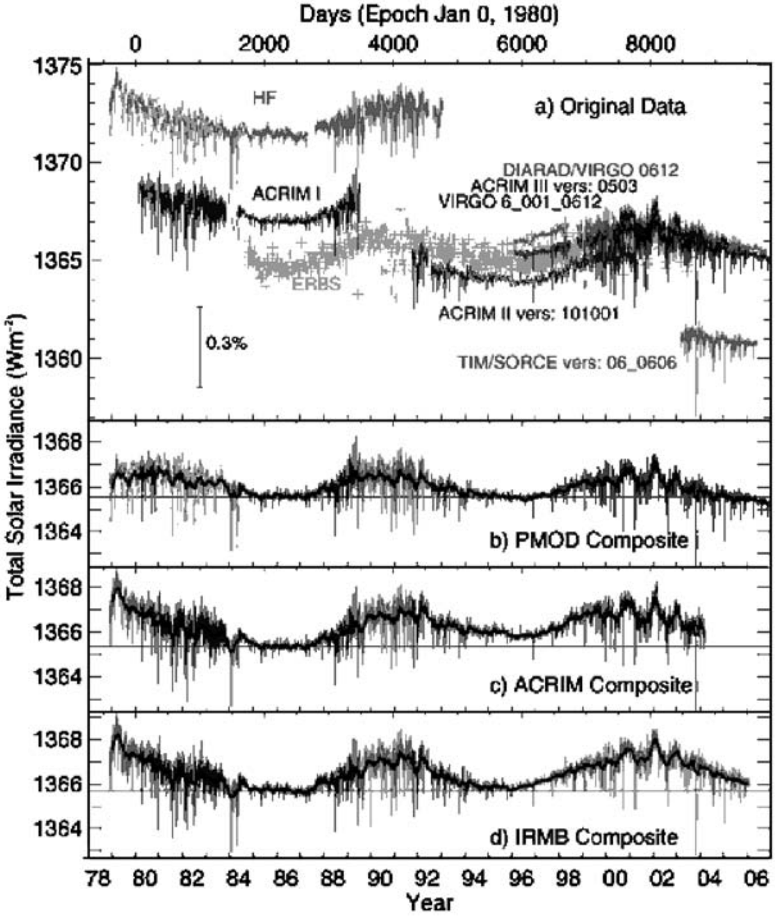


Fig. 15. – Plot showing all TSI measurements made via various instruments all nominally traceable to SI. The data presented in the upper section is largely uncorrected (although ACRIM II data has been normalised to that of ACRIM I by addition of 0.12% offset). The data in the lower three curves is what is often shown as the composite TSI shown here normalised to either of three instruments [89].

nator. Since such sensors cannot spatially resolve the surface with sufficient resolution to do this in the way the human observer could, reliance has to be made on the use of key spectral features. Everything has a unique spectral signature, although this might also vary as a function of illumination conditions, allowing it to be characterised. In many cases, sufficient information can be determined by simple ratios of two spectral bands, in others full spectral information is required.

Clearly such imaging spectrometers (cameras) need to be calibrated, and whilst often challenging, due to the physical size of the instruments, and their optics, this can be done on the ground before launch fully traceable to SI. However, while pre-launch activities

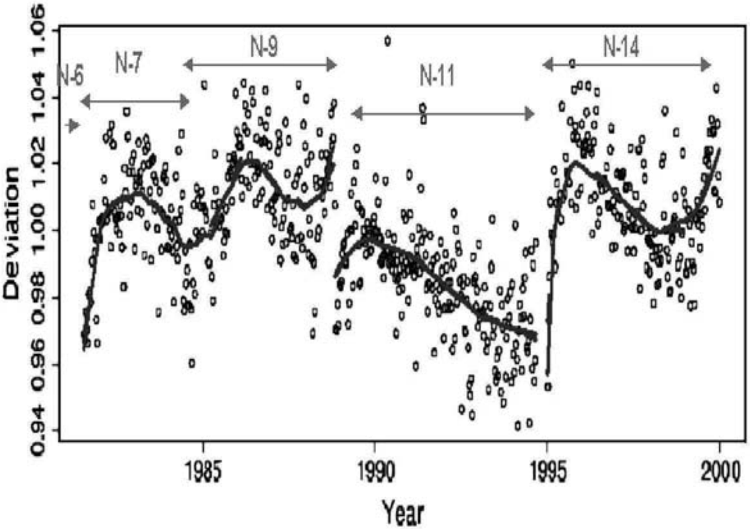


Fig. 16. – Plot showing change in NDVI of a nominally stable desert site as measured by AVHRR on the NOAA series of satellites (N6, N7, N9, N11 and N14).

help in evaluating the extent to which the instrument meets specifications, it is in the post-launch environment that the issue of traceability to SI units becomes critical. This is particularly so for the post-launch calibration of satellite sensors in the visible and near-infrared where there are many examples of pre-launch calibration coefficients needing revision due to changes in the sensor caused by storage and launch into orbit. Frequently revisions need to continue to be made through the life of the mission due to degradation of the instrument in the hostile space environment.

Figure 16 (taken from [90]) shows measurements of NDVI (Normalised Difference Vegetation Index) (based on a ratio of signal at two wavelengths) of a nominally stable desert site (results should be linear with time as there is no vegetation and thus very little change) as taken by the NOAA AVHRR series of instruments. N-7, N-9, etc. relate to flights of nominally the same instrument on board different satellites but viewing the same surface. It demonstrates the difficulty in ensuring long-term stability in-flight and consistency between instruments, even when they are of the same design. In this example, as there are no on-board calibration systems, the results can actually be used in reverse as a vicarious ground calibration to apply correction factors to normalise results to a common baseline.

How such calibration updates should best be determined is a fairly hot debate between the advocators of on-board calibration systems and those of vicarious techniques. The latter offers a range of options for determining correction factors usually based on solar reflected radiation from a "stable" natural target such as an Earth desert, snowfield or the Moon. In most cases there is little to choose between the techniques on offer, all are pretty good at monitoring change, but when determining radiometric accuracy in terms of SI units, all are relatively poor, with accuracies ranging from 3 to 10%. If such uncertainties are to be improved (and the goal is $< 1\%$), then the following issues need to be addressed [91].

- What is the attainable accuracy of radiation measurements in the visible and near-infrared with on board calibrators?
- What is attainable through vicarious techniques?
- Are the results from on-board and vicarious techniques in agreement?
- What is the accuracy required by and delivered to end users?

10˙4. *Traceable Radiometry Underpinning Terrestrial and Helio Studies (TRUTHS).* – One option for the future which would significantly improve accuracy and traceability of EO data, would be the implementation of a dedicated calibration mission such as proposed in TRUTHS (Traceable Radiometry Underpinning Terrestrial- and Helio-Studies) [41]. This proposal is for a mission to establish a set of calibration reference targets, Earth deserts, Sun and Moon to transfer calibrations of radiance or irradiance to other in-flight sensors. This is broadly similar to current best practice for vicarious calibrations. However, in TRUTHS, the calibration coefficients of these targets would be regularly updated through observation and calibration by instruments onboard a small satellite. TRUTHS instruments would be calibrated in-flight using a novel on-board procedure which mimicks that performed on the ground by NMIs when establishing primary scales, see fig. 17, left-hand panel. This has been discussed in more detail in earlier sections of this paper. This procedure includes the flight of a primary standard and so traceability to SI can be fully and regularly established in-flight, with very high radiometric accuracy ($< 0.5\%$ for spectral radiance) avoiding any problems due to drift either pre- or post- launch. A more detailed description of the concept can be found in [41].

Whilst TRUTHS offers the complete solution, the principles and techniques it proposes can be used independently. For example, the in-flight calibration system can be incorporated onto any earth viewing imager.

Similarly, it is perhaps timely to consider the establishment of a global network of a small number of calibration test sites to acquire benchmark data sets, for example GIANTS (Global Instrumented And Networked Test Sites) as proposed by Teillet *et al.* [92]. The use of "standard" ground reference calibration test sites as a means of cross-calibration and validation of satellite sensors is well established. In many cases, dedicated campaigns have been organized using teams supporting their respective instruments. In some cases, particularly atmospheric chemistry applications, use has been made of existing ground networks of validation equipment, much of which is automated. In the case of land imagers, some test sites have become recognized "standards", *e.g.*, White Sands alkali flats and Railroad Valley Playa in the Central USA and La Crau in Southern France. These and other sites have been well characterized and shown to be relatively homogeneous spatially and temporally stable (at least in the short-term). However, significant differences have been observed by different sensor teams when using

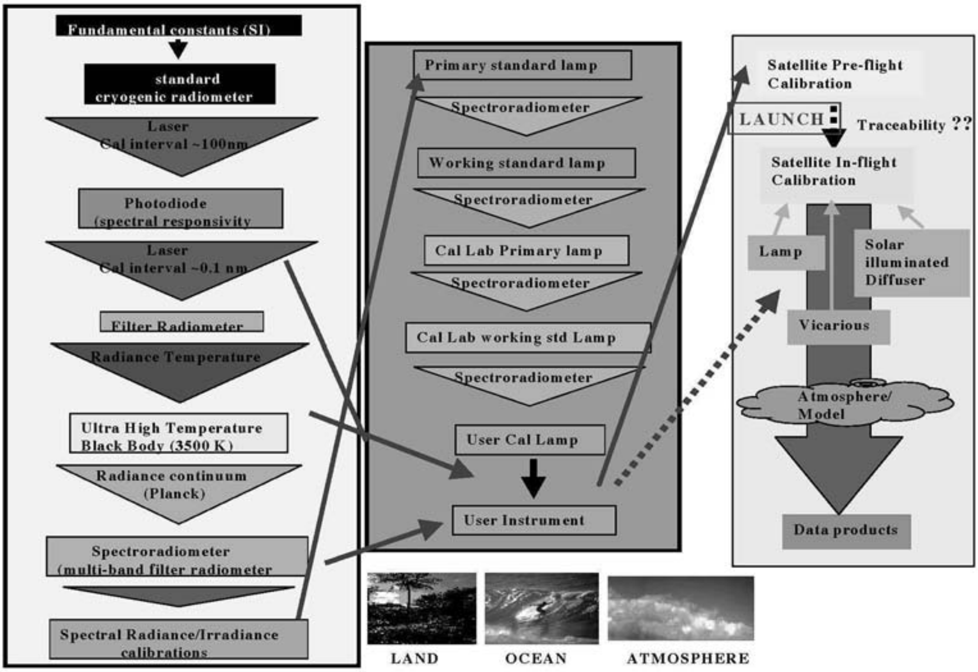

Fundamental constants (SI)
standard cryogenic radiometer
Laser Cal interval ~100nm
Photodiode (spectral responsivity
Laser Cal interval ~0.1 nm
Filter Radiometer
Radiance Temperature
Ultra High Temperature Black Body (3500 K)
Radiance continuum (Planck)
Spectroradiometer (multi-band filter radiometer
Spectral Radiance/Irradiance calibrations
Primary standard lamp
Spectroradiometer
Working standard lamp
Spectroradiometer
Cal Lab Primary lamp
Spectroradiometer
Cal Lab working std Lamp
Spectroradiometer
User Cal Lamp
User Instrument
Satellite Pre-flight Calibration
LAUNCH
Traceability ??
Satellite In-flight Calibration
Lamp
Solar illuminated Diffuser
Vicarious
Atmosphere/ Model
Data products
LAND
OCEAN
ATMOSPHERE

Fig. 17. – Schematic representation of the route of traceability for the EO community. The left and middle panels describe the typical procedure from the primary standard to the user. The right-hand panel illustrates the difficulty in obtaining traceability for a satellite instrument post-launch. The solid lines show good traceability routes, the dashed the best effort.

the same target area for vicarious calibration activities because of biases originating from subtle differences in the methodologies used, instrumentation and calibration traceability. Such biases can also occur for networked sites, although these can be reduced by the use of common instrumentation and standard methodologies. Each site requires a common set of automated instrumentation, including Sun photometers; standard meteorological parameters; video images of the site in real time; down-welling solar irradiance; and surface spectral reflectance/radiance. All instruments should be automated and transmit data independently. Continuous year round availability of a single calibration site is difficult to achieve and highly susceptible to local weather conditions, for an extreme example, snow. However, having a global network of essentially interoperable test sites overcomes this limitation.

In the context of the TRUTHS mission, data from the ground will correlate with absolute information from the TRUTHS satellite such that other satellite sensors need only be stable in the short term, which is easier to achieve than absolute calibration. However, in the near term in the absence of such a satellite, calibration updates would need to be carried out by ground support teams or aircraft overpass.

11. – Spectrophotometry, colour and sensory metrology

11˙1. *Spectrophotometry.* – Over the last few decades, the primary focus of many NMIs, including NPL, has been to establish new realisations of the various base and derived radiometric quantities linked to a cryogenic radiometer. It should however be noted that in terms of customer demand these are not the most commonly used calibration services. In fact there is a greater demand for those only requiring ratio or comparative measurements, *e.g.*, spectral reflectance, transmittance, colour, etc. The improvements in detectors and instrumentation to support the radiometric quantities are of course being utilized in these disciplines but given the nature of the measurements, little radical change has occurred and the services offered by NPL are similar to those of many NMIs. This is not to say that NPL does not also posses a range of unique and special capabilities. For example in the Infrared region of the spectrum, until recently, NPL was the only NMI providing a complete range of spectrophotometric services including reflectance, transmittance and hemispherical diffuse reflectance to wavelengths as long as $55\,\mu$m [93, 94] (more recently this has been extended to $100\,\mu$m).

11˙2. *Colour.* – A related topic is the area of colour, which is of course a measure of reflectance weighted by the eyes spectrally selective response, to give a measurand, which, in origin, has high-level human perception, associated with it. However in its simplest form it can be reduced to reflectance. NPL in conjunction with CERAM Research developed a set of colour reference tiles to improve the calibration and traceability of colour measuring instrumentation. The series II tiles were launched in 1983 as a set of 12 tiles representing different hues and lightnesses [95]. They were produced as both gloss or matt to suit different applications. In addition, black, white and four metameric pairs were also produced, the latter to check for discrepancies under different illumination conditions. These tiles have been widely used throughout the world with more than 4000 sets in circulation, although often such tiles are cut into small samples increasing their number still further. The use of such tiles has led to significant improvements in the ability of instrumentation to replace the human observer. However, there are still significant differences in absolute terms between users of different instrumentation caused by subtle differences in the operational characteristics of the spectrometers and colorimeters. NPL has recently carried out studies in conjunction with European partners to try to understand the causes of the differences and to develop best practice procedures to alleviate the problem [96].

11˙3. *Sensory metrology.* – NPL and other NMIs have started work to investigate the more direct measurement problem associated with colour and vision, that of "appearance" an aspect of "sensory metrology".

Sensory metrology (sometimes called "soft metrology") covers the development of measurement techniques and mathematical models that enable objective quantification of the properties of materials, products and activities that are determined by human response.

Sensory metrology, in its broadest sense, is not yet an established branch of metrology and, at present, it does not find a unique place within the structure of the SI or national

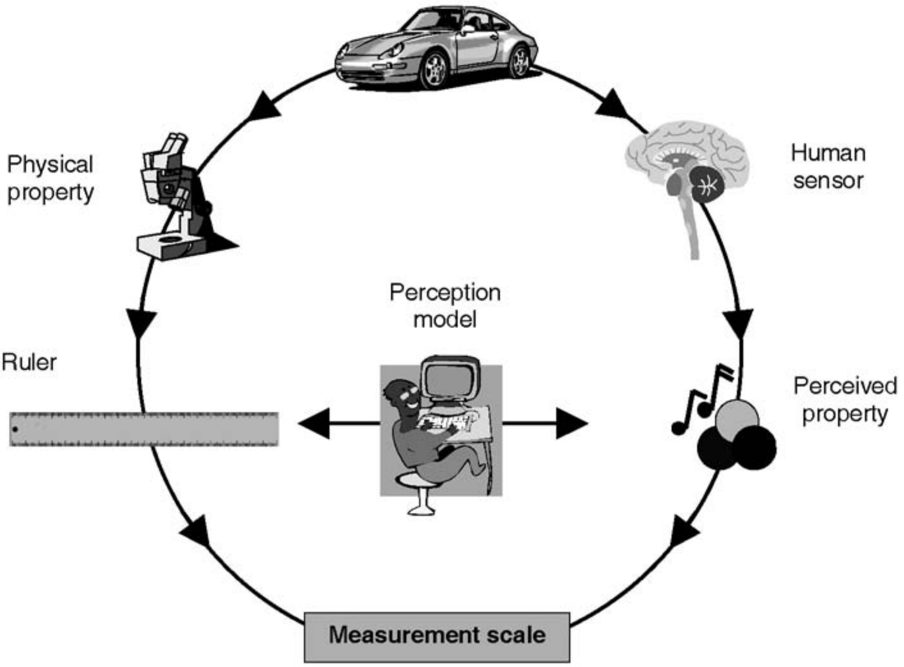


Fig. 18. – A perception model that relates a physical property of an object as measured to an aspect of that object that is defined by a human response.

measurement infrastructures. This is not to say that measurement scales do not exist or that research is not being conducted that falls within the definition of sensory metrology, but rather that these projects find a place in several of the established technical programmes.

Sensory metrology can be formally defined as:

The measurement of parameters that, either singly or in combination, correlate with attributes of human response.

Note. The human response may be in any of the five senses: sight, smell, sound, taste and touch.

Sensory metrology entails the measurement of appropriate *physical* parameters and the development of models to correlate them to *perceptual* quantities. See fig. 18. Traceable sensory metrology can be achieved both through traceable measurement of the physical parameters and the development of accurate correlation models. The most advanced topic within this domain is that related to the human visual system.

We use our visual sense constantly to process the complex patterns of light around us into objects, space, location and movement, and with that information make judgements leading to a particular course of action. For example, how we perceive the visual "appearance" of a product may cause us to buy it, reject it or even write in and complain about it! We may choose to buy a particular food because its appearance suggests freshness. Our choice of car may have been influenced because its glossy/sleek appearance made us think of quality and prestige.

The sophistication of our visual sense works against us when considering the measurement of appearance because we perceive appearance so easily that it is often difficult for us to analyse what actual physical attributes contribute to our observations and perceptions. For example, objects can exhibit, combinations of diffuse and regular reflectance, transparency, translucency etc. In many ways the physical properties being measured are the same as those for which standard techniques and reference standards already exist. However, because more complex visual effects are being sought by consumers, then many of these properties exist together making it difficult to utilize existing techniques to isolate a single measurable parameter.

Current efforts centre on the measurement of spatially and spectrally resolved reflectance (transmittance) as a function of angle of illumination and/or viewing. These physical measurands will then be correlated, using human observers, with what is perceived. It is then hoped that a model can be developed to allow prediction of the observables based on some, to be defined, sub-set of measurements [97].

12. – Conclusion

The paper has reviewed the adequacy of the establishment and interdependency of the primary radiometric quantities. It has discussed how transfer standards have been or are being developed to meet the needs of all sectors of the user community. It also gives some examples of the use of primary standards directly, to provide high-accuracy calibrations to the user.

The overlap between the thermal and radiometric communities and the need to review the methodologies used by the communities in disseminating quantities to the different user communities is discussed. In this context it proposes an alternative to ITS 90 and how its adoption might provide the user with a closer approximation to thermodynamic temperature at a lower cost and in a more convenient form.

The use of filter radiometers and their critical importance to modern radiometry can be seen as a common thread through the paper as can the use of "reflectors" to enhance performance. The paper shows how such filter radiometers not only link the primary quantities together but also provide the best means of their dissemination. Improving their performance provides a good example of the technology areas which remain a challenge to the community:

- improved measurement of geometric defining systems *e.g.* apertures and diffraction,

- stable well-blocked filters,

- efficient calibration methodologies.

Of course the most critical application needing improved optical radiation metrology is that of climate change, in order to establish accurate reliable baselines from which to monitor change. However, perhaps the most challenging will be the attempt to correlate physical observables with those perceived by humans and in effect develop a "quality" meter.

REFERENCES

[1] Ångström K., *Nova Acta Soc. Sci. Ups. Ser.* **3**, **16** (1893) 1.

[2] Kurlbaum F., *Ann. Phys. (Leipzig)*, **287** (1894) 591.

[3] *Proceedings 16th Conference General Poids et Measures* (1979). www.bipm.org.

[4] Wende B., *Metrologia*, **32** (1995/96) 419.

[5] Hollandt J., Seidel J., Klein R., Ulm G., Migdall A. and Ware M., *Primary sources for use in radiometry*, in *Optical Radiometry*, edited by Parr A. C., Datla R. U. and Gardiner J. L. (Elsevier Academic Press, Amsterdam) 2005.

[6] Metzdorf J., *Metrologia*, **30** (1993) 403.

[7] Yoon H. W., Gibson C. E. and Barnes P. Y., *Appl. Opt.*, **41** (2002) 5879.

[8] Woolliams E. R., *Development and evaluation of a high temperature blackbody source for the realisation of NPL's primary spectral irradiance scale*, PhD Thesis, University of Manchester (2003).

[9] Sapritsky V. I., *Metrologia*, **32** (1995/96) 411.

[10] Yamada Y., Sakate H., Sakuma F. and Ono A., *Metrologia*, **36** (1999) 207.

[11] Schwinger J., *Phys. Rev.*, **75** (1949) 1912.

[12] Wende B., *Metrologia*, **32** (1995/96) 419.

[13] Rastello M. L., this volume, p. 611.

[14] Migdall A. M., *Phys. Today*, **52** (1999) 41.

[15] Klyshko D. N., *Photon and Non-linear Optics* (Gordon and Breach, New York) 1988.

[16] Cheung J. Y., Chunnilall C. J., Woolliams E. R., Fox N. P., Mountford J. R., Wang J. and Thomas P. J., *J. Mod. Opt.*, **54** (2007) 373.

[17] Geist J., Lind M. A., Schaefer A. R. and Zalewski E. F., *Natl. Bur. Stand. (U.S.) Tech. Note* 954 (1977).

[18] Anderson V. E., Fox N. P. and Nettleton D. H., *Appl. Opt.*, **31** (1992) 536.

[19] Zalewski E. F. and Geist J., *Appl. Opt.*, **19** (1980) 1214.

[20] Geist J., Zalewski E. F. and Schaefer A. R., *Appl. Opt.*, **19** (1980) 3795.

[21] Boivin L. P. and Smith T. C., *Appl. Opt.*, **17** (1978) 3067.

[22] Martin J. E., Fox N. P. and Key P. J., *Metrologia*, **21** (1985) 147.

[23] Geist J. and Baltes H., *Appl. Opt.*, **28** (1989) 3929.

[24] Geist J., Chandler-Horowitz D., Robinson A. M., James C. R., Köhler R. and Goebel R., *J. Res. Natl. Inst. Stand. Technol.*, **96** (1991) 463, Parts I-III.

[25] Geist J., private communication (1999).

[26] Gran J., *Accurate and independent spectral response scale based on silicon trap detectors and spectrally invariant detectors*, PhD Thesis University of Oslo (2005).

[27] Bittar A., *Metrologia*, **32** (1995/96) 497.

[28] Kubarsepp T., Karha P. and Ikonen E., *Metrologia*, **37** (2000) 441.

[29] Quinn T. J. and Martin J. E., *Philos. Trans. R. Soc. London*, **316** (1985) 85.

[30] Fox N. P. and Rice J. P., *Absolute radiometers*, in *Optical Radiometry*, edited by Parr A. C., Datla R. U. and Gardiner J. L. (Elsevier Academic Press, Amsterdam) 2005.

[31] Ginnings D. C. and Reilly M. L., *Temp.*, **4** (1972) 339.

[32] *CODATA Bulletin*, **63** (1986) 11 (Oxford Pergamon Press).

[33] Blevin W. R. and Brown W. J., *Metrologia*, **7** (1971) 15.

[34] Blevin W. R., Geist J. and Quinn T. J., private communication.

[35] Zalewski E. F., Key P. J., Martin J. E. and Fowler J. B., private communication.

[36] Fox N. P., Haycocks P. R., Martin J. E. and Ul-haq I., *Metrologia*, **32** (1995/96) 581.

[37] Hoyt C. C. and Foukal P. V., *Metrologia*, **28** (1991) 163.

[38] Varpula T., Seppa H. and Saari J.-M., *IEEE Trans. Instrum. Meas.*, **38** (1989) 558.

[39] Datla R. U., Stock K., Parr A. C., Hoyt C. C., Miller P. J. and Foukal P. V., *Appl. Opt.*, **31** (1992) 7219.

[40] Martin J. E. and Fox N. P., *Solar Phys.*, **152** (1994) 1.

[41] Fox N. P., Aiken J., Barnett J. J., Briottet X., Carvell R., Frohlich C., Groom. S. B., Hagolle O., Haigh J. D., Kieffer H. H., Lean J., Pollock D. B., Quinn T., Sandford M. C. W., Schaepman M., Shine K. P., Schmutz W. K., Teillet P. M., Thome K. J., Verstraete M. M. and Zalewski E., *Adv. Space Res.*, **32** (2003) 2253.

[42] Foukal P. V. and Jauniskis J, *Metrologia*, **30** (1993) 279.

[43] Lau-Främbs A., Rabus H., Kroth U., Tegeler E., Ulm G. and Wende B., *Metrologia*, **32** (1995/96) 571.

[44] Boivin L. P. and Gibb K., *Metrologia*, **32** (1995/96) 565.

[45] Schrama C. A., Bosma R., Gibb K., Reijn H. and Bloembergen P., *Metrologia*, **35** (1998) 431.

[46] Goebel R. and Stock M., *Metrol. Tech. Suppl.* (2004) 02004.

[47] `http://www.bipm.org/en/cipm-mra/`.

[48] Köhler R., Goebel R., Pello R., Touayar O. and Bastie J., *Metrologia*, **32** (1995/96) 551.

[49] Goebel R., Pello R., Haycocks P. and Fox N. P., *Metrologia*, **33** (1996) 177.

[50] Goebel R., Pello R., Stock K. D. and Hofer H., *Metrologia*, **34** (1997) 257.

[51] Stock K. D., Hofer H., White M. and Fox N. P., *Metrologia*, **37** (2000) 437.

[52] Fox N. P. and Martin J. E., *Appl. Opt.*, **29** (1990) 4686.

[53] Martin J. E and Haycocks P. R., *Metrologia*, **35** (1998) 229.

[54] Quinn T. J. and Martin J. E., *Metrologia*, **28** (1991) 155.

[55] Nettleton D. H., Prior T. R. and Ward T. H., *Metrologia*, **30** (1993) 425.

[56] Fox N. P., Theocharous E. and Ward T., *Metrologia*, **35** (1998) 535.

[57] Werner L. and Fischer J., *Metrologia*, **35** (1998) 403.

[58] Zalewski E. F. and Duda C. R., *Appl. Opt.*, **22** (1983) 2867.

[59] Manufactured by Graseby Optronics Inc., Florida, USA.

[60] Fox N. P., *Metrologia*, **28** (1991) 197.

[61] Cromer C., private communication.

[62] Gardner J. L., *Appl. Opt.*, **33** (1994) 5914.

[63] Köhler R., Goebel R. and Pello R., *Metrologia*, **32** (1995/96) 463.

[64] Fox N. P., *Metrologia*, **30** (1993) 321.

[65] Kuschernus P., Rabus H., Richter M., Scholze F., Werner L. and Ulm G., *Metrologia*, **35** (1998) 355.

[66] Durant N. M. and Fox N. P., *Metrologia*, **30** (1993) 345.

[67] Richter M., Johannsen U., Kuschnerus P., Kroth U., Rabus H., Ulm G. and Werner L., *Metrologia*, **37** (2004) 515.

[68] Fox N. P., Prior T. R., Theocharous E. and Mekhontsev S. N., *Metrologia*, **32** (1995/96) 609.

[69] Fischer J., this volume p. 427.

[70] CIE 1970, **18**, E-1.2. `www.cic.co.at`.

[71] Fox N. P., Martin J. E. and Nettleton D. H., *Metrologia*, **28** (1991) 357.

[72] Friedrich R., Fischer J. and Stock M., *Metrologia*, **32** (1995/96) 509.

[73] Woolliams E. R., Machin G., Lowe D. H. and Winkler R., *Metrologia*, **43** (2006) 11.

[74] Fox N. P., *Proc. TempMeko* (VDE, Berlin) 2002, p. 27.

[75] Johnson B. C., Cromer C. L., Saunders R. D., Eppeldauer G., Fowler J., Sapritsky V. I. and Dezsi G., *Metrologia*, **30** (1993) 309.

[76] White M. G., Fox N. P., Ralph V. E. and Harrison N. J., *Metrologia*, **32** (1995/96) 431.

[77] Sperfeld P., Raatz K. H., Nawo B., Möller W. and Metzdorf J., *Metrologia*, **32** (1995/96) 435.

[78] Madden R. P., O'Brian T. R., Parr A. C., Saunders R. D. and Sapritsky V. I., *Metrologia*, **32** (1995/96) 425.

[79] Sperfeld P., Yousef G. S., Metzdorf J., Nawo B. and Moller W., *Metrologia*, **32** (1995/96) 435.

[80] Goodman T. M. and Key P. J., *Metrologia*, **25** (1988) 29.

[81] Ohno Y., *Metrologia*, **35** (1998) 473.

[82] Ohno Y., Kohler R. and Stock M., *Metrologia*, **37** (2000) 583.

[83] Metzdorf J., Sperling A., Winter S., Raatz K. H. and Moller W., *Metrologia*, **35** (1998) 423.

[84] Harrison N. J., Wooliams E. R. and Fox N. P., *Metrologia*, **37** (2000) 453.

[85] Johnson B. C., Brown S. W. and Yoon H. W., *Metrologia*, **37** (2000) 423.

[86] Fox N. P., Chunnilall C. J. and White M. G., *Metrologia*, **35** (1998) 555.

[87] Murdock T. L. and Pollock D. B., NIST GCR 98-748 1998.

[88] Fröhlich C., *Solar Irradiance Variability Since 1978*, *Space Sci. Rev.*, **125** (2006) 53.

[89] `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`.

[90] Los S. O., *IEEE Trans. Geosci. Remote Sensing*, **36** (1998) 206.

[91] Fox N. P., *Validated data and removal of bias through traceability to SI units*, in *Post-launch calibration of satellite sensors*, edited by Morain S. A. *et al.* (Taylor and Francis, London) 2004, pp. 31-42.

[92] Teillet P. M., Thome K. J., Fox N. and Morisette J. T., *Proc. of SPIE*, **4550** (2001) 246.

[93] Clarke F. J. J., *Proc. Soc. Photo-Opt. Instrum. Eng.*, **2776** (1996) 184.

[94] Clarke F. J. J. and Larkin J. A., *High Temp. High Pressures*, **17** (1985) 89.

[95] Malkin F. and Verrill J., *Proceedings of CIE 20th Session, Amsterdam* (1983), `www.cic.co.at`.

[96] Clarke P. J., NPL *Good Practice Guides*, *GPG(096) Surface Colour Measurements* (2006).

[97] Pointer M. J., CIE Technical Report 175:2006 (2006).

Classical and quantum techniques for photon metrology

M. L. Rastello

Istituto Nazionale di Ricerca Metrologica - 91 strada delle Cacce, Turin 10135, Italy

The quantum theory of electromagnetic radiation is reflected in the field of metrology by considering quantities related to the number of photons instead of conventional radiant quantities, intended in a Maxwellian frame. This paper is intended to introduce some basic methods for photon radiometry. First, classical standards are discussed for operation at optical power levels higher than some nanowatt, when detectors average on the incident radiation with a continuous quantity as an output. The way how to validate them is also discussed. Then, some fundamental concepts are introduced to bridge between classical radiometry and quantum photon radiometry. Finally, quantum photon techniques are introduced and their main characteristics are discussed, in particular the photon metrology with entangled photons.

1. – Introduction

Advances in quantum optical technology need access to accurate measurements traceable to the International System (SI), on the one hand, and open up the possibility of counting and controlling photons one by one, on the other. This revolution may lead to new realizations of the SI units with improved accuracy as managing and counting photons one by one or engineering electrically driven single-photon sources will probably lead to a given flux with a well-established number of photons per second with an unexpected accuracy.

Moreover, manipulating photons one by one will determine new instrumentation, claiming for appropriate methods and units for calibration, not to forget that advances in metrology always enable new technology. In fact, the development of optical quantum devices is expected to grow exponentially in particular in the light of the potential of optical quantum computing and quantum information, not to mention the progress done in experiments with individual quantum systems, and their applications in quantum cryptography, teleportation, frequency standards, distributed quantum computation, multi-particle entanglement swapping, multi-particle entanglement purification and interaction free measurement. Also the result of optical radiation acting on human eyes can be described basically as one photon interacting with one molecule. For instance, the minimum luminous signal that has to enter the human eye to be detected by the brain is around 600 photons per second. The collection efficiency is about 10% so the minimum number of photons converted into signal to the brain is around 60. Just the interaction of light with the human eye in vision is the only photobiological quantity considered in metrology up to now. Despite of this, photometry and radiometry use mainly the wave description, even if with some exceptions, *i.e.* when dealing with detector properties like quantum efficiency.

Classically, radiometry evaluates the energy propagating as radiation according to Maxwell equations as an interaction between oscillating electric and magnetic fields, from a source through a transparent medium to a suitable receiver. Classical standards for photon quantities operate at an optical power level higher than some nanowatt, equivalent to 10^9 photons/s in the visible spectrum. The detectors are analogue ones, *i.e.* unable to resolve any information about the arrival of photons to discriminate the behavior of light emission, but averaging on the incident radiation with a continuous quantity (a current or a voltage) as an output. Section **2** is devoted to classical standards for photon radiometry and the way to validate them.

The quantum theory of electromagnetic radiation reflects in the field of metrology by the replacement of both radiant and luminous quantities, intended in a classical frame, with quantities related to the number of photons. In other words, photon radiometry is the measurement of electromagnetic radiation in terms of photon quantities. In particular, the quantum correlation of paired photons allows the development of new quantum techniques for photon metrology. The quantum aspects of radiometry give the chance to perform investigation in the very low intensity domain, known as photon-counting regime, where the classical reference standards are inadequate, and entanglement represents a unique resource. For photon fluxes below 10^9 photons/s, counting photons directly will not work as a primary method, because the detector has an uncertainty in its quantum efficiency. In the last decade new primary techniques for photon counters and detectors for extremely low optical powers (10^{-12} W to 10^{-18} W) have been realized through correlated photons. In sect. **3**, some fundamental concepts are introduced to draw a bridge between the classical radiometric arguments and the fundamental quantity of quantum photon radiometry, namely the number of photon per mode. Finally in sect. **4** quantum photon techniques are discussed and the main aspects of this counterpart of conventional optical radiometry in quantum domain are discussed, in particular the photon metrology with entangled photons.

2. – Classical primary methods

Photon standards refer any way to the quantum theory of light by the definition of a single quantum of radiation as a photon. Each photon is characterized by a specific frequency ν and its energy $E = h\nu$, where h is the Planck constant. According to International Lighting Commission (CIE), the photon number is the number of photons emitted by a source or propagating onto, through or emerging from a specified surface of a given area in a given period of time and in a given interval of frequencies. Radiant and photon quantities are related by the following

$$X_p = \int_\nu X_{e,\nu} \frac{1}{h\nu} \mathrm{d}\nu \,,$$

where X_p is the photon quantity and $X_{e,\nu}$ is the related spectral radiant quantity.

Primary methods are needed for relating the photon quantity to some fundamental constants and/or independent physical quantities expressed in SI unit [1-3]. Basically, there are two approaches for assessing a primary photon standard, absolute sources and absolute detectors [4, 5]. Here absolute means that their properties are linked through fundamental constants and physical laws, therefore avoiding the need for a calibration by comparison with already existing radiometric standards. For instance, a black-body radiator is a standard source, as the Planck law, involving the Planck constant h, the speed of light c, the Boltzmann constant k and the thermodynamic temperature T [6], predicts its spectral photon radiance.

In the same way, a detector is considered absolute if fundamental physical laws can estimate its responsivity. As suggested in 1980 by J. Geist and E. Zalewski in classical radiometry, silicon photodiodes can play also the role of primary photon standards [7], their ideal photon responsivity being linked to the frequency by fundamental constants only. The operation principle of semiconductor photon detectors underpins on the photoelectric effect, which is the generation of a free-electron–hole pair at the absorption of a photon.

The average number of free-electron–hole pairs produced per incident photon defines the quantum efficiency of the device. In commercial silicon photodiodes the quantum efficiency in the visible range is close to 1 to within a few tenths of one percent, because some deviations from the ideal behavior are caused by loss mechanisms, mainly the non-ideal conversion of photons into measurable electrons and the reflectance from the diode surface.

An actual silicon photodiode can be modeled as a modified ideal quantum detector, where the two loss mechanisms, reflectance from the surface and internal losses, are taken into account. The responsivity is then modeled as

$$R(\lambda) = \frac{e}{h\nu} \cdot [1 - \rho(\nu)] \cdot [1 - \delta(\nu)],$$

where e is the elementary charge, h is the Planck constant, c is the speed of light in vacuum, and ν is the frequency of the impinging radiation. These quantities form the

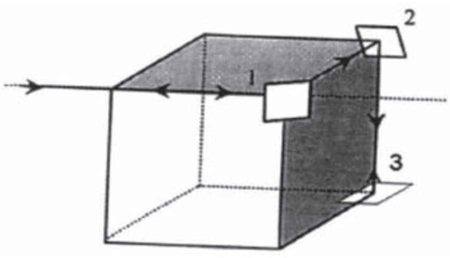


Fig. 1. – Schematic drawing of a reflection trap detector. The figure is taken from [28].

ideal term of a quantum detector. Reflectance $\rho(\nu)$ and quantum deficiency $\delta(\nu)$ depend on frequency and are to be estimated separately to get a primary standard. Both loss mechanisms can be modeled and described by a few parameters.

As shown in [8], reflection losses can be reduced to negligible levels or measured with purely relative methods. Fresnel equations allow calculating the spectral reflectance over the whole spectrum with an uncertainty of less than $1 \cdot 10^{-2}$ from a single measurement, when we know the spectrally dependent refractive indices [9-11], polarization, angle of incidence and oxide thickness. Moreover, a polarization insensitive silicon detector can be designed in a trap configuration, as shown in fig. 1, with a reflectance given by

$$\rho(\nu, d) = \rho(\nu, 0, d)\rho_s^2(\nu, \pi/4, d)\rho_p^2(\nu, \pi/4, d),$$

where indexes p and s indicate polarization directions and d is the oxide thickness.

Internal losses can be evaluated by applying suitable biases: a reverse bias for the losses in the deepest layers of the diode, and a front oxide bias for the losses in the shallow part of the diode. By increasing the applied bias, the diode responsivity increases until a saturation level is reached, the difference between saturated and unbiased response being a direct measurement of $\delta(\nu)$.

The method is called *self-calibration* and extensive work has been done on this technique [12-21] since its introduction. In particular, to avoid the change in the diode quantum efficiency after biasing the oxide, inversion layer n on p diodes were developed, showing a quantum efficiency very close to 1 only over a reduced spectral range.

Self-calibration performs at the best in an experimental scheme involving laser frequencies. Then, this set of discrete calibration points are interpolated by suitable models, based upon the same philosophy of the self-calibration procedure where the loss mechanisms, reflectance and $\delta(\nu)$ were treated separately [17, 22, 23]. The model developed by Gentile [23] allows interpolating spectral responsivity continuously between 400 and 920 nm [24-26]. This model is based on the assumption that the recombination probability for electron hole (e-h) pairs is given by the position where they were created. By adding two more parameters Werner extended the model beyond 1015 nm [24]. Kübarsepp developed a model of the quantum yield in the UV region with two free parameters, extending the model down to 250 nm [25].

To be a primary standard, the diode spectral responsivity should be linked to fundamental constants with a sufficiently known accuracy. The full covariance analysis of the results assures this link, as reported in ref. [27]. Moreover, the calculated uncertainty given from the method is at best comparable to that given at laser frequency cryogenic radiometers, due to the low $\delta(\nu)$ of the trap detector and the high stability of the source [28]. A more dense spacing between the measurements of the relative response enabled to lower the uncertainties in the responsivity based on these experiments as well. In this way an independent high-accuracy spectral responsivity scale has been established over a broad spectral range by a primary method, where the responsivity of the detector is linked to fundamental constants.

The original self-calibration procedure [7] was not optimized for high accuracy but for simple, low-cost calibrations of commercial silicon photodiodes. Uncertainties of a few parts in 10^4 in the mid-visible appear to be the limit of conventional self-calibration with commercial photodiodes. Some comparisons at some discrete frequencies [16, 29] showed a good agreement between the self-calibration procedure and the results by cryogenic radiometers, which have been proved to be able to measure the optical power of a laser beam with an uncertainty of $4 \cdot 10^{-5}$ [30].

Now, uncertainty can improve to 1 part in 10^6 or better for reflectance traps of custom photodiodes optimized for both self-calibration and operation under reverse bias at cryogenic temperature [31]. Eleven physical mechanisms have been identified to cause the departure of the diode quantum efficiency from its ideal value of unity. Provided that the deviations caused by each of these is much less than 1, then their effects will be additive, and the quantum deficiency can be written as

$$\delta = \delta_{ad} + \delta_\rho + \delta_{ap} + \delta_{af} + \delta_\tau + \delta_{ri} + \delta_{rA} + \delta_{rbb} + \delta_{ai} + \delta_{iv} + \delta_g \,, \tag{1}$$

where δ_{ad} is the fraction of the incident photons that are absorbed by dirt particles on the front surface of the photodiode, δ_ρ is the fraction of the photons that are lost due to front-surface reflectance from the photodiodes in a trap configuration, δ_{ap} is the fraction of the incident photons that are absorbed in the photodiode passivation layer at the front of the photodiode, δ_{af} is the fraction of the incident photons that are absorbed by free carriers in the photodiode, δ_τ is the fraction of the incident photons that are transmitted out the rear of the photoactive material of all of the photodiodes in a trap configuration, δ_{ri} is the fraction of the photogenerated carriers that recombine through transitions involving an intermediate imperfection state in the silicon lattice, δ_{rA} is the fraction of the photo-generated carriers that recombine through an inverse Auger process that transfers the recombination energy to a free carrier, δ_{rbb} is the fraction of the photo-generated carriers that recombine by a momentum-conserving direct transition of a conduction-band electron into an empty state (a hole) in the valence band, δ_{ai} is the fraction of the photo-generated carriers that are absorbed by imperfection states without the creation of free carriers, δ_{iv} is the fraction of the photo-current that is lost due to photo-current-induced forward bias of the photodiode, and δ_g is the fraction of the photocurrent that is produced by the acceleration of free carriers by the built-in photodiode electric field

Table I. – *Tentative loss terms for photodiode meeting preliminary specifications.*

Loss mechanism	Symbol	Nominal value ($\times 10^{-9}$)	Reference
Dirt-particle absorption	δ_{ad}	< 1	(none)
Trap reflectance	δ_{ρ}	≤ 6	[33,34]
Thermal-oxide absorption	δ_{ap}	$\ll \delta_{af}$	(none)
Free-carrier absorption	δ_{af}	≈ 1000	[35,36]
Silicon-substrate transmittance	δ_{τ}	< 1	[34]
Interface recombination	δ_{ri}	≤ 2.5	[32]
Auger recombination	δ_{rA}	0	[32]
Transition recombination	δ_{rbb}	0	[32]
Imperfection absorption	δ_{ai}	$\ll \delta_{ri}$	(none)
Volume recombination	δ_{rb}	≈ 0.1	(none)
Reverse-bias dark current	$\lvert\delta_{iv}\rvert$	$\ll 1$	[32]
Field-induced gain	δ_{g}	≤ 15	[37]

to sufficient velocity to cause Auger (impact) ionization of valence-band electrons. The last two terms are actually gain terms (negative loss terms).

Simulation [32] was used to study the dependence of δ_{rA}, δ_{rbb}, δ_{ri}, and δ_{iv} on photodiode design and operating parameters. Contributions to δ_{ri} from both interface states at the oxide-silicon interface and localized states of the bulk silicon were modeled. Literature data were used to study the other loss terms. The results of the loss study were used to define the set of preliminary photodiode specifications listed in ref. [31].

The photodiode described in the preliminary specifications is unusual in that it has no intentional dopants. The built-in field is created entirely by oxide-trapped charge on the front surface, oxide-bias charge on the rear surface, and is augmented by reverse bias. All but the last three specifications are well within the current state of the art. It is not clear that the last three can actually be met with real photodiodes. What is clear is that producing photodiodes to meet these specifications will be a major effort. It will require the development of fabrication capability and techniques specifically designed for this purpose. Similarly, new measurement capability will have to be developed specifically to allow characterization of the various loss terms.

Specifications in [31] are sufficient to reduce the loss terms to the values given in table I, when the custom photodiode is operated at 72 K and 16 V reverse bias while being irradiated by 2.72 mW/cm of 730.5 nm radiation.

Lower temperatures or higher reverse-bias voltages may provide even better performance. The combination of 2.72 mW/cm and 730.5 nm was chosen because it gives a good approximation to $1 \cdot 10^{16}$ photons per (cm second). The simulations also predicted that the dark current at 72 K would be negligible. In fact, the dark current is so low that is predicted to remain below one part in 10^9 even when the photodiode area exceeds 1000 cm^2 and the irradiance is confined to an area of 1 cm^2 for a total radiant power of 2.72 mW. This result facilitates the construction of very low-reflectance photodiode traps [8] for very accurate applications while providing a shot-noise uncertainty of 1 part in 10^8 for a one-second-measurement period. It is the combination of cryogenic temperatures, reverse-bias operation, and low doping that increases the upper end of the linearity range of the photodiode well beyond what is available from induced-junction photodiodes at room temperature.

Because the target photodiodes will be custom fabricated and because the dark current is so low at cryogenic temperatures, the constraints on photodiode size and shape are very much relaxed compared to those with commercial photodiodes. With this design freedom, a 15-reflection trap that collects most of the specularly reflected radiation is easily designed to minimize ρ. Furthermore, the small fraction of the scattered (non-specularly reflected) radiation that leaves the trap can be accurately measured with a scattermeter that has been described previously [38]. Some modifications of the conventional self-calibration appear desirable. Specifically, photodiodes should be used under maximum reverse bias rather than unbiased, and the oxide-bias experiment should be performed on test photodiodes produced specifically on the same wafer as the trap photodiodes. The proposed modifications and extensions of the self-calibration procedure include electrically calibrated photo-acoustic radiometry, a modified water drop experiment, and an extension of the oxide-bias experiment to witness sample photodiodes. The first is needed to measure extremely low absorption losses in thin silicon-dioxide films, the second to measure absorption losses in accumulation layers in semiconductors, and the third to detect the onset of gain with increasing oxide bias.

So far these ideas are just challenges for high-accuracy radiometry, but they may also turn out to be opportunities. These results may also be useful for developing transfer standards with reduced uncertainties for three reasons. First, the reduction of δ in eq. (1) without the need for any supporting self-calibration experiments is a sufficient condition for improved transfer accuracy. Second, a spatial uniformity map is sufficient to determine the major contribution to the transfer uncertainty. Third, the results at liquid-nitrogen temperature should be very similar to those reported here at 72 K.

Nevertheless, a fascinating measurement scheme can be proposed to check the level of accuracy attainable with the custom diodes. Improvements in the accuracy of both silicon self-calibration and cryogenic radiometry might allow a radiometric measurement of the ratio R of the Josephson constant K_J to $2e/h$, that is the experimental check of the exactness of the relation $K_J = 2e/h$.

The theory of Josephson effect predicts the phenomenological constant K_J to be equal to the ratio $2e/h$, where e is the elementary charge and h is the Planck constant. This relation has been shown to be consistent theoretically [39], whereas Taylor and Cohen [40] found an inconsistency in the equality of 2 parts in 10^7 on the basis of the most accurate data available at that time. On the other hand, it has been shown to be independent of experimental parameter [41], thus confirming its universality. Unfortunately, the values K_J and $2e/h$ have never been compared experimentally to prove the Josephson prediction.

The idea to evaluate the ratio $R = 2eK_J/h$ is based on the measurement of sufficiently monochromatic and stable radiation both with a silicon photodiode in the photoamperometric mode and with a cryogenic radiometer and to calculate the ratio of the results.

The proposed radiometric determination is based on a few simple physical laws. Let r be the rate at which photons of frequency ν are flowing through an aperture. According to Einstein theory of the photoelectric effect, the radiant power flowing through the aperture is given by $\Phi = rh\nu$. Also, according to this theory, the maximum photocurrent (assuming a pair-creation energy greater than $h\nu/2$) that this flow can create is given by $J = re$.

According to the Helmholtz-Joule-Mayer theory of energy conservation, electrical power is given by $P = IV$, where V is the voltage drop across a resistor carrying a current I. As a result of these laws, a photodiode can be used to measure the photon rate as

$$r = \varepsilon(\nu) \cdot \frac{J}{e}, \tag{2}$$

where $\varepsilon(\nu)$ is the external quantum efficiency of the photodiode at photon frequency ν. Similarly, a cryogenic radiometer can be used to adjust the electrical power P to cause the same heating effect as the radiant power Φ, so that

$$\Phi = f(\nu) IV, \tag{3}$$

where $f(\nu)$ is the correction factor for the radiometer at photon frequency ν. Finally, eqs. (2), (3) can be combined to give

$$(h/e) = \left[\frac{V}{\nu}\right] \cdot \left[\frac{I}{J}\right] \cdot \left[\frac{f(\nu)}{\varepsilon(\nu)}\right]. \tag{4}$$

At first, it looks like eq. (4) should give a new value for h/e when all the ratios on the right side of the equation are determined with sufficient accuracy. However this is not the case because the voltage V is defined in terms of the Josephson effect by

$$V = K_J \nu_J, \tag{5}$$

where ν_J is the frequency of the microwave radiation used to stimulate the Josephson junction. Multiplication of both sides of eq. (4) by $K_J/2$ and use of eq. (5) gives a result that can be written as

$$K_J = R \cdot \left[\frac{h}{2e}\right], \tag{6}$$

where

$$R = \left[\frac{2\nu}{\nu_J}\right] \cdot \left[\frac{J}{I}\right] \cdot \left[\frac{\varepsilon(\nu)}{f(\nu)}\right]. \tag{7}$$

Thus what is actually measured in the proposed experiment is the dimensionless quantity $R = 2eK_J/h$.

If K_J were more accurately known than $h/2e$, then a radiometric determination of R of sufficient accuracy would provide a more accurate value for $h/2e$. This is not the case, as, according to Josephson theory, R is exactly one, and this is true independently of how accurately either K_J or $h/2e$ is known from experiment. As a result, eq. (5) cannot be used to determine a new value of $h/2e$. Instead it just approximates whatever value is assigned to K_J. Indeed, if Josephson theory is correct, an error-free measurement of R will just reproduce the assigned value of K_J exactly.

With sufficient accuracy, the proposed measurement is both a test of Josephson theory and a determination of a value of the Josephson constant consistent with the value $h/2e$ obtained by other experiments.

The latest adjustment of the atomic constants to experimental measurements gives $h/2e$ to about 3 parts in 10^7. Thus the proposed measurement starts to be interesting somewhere between 1 and 10 parts in 10^8. As is well known, optical frequency ν in eq. (7) can be measured to much better than 1 part in 10^9. Similarly, the Josephson (microwave) frequency ν_J can be measured to much better than 1 part in 10^9, even though the uncertainty in K_J is much larger than this. A suitably designed experimental configuration should produce a very small uncertainty in the measurement of the ratio (J/I) directly. According to above, there should be no problem in designing a measurement in which the two left-most factors on the right-hand side of eq. (7) can be measured much more accurately than 1 part in 10^9. There may be good reasons to attempt the measurement if one can determine $\varepsilon(\nu)$ and $f(\nu)$ to well within 100 parts in 10^9. Top level cryogenic radiometry [42] claims for a total uncertainty in the measurement of laser radiation within some parts in 10^6 when the window contribution could be disregarded, as it can be for the proposed experiment. However, a proper choice of the cavity geometry and coating could minimize the fraction of radiation lost out. Ideally considerably less than 1 part in 10^9 will be lost when a multiple reflection scheme would be adopted. Moreover, when heaters will be located where the first reflections of the incident radiation hit the cavity wall and designed to dissipate approximately the same amount of heat as will be absorbed, then the non-equivalence correction between radiant and electrical heating will be reduced to below 1 part in 10^9.

3. – Quantum techniques for photon metrology

The quantum theory of electromagnetic radiation is the most successful and comprehensive theory in optics and none of its predictions has been contradicted by experiments up to now. By taking in account light as discrete amounts of energy, it provides a powerful tool for addressing physical systems of few photons, *i.e.* when the average number of photons per mode is at the counting rate level.

The first indication that light might be quantized came from Max Planck when he correctly modeled black-body radiation [43] by assuming that the exchange of energy between light and matter only occurred in discrete amounts he called quanta. It was not known whether the source of this discreteness was the matter or the light. In 1905, Albert Einstein published the theory of the photoelectric effect [44], one possible explanation for the effect being the existence of particles of light called photons. Later, Bohr showed that also atoms are quantized, in the sense that they can only emit discrete amounts of energy. It was up to Dirac to combine the wave- and particle-like aspect of light. Following his work in quantum field theory [45], R. J. Glauber [46] and L. Mandel [47] applied quantum theory to the electromagnetic field to gain a more detailed understanding of the detection and the statistics of light, by studying the correlation functions to characterize the coherence properties of an electromagnetic field. This led to the introduction of the coherent state as a quantum description of laser light and the conclusion that only few states of light could not be described with classical waves. In the 1970s, Kimball demonstrated the first source of light requiring a quantum description: a single atom that emitted one photon at a time. This was the first conclusive evidence that light was made up by photons.

3˙1. *Quantization.* – According to quantum mechanics, light may be considered not only as an electromagnetic wave but also as a beam of photons traveling at the vacuum speed of light c. These particles should not be considered to be classical bullets, but as quantum-mechanical particles described by a wave function spread over a finite region.

At the heart of the quantum theory of radiation, the quantization associates each mode of the radiation field with a quantized harmonic oscillator. Light is described in terms of field operators for creation and annihilation of photons. The energy operator is expressed by the Hamiltonian

$$\mathcal{H} = \sum_{\mathbf{k},s} \hbar\omega \left[a^{\dagger}_{\mathbf{k},s}(t) a_{\mathbf{k},s}(t) + \frac{1}{2} \right],$$

where the contribution $(1/2)\hbar\omega$ to the energy of each $\mathbf{k}, s$ oscillator mode is the so-called zero-point contribution and the Hermitian operator $[a^{\dagger}_{\mathbf{k},s}(t) a_{\mathbf{k},s}(t)]$ defines the *Photon-Number* Operator:

$$\mathbf{n}_{\mathbf{k},s} = \mathbf{a}^{\dagger}_{\mathbf{k},s}(t)\mathbf{a}_{\mathbf{k},s}(t)$$

whose eigenvalues $n_{\mathbf{k},s}$ are the number of photons excited in the cavity mode; $\mathbf{k}$ represent the wave vectors, or field modes, and $s = 1, 2$ are the two possible polarization mode states. The eigenstates $|\mathbf{n}_{\mathbf{k},s}\rangle$ form an orthonormal basis and for them hold:

$$\mathbf{n}_{\mathbf{k},s}|n_{\mathbf{k},s}\rangle = n_{\mathbf{k},s}|n_{\mathbf{k},s}\rangle \quad n_{\mathbf{k},s} = 1, 2, \ldots$$

Since $\mathbf{n}_{\mathbf{k},s}$ is a Hermitian operator, the number $n_{\mathbf{k},s}$ is real. Looking at the commutation relations

$$[\mathbf{a}_{\mathbf{k},s}(t), \mathbf{n}_{\mathbf{k}',s'}(t)] = \mathbf{a}_{\mathbf{k},s}(t)\delta^3_{kk'}\delta^3_{ss'},$$

$$\left[\mathbf{a}^{\dagger}_{\mathbf{k},s}(t), \mathbf{n}_{\mathbf{k}',s'}(t)\right] = -\mathbf{a}^{\dagger}_{\mathbf{k},s}(t)\delta^3_{kk'}\delta^3_{ss'},$$

the so-called annihilation and creation operators take the form

$$\begin{aligned}\mathbf{a}_{\mathbf{k},s}|\mathbf{n}_{\mathbf{k},s}\rangle &= \sqrt{n_{ks}}|\mathbf{n}_{\mathbf{k},s}-1\rangle,\\ \mathbf{a}^{\dagger}_{\mathbf{k},s}|\mathbf{n}_{\mathbf{k},s}\rangle &= \sqrt{n_{ks}+1}|\mathbf{n}_{\mathbf{k},s}+1\rangle,\end{aligned}$$

their effect over a state being the lowering and raising the number of quanta in the state to which they are applied. The eigenvalues of $\mathbf{n}_{\mathbf{k},s}$ cannot be negative, and as a consequence the spectrum of $\mathbf{n}_{\mathbf{k},s}$ is the set of integers $0, 1, 2, 3, \ldots$.

So far we have discussed a single mode of the radiation field, but it is true that the whole set of operators $\mathbf{n}_{\mathbf{k},s}$ forms a complete ensemble of commuting observables for the field. The operator corresponding to different modes $(\mathbf{k}, s)$ operates on different subspaces of Hilbert space, therefore the state vector characterizing the entire fields is the direct product of $\mathbf{n}_{\mathbf{k},s}$ state vectors over all the modes:

$$\prod_{\mathbf{k},s}|\mathbf{n}_{\mathbf{k},s}\rangle = |\{n_{k,s}\}\rangle.$$

This is called the photon-number state or Fock state of the electromagnetic radiation field, and it is featured by the (infinite) set of occupation numbers $n_{k_1,s_1}, n_{k_2,s_2}, \ldots$ for all the modes. So, the Fock state $|\{n_{k,s}\}\rangle$ is an eigenstate of the number operator for the mode $(\mathbf{k}, s)$:

$$\mathbf{n}_{k,s}|\{n_{k,s}\}\rangle = n_{k,s}|\{n_{k,s}\}\rangle$$

and, if we define a total number operator as $\mathbf{n} \equiv \sum_{k,s}\mathbf{n}_{k,s}$, we obtain

$$\mathbf{n}|\{n\}\rangle \equiv \left(\sum_{k,s}\mathbf{n}_{k,s}|\{n\}\rangle\right) = \mathbf{n}|\{n\}\rangle,$$

thus indicating that the Fock state is also an eigenstate of $\mathbf{n}$, with an eigenvalue which is the total occupation number n summed over all modes. At the same time the state $|\{0\}\rangle$ for which all the occupation numbers are zero, has the lowest eigenvalue of n. Any Fock state can be built up by repeating the application of the raising operators on the vacuum in such a way:

$$|\{m\}\rangle = \prod_{k,s}\left[\left(\mathbf{a}^{\dagger}_{k,s}\right)^{m_{k,s}}\right]|\mathrm{vac}\rangle.$$

The discrete excitations or quanta of the electromagnetic field, corresponding to the occupation numbers $\{n\}$, are called photons, and a state $|\ldots, 0, 0, 1_{k,s}, 0, \ldots\rangle$ is described as a state with one photon of wave vector $\mathbf{k}$ and polarization s. The eigenvalues of the photon number operator $\mathbf{n}_{k,s}$ are unbounded, so that an arbitrarily large number of photons may be found in the same quantum state, which means that the photons are bosons and obey the Bose-Einstein statistics. The energy eigenvalue can be interpreted to mean that each of the $n_{k,s}$ photons belonging to the mode k, s carries the energy $h\omega$, which is independent of both the polarization s of the photon and the direction of the wave vector $\mathbf{k}$: it depends only on the frequency ω.

Then, photons are defined as quantum excitations of the normal modes of the electromagnetic field, and are associated with plane waves of definite wave vector $\mathbf{k}$ and polarization s. A plane wave is not localized in space or time, but sometimes it is necessary to deal with photons that are localized to a certain extent and propagate with the velocity of light (the concept of position is meaningful only in a restricted sense). A state can be introduced corresponding to an approximately localized photon at a given time:

$$|\Psi\rangle = C \sum_{\mathbf{k}} e^{-\frac{-(\mathbf{k}-\mathbf{k}_0)}{\sigma^2}} e^{-i\mathbf{k}\cdot\mathbf{r}_0} |\mathbf{1}_{k,s}\{0\}\rangle,$$

which is a linear superposition of one-photon Fock states with different wave vectors $\mathbf{k}$, where C is the normalization constant. The state $|\Psi\rangle$ is an eigenstate of the total photon number operator with eigenvalue unity, so that is a one-photon state. In particular the wave vector of this state has no definite value, but a Gaussian spread about $\mathbf{k}_0$, so that the corresponding photon can be regarded as being approximately localized in the form of a wave packet centered at position r_0 at a given time.

A direct consequence of the quantization of radiation is the so-called vacuum-fluctuation, associated with the zero-point energy, These have no counterpart in the classical description of radiation, and represents a basic concept to a proper interpretation of phenomena like spontaneous emission, photon statistics of laser, two-photon interferometry and consequent production of entangled states, squeezed states, sub-Poissonian statistics and photon antibunching.

The state $|\{0\}\rangle$, for which all the occupation numbers are zero, has the lowest eigenvalue of $\mathbf{n}$ and is known as the *vacuum state* $|\text{vac}\rangle$. The zero-point energy is linked to a vacuum state $|\text{vac}\rangle$, the state of lowest energy and corresponds to the state for which all the occupation numbers n_k are zero. This reflects the fact that, according to the uncertainty principle, a quantum-mechanical harmonic oscillator can never come to rest, not even in the ground state. Although no excitations are present, the total energy does not vanish. The energy of this state is the sum of the energies of the ground state of each harmonic oscillator and it is infinite because there is no upper boundary to the field frequency even if the volume is finite. It is possible to eliminate this term by shifting the energy of the ground state to allow the solution to remain finite and well behaved when the number of modes is allowed to become infinite at a later stage of the calculation. This infinite energy shift cannot be detected, since the experiments measure only energy differences from the ground state of $\mathcal{H}$.

Moreover, as a consequence of the quantization of the radiation field, the commutator between the field energy and the electric field E does not vanish, and the same happens for the magnetic field H. This means that in the vacuum state the E and H fields fluctuate. These fluctuations are proportional to the square of the fields because the expectation value of the fields vanishes for the vacuum state, so in the case of an electric field E there holds

$$\langle 0|\mathbf{E}^2|0\rangle = \frac{\hbar}{\epsilon_0 V}\sum_{\mathbf{k}}^{+\infty}\omega_{\mathbf{k}} = \frac{\hbar}{\epsilon_0(2\pi)^3}\int_0^{\infty}\omega_{\mathbf{k}}\mathrm{d}^3k,$$

for which the fluctuations are infinite for an unbounded set of modes, but in practice, measurements are made over a finite region of time, frequencies and space. Detectors are sensitive to a finite number of frequencies and a measurement is the average over a space-time region and bandwidth. The fluctuations are then finite.

Vacuum Fluctuations are responsible of counter-intuitive phenomena like the generation of entangled photons by Spontaneous Parametric Down-Conversion (SPDC) or Parametric Scattering (PS), when a high-energy photon of the pump field spontaneously decays into correlated pairs.

3˙2. *Entanglement.* – Entanglement is the characteristic trait of quantum mechanics, the term Entanglement being a free translation of the word Verschränkung, introduced in 1935 by Schrödinger to characterize this special feature of composite quantum systems. The entangled photon states have been applied to search for definitive results also in quantum mechanics foundations field [48], as for example in the Einstein-Podolsky-Rosen (EPR) paradox, Bell's inequalities tests, entangled photon-induced transparency.

Modern laser techniques enable to prepare the light field in a two-photon state (plus the vacuum component) which in some approximation is a pure state [49]. The behavior of this conjugate pair of photons [50], known as an entangled state, is that though each individual particle exhibits an inherent uncertainty, the joint entity of an entangled pair can exhibit no such uncertainty: while the time of arrival of an individual particle may be totally random, an entangled pair must always arrive simultaneously. The entanglement exhibited by these states is reflected in the high correlation in energy, momentum, polarization and simultaneous emission of the two entities, and it has been recognized as a fundamental resource for quantum technology, in particular because the realization of efficient sources of entangled states is of the utmost relevance.

The emerging fields of quantum communication and quantum information rely basically on the entangled properties linking two or more quantum variables. For instance, in quantum cryptography the use of entangled photon states make the fundamental laws of quantum mechanics protect the communications, the key feature relying on the impossibility to clone the quantum state or extract information without destroying it.

SPDC is a purely quantum phenomenon due to an induced material polarization proportional to the second power of the electric field (second-order non-linear parametric process) and involves the interactions between three electromagnetic fields inside a non-linear medium; as a consequence of the dielectric properties the response of the medium implies an exchange of energy between electromagnetic fields of different frequencies through annihilation and creation of photons: in particular when a high-energy photon, the pump, enters a non-linear crystal, it may spontaneously decay with a certain very

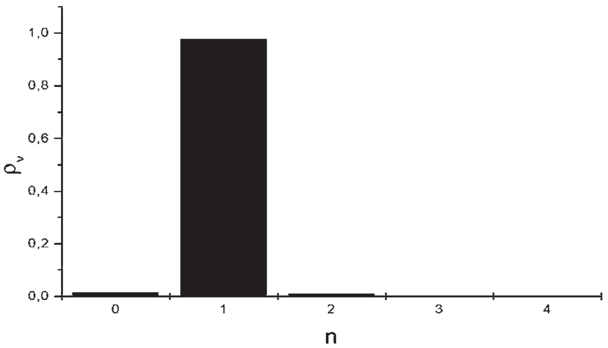


Fig. 2. – Photon distribution for heralded single-photon states from a type-II SPDC.

low probability into two lower energy photons, the signal and the idler, fundamentally related by the laws of energy and momentum. The detection of an idler photon heralds the existence of its twin signal photon: the emission time, direction, wavelength and polarization of the so-called bi-photon are correlated, or entangled, that means given the values for the idler, those of the signal can be inferred.

In metrology the quantum correlation of paired photons allows the design and the realization of absolute measurements without requirement of any radiometric absolute standards, giving rise to the application of new quantum techniques to photon metrology. This quantum photon radiometry performs investigation in the very low-intensity domain, known as photon-counting regime, where the classical reference standards are inadequate, and the entanglement by SPDC represents a unique resource. On the other hand, SPDC offers the opportunity of developing a practical standard photon generator that radiates a known number of photons, making it possible to realize also "gedanken" experiments, and at the same time photon counting can be used to determine the number of photons which arrive at a device: in this way the reference at photon-counting level is realized, to perform measurements with very low intensity signal levels below 10^6 photons per second.

Quantum photon measurements aim is to explore the application of entanglement in some innovative experimental realizations that can be regarded as the research frontier in the field of quantum metrology.

An innovative experiment proposed in [51] consists in a technique for the reconstruction of diagonal elements of density matrix of quantum optical states, or better the statistics reconstruction of photon distribution for free-propagating fields in continuous-wave and in pulsed light beams, for both single-mode semiclassical and quantum states as well as for multimode states, via measurements performed at variable quantum efficiencies of a single on/off avalanche photodetector, assisted by maximum-likelihood estimation and without involving photon counting. The first experimental implementation of the method has been performed for the measurements concerning the photon reconstruction for heralded single-photon Fock states from Type-II SPDC source and for a weak coherent state, as shown in figs. 2 and 3, respectively.

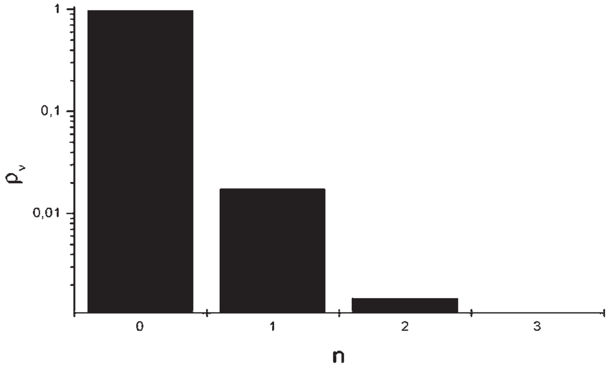


Fig. 3. – Photon distribution for coherent states from a strongly attenuated He-Ne laser.

The relevance of the measurement of statistical distribution of the number of photons provides fundamental information on the nature of optical fields and finds relevant applications in the field of quantum optics.

The realization of efficient sources of bi-photon states is widely recognized as a main resource for quantum metrology [52]. The application of entangled states to radiometry originated by a clever idea of Klyshko [49], and it has been mostly implemented in terms of national metrology institutes from Migdall at NIST [52, 53], with significant contribution from INRIM [54, 55]. The technique is based on the decay of high-energy photons into a non-linear material creating a pair of strictly correlated photons, the presence and properties of one being predicted by those of the other.

The SPDC method can be considered to act like an absolute source or, in other respects, like an absolute detector: for this reason it is preferable to refer to it as a technique. The attractive aspect of its implementation relies on the fact that the method involves only the event counting, independently of other artifacts or standards.

So that the efforts of scientists is focused on the developing techniques for handling measurement issues to transform the philosophy of traditional dissemination of a radiometric quantity, laying aside in this case the distribution of physical reference standards for comparison, and favoring on the other side, the teaching of a technique, to allow the end user to take the maximum advantage of this inherently absolute technique. The present uncertainties estimates for the method is close to 0.2%, but recent works at NPL on the use of laser-based spectrophotometer indicate that it is possible to reach the 0.01% level or lower.

Anyway, even at the current uncertainty level there are great advantages in the use of the method, first for the measurement of very weak signals, *i.e.* at the counting rate regime lower than 10^6 photons/s, where no direct reference standard exists, and finally for the capability, through SPDC to transfer the measurement from a more difficult spectral region, as the infrared one, to the visible, where more efficient detectors are available. The detection operation obviously involves photon-counter rather than analogue devices, because the response to incident radiation is a pulse corresponding to a photon incident on the sensitive area of the device.

3˙3. *Photodetection.* – As photo-detectors suited for an effective discrimination among different numbers of incident photons are not available up to now, measurements are made with single-photon semiconductive detectors, in which the photoelectric effect plays an important role in the detection of light.

As discussed in the previous section, linked to the photoelectric effect there is the concept of quantum efficiency, describing the non-ideal feature of a semiconductive device, and expressing the percentage of photons hitting the photo-active surface that will produce an electron-hole pair.

The response of such a device to the incident radiation is directly a pulse corresponding to one photon incident on the detector sensitive area. In quantum-mechanical terms, the measurement interferes with the measured system. By determining the rate of photon absorption, the photon number reduces by 1, and any successive measurement will find only $n-1$ photons in the photon beam. This is the link between the statistics of the measured quantity and the statistics of the light under measurement.

The photon entangled states, due to their distinctive characteristic of simultaneous creation of conjugated photons, give a unique and powerful means to measure the absolute quantum efficiency of detectors, by performing intrinsically absolute measurements based in principle only upon event counting, which are not tied to any other standard: this led to a well-known technique of coincidence measurement for the calibration at photon counting regime, outgoing the requirements for conventional standards of optical radiation. The process of measuring a light field relies on the absorption of the light impinging on a detector, whether this is a sophisticated electronic detector or the human eye. The end result is that in an ideal measurement, all of the light is absorbed, and on the consequence the measurement destroys the state being measured.

The general properties of ideal photodetection were stated by Glauber [46]. It is not the purpose of this paper to discuss the details of photodetection, therefore a large-scale macroscopic device description will not be considered in this discussion, and anyway a brief description of the general ideas will be illustrated to give the basis to study the quantum statistical correlations of light.

Taking into account that a field operator can be separated into the sum of its positive and negative frequency parts as $\mathbf{E}(\mathbf{r},t) = \mathbf{E}^{(+)}(\mathbf{r},t) + \mathbf{E}^{(-)}(\mathbf{r},t)$, and considering an ideal detector (in a practical photodetection process it is necessary take into account the efficiency of the device), working on an photoelectric effect absorption mechanism, which is sensitive to the field $\mathbf{E}^{(+)}(\mathbf{r},t)$ at the space point $(\mathbf{r},t)$ (only this annihilation operator contributes due to the fact that the measurements are destructive as the photons responsible for producing photoelectrons disappear), the transition probability of the device for absorbing a photon in a time dt is proportional to $\mathcal{T}_{if} = |\langle f|\mathbf{E}^{(+)}(\mathbf{r},t)|i\rangle|^2$, that classically would represent the intensity of the radiation field in the sense of Poynting vector or irradiance in metrological terms, and where $|i\rangle$ and $|f\rangle$ are the initial and final states of the field.

Since the precise knowledge of the field does not usually exist, the statistical description by averaging over all the possible realizations of the initial field can be restored, and it is possible to write

$$\mathcal{T}_{if} = \sum_i P_i \langle f|\mathbf{E}^{(-)}(\mathbf{r},t)\mathbf{E}^{(+)}(\mathbf{r},t)|i\rangle = \mathrm{Tr}\,\varrho\left[\mathbf{E}^{(-)}(\mathbf{r},t)\mathbf{E}^{(+)}(\mathbf{r},t)\right] \equiv W_1(\mathbf{r},t)$$

defining the photon counting rate W_1 (in which the Hermitian operators are put in normal order, with all the destruction operators on the right and all the creation ones on the left). The completeness relation $\sum_f |f\rangle\langle f| = 1$ on the final states has been used, and the density operator for the field $\varrho = \sum_i |i\rangle\langle i|$ has been introduced. The correlation function at the first order of the field is defined by

$$G^{(1)}(\mathbf{r}_1,\mathbf{r}_2;t_1,t_2) = \mathrm{Tr}\left[\varrho\mathbf{E}^{(-)}(\mathbf{r}_1,t_1)\mathbf{E}^{(+)}(\mathbf{r}_2,t_2)\right] \equiv G_1(\mathbf{r}_1,\mathbf{r}_2,\tau),$$

where $\tau = t_1 - t_2$. In terms of $G^{(1)}$ the counting rate therefore results to be $W_1 = G^{(1)}(\mathbf{r}_1,\mathbf{r}_2;0)$.

The description of a photoelectric correlation involving two photodetectors, in which the field at each device results from the superposition of two (or eventually more) light beams at two space-time points $(\mathbf{r}_1,t_1)$ and $(\mathbf{r}_2,t_2)$, with $t_1 \le t_2$, leads to the following counting rate for the intensity:

$$W_2(\mathbf{r}_1,\mathbf{r}_2;t_1,t_2) = \mathrm{Tr}\left[\varrho\mathbf{E}^{(-)}(\mathbf{r}_1,t_1)\mathbf{E}^{(-)}(\mathbf{r}_2,t_2)\mathbf{E}^{(+)}(\mathbf{r}_1,t_1)\mathbf{E}^{(+)}(\mathbf{r}_2,t_2)\right],$$

so that this joint probability of observing a photoionization at each of two photodetectors is governed by the second-order quantum-mechanical correlation, that in the most general case is defined as

$$G^{(2)}(\mathbf{r}_1,\mathbf{r}_2;t_1,t_2) = \langle\mathbf{E}^{(-)}(\mathbf{r}_1,t_1)\mathbf{E}^{(-)}(\mathbf{r}_2,t_2)\mathbf{E}^{(+)}(\mathbf{r}_1,t_1)\mathbf{E}^{(+)}(\mathbf{r}_2,t_2)\rangle$$

that is a normally ordered correlation function of the 4th order, and in general it is possible to define the n-th–order correlation function.

4. – Quantum primary methods

For single-photon detectors [56], classical calibration techniques are based on the use of a strongly attenuated laser source whose (unattenuated) intensity has been measured by means of a power-meter. The uncertainty of this kind of measurement is principally limited by the uncertainty in the calibration of the very low transmittance required for reaching single-photon level.

This limitation has prompted the study of an alternative scheme, based on the use of photons produced by means of spontaneous parametric down-conversion (SPDC), where photons are emitted in pairs strongly correlated in direction, wavelength and polarization.

Furthermore, photons of the same pair are emitted within tens of femtoseconds. Since the observation of a photon of a pair in a certain direction (signal) implies the presence of the other one in the conjugated direction (idler), when this last is not observed this is due to non-ideal quantum efficiency of the idler detector, which can be measured in this way [52, 57-64]. This absolute technique (and others related [65-67]) is becoming attractive for national metrology institutes to realize absolute radiometric standards because it relies simply on the counting of events, involves a remarkably small number of measured quantities, and does not require any reference standards.

Because of the success of the SPDC scheme for calibrating single-photon detectors, we analyze the possibility to extend this technique to higher photon fluxes for calibrating analog detectors.

A seminal attempt in this sense was made in [68] following the theoretical proposal of [59]. Nevertheless, these results were limited to the case of very low photon flux (as we will show in detail later). A systematic analysis of the measurement method for increasing photon flux produced by SPDC is addressed to overcome this limit.

Incidentally, the quantum efficiency η of analog detectors appears also in equations describing suppression of photon noise in parametric down-conversion using the feed-forward [69] or feedback [70, 71] transformations. Hence, such experiments could eventually be used to develop an alternative scheme for analog detector calibration.

Following a theoretical description of SPDC [72, 73], some measurement methods are analyzed for increasing values of the photon flux produced by SPDC, showing how to estimate the quantum efficiency in both counting and analog regimes. Due to the intrinsic limitation of SPDC for calibration when large parametric gains are required, the stimulated parametric down conversion is introduced as the alternative bright source of correlated photons to use, under appropriate conditions, for evaluating quantum efficiency in analog regime.

As the aim is to analyze the SPDC pair production at different flux regimes and its application to the calibration techniques, in the following systematic issues will not be discussed related to the experimental implementation of measurement techniques, such as losses in the idler optical path or electronics and dark noise. These issues have been explored in depth in the low photon flux regime [54, 74, 75] and are not expected to change at higher fluxes for the uncertainty budget determined for SPDC calibration in photon-counting regime (see [75] for numerical estimates). It is also worth mentioning that, analogously to SPDC calibration in the counting regime, various factors as the pump stability or the crystal homogeneity are irrelevant [63].

4·1. *The SPDC scheme for calibrating single-photon detectors.* – The scheme for calibrating single-photon detectors by using SPDC is based on the specific properties of this process, where a photon of the pump beam (usually a laser beam) “decays” inside a non-linear crystal into two lower-frequency photons, 1 and 2 (conventionally dubbed “idler” and “signal”), such that energy and momentum are conserved:

$$\omega_{\text{pump}} = \omega_1 + \omega_2, \qquad \vec{k}_{\text{pump}} = \vec{k}_1 + \vec{k}_2. \tag{8}$$

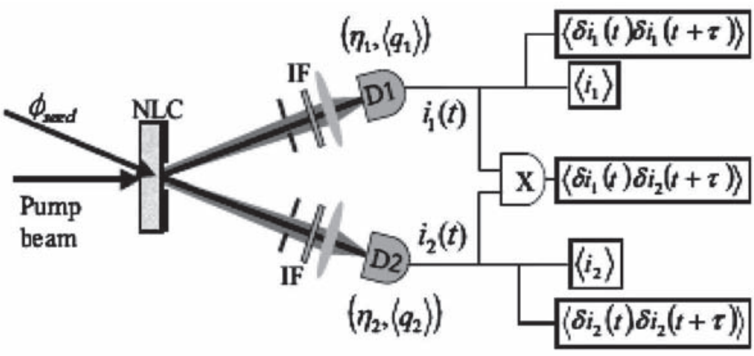

ϕ_{seed}
NLC
Pump beam
IF
IF
$(\eta_1, \langle q_1 \rangle)$
D1
$i_1(t)$
D2
$i_2(t)$
$(\eta_2, \langle q_2 \rangle)$
$\langle \delta i_1(t) \delta i_1(t+\tau) \rangle$
$\langle i_1 \rangle$
X
$\langle \delta i_1(t) \delta i_2(t+\tau) \rangle$
$\langle i_2 \rangle$
$\langle \delta i_2(t) \delta i_2(t+\tau) \rangle$

Fig. 4. – Scheme for calibrating detectors by PDC.

These relations are usually called perfect phase-matching conditions. Moreover, the two photons are emitted within a coherence time τ_{coh} tens of femtoseconds from each other. The process can be spontaneous (SPDC) when no mode of radiation except the pump modes are injected through the input face of the crystal. If a seed mode $\overrightarrow{k}_2$ is injected, its presence stimulates the process and many more photons of the pump are converted. The scheme is schematically depicted in fig. 4.

In essence, the calibration procedure consists [61] of placing a couple of photon-counting detectors, D_1 and D_2 down-stream from the nonlinear crystal, along the direction of propagation of correlated photon pairs for a selected pair of frequencies: the detection of an event by one of the two detectors guarantees with certainty, due to the SPDC properties, the presence of a photon with a fixed wavelength in the conjugated direction. If N is the total number of photon pairs emitted from the crystal in a given time interval T_{gate} and $\langle N_1 \rangle$, $\langle N_2 \rangle$ and $\langle N_c \rangle$ are, respectively, the mean numbers of events recorded during the same time interval T_{gate} by the signal detector, the idler detector, and in coincidence, we have the following obvious relationships [59]:

$$\langle N_1 \rangle = \eta_1 N; \quad \langle N_2 \rangle = \eta_2 N, \tag{9}$$

where η_1 and η_2 are the detection efficiencies in the signal and idler arms. The number of coincidences is

$$\langle N_c \rangle = \eta_1 \eta_2 N, \tag{10}$$

due to the statistical independence of the two detectors. Then the detection efficiency can be found as

$$\eta_1 = \langle N_c \rangle / \langle N_2 \rangle. \tag{11}$$

This simple relation, slightly modified by taking into account the subtraction of background counts and corrections for acquisition dead-time, is the basis for the scheme for the absolute calibration of single-photon detectors by means of SPDC, which reaches now measurement uncertainty competitive with traditional methods [52].

4·2. *Calibration of analog detectors by SPDC correlations.* – In order to study the absolute calibration of analog detectors a PDC model has been developed in [76] working at different values of parametric gain.

In the following a few millimeters non-linear crystal pumped by a CW laser will be considered. If the waist of the pump, identified with the transverse cross-section S_A of the system, is relatively large, namely of the order of a millimeter or more, the real system fits the model of SPDC discussed in [76].

Since the incident photon fluxes $F_1(t)$ and $F_2(t)$ are correlated within $10^{-13}\,\mathrm{s}$, the fluctuation of the registered currents $i_1(t)$ and $i_2(t)$ are supposed to be strictly correlated. The non-ideal quantum efficiency of the detectors makes some photons missed sometimes by D_1 sometimes by D_2, spoiling the correlation. The techniques for estimating the quantum efficiency, both in counting and in analog regime, consist in measuring this effect.

In the following, the photodetection process in the analog regime will be modeled as a random pulse train [77],

$$i(t) = \sum_n q_n f(t - t_n),$$

i.e. a very large number of discrete events at random times of occurrence t_n. The pulse shape $f(t)$ is determined by the transit time of charge carriers. We assume that $f(t)$ is a fixed function with the characteristic width τ_p and a unit area. $\tau_p \sim 10\,\mathrm{ns}$ represents a typical value in analog detection. The pulse amplitude q_n is a random variable in order to account for a possible current gain by avalanche multiplication. The statistical nature of the multiplication process gives an additional contribution to the current fluctuations. In an ideal instantaneous photocell response, without avalanche gain, all values q_n are equal to the charge e of a single electron and $f(t) \sim \delta(t)$.

In the case of ideal quantum efficiency, since the probability density of observing a photon at time t at the detector D_k $(k = 1, 2)$ is related to the quantum mean value of the photon flux $\langle F_k(t)\rangle$, we calculate the average current output of D_k as

$$\langle i_k\rangle = \sum_n \langle q_{kn} f(t - t_{kn})\rangle = \int \mathrm{d}t_k \langle q_k\rangle f(t - t_k)\langle F_k(t_k)\rangle, \tag{12}$$

where the factor $\langle q_k\rangle$ is the average charge produced in a detection event. We have assumed the response function for the two detectors to be the same, $f_1(t) = f_2(t) = f(t)$. Now we introduce the quantum efficiency η_k of detector D_k, defined as the number of pulses generated per incident photon. In [76] a real detector is modeled as an ideal one $(\eta = 1)$ interfaced with a beamsplitter, the transmittance of which equals the quantum efficiency of the real detector. Following the results reported there, $F_k(t_k)$ can be replaced by $\eta_k F_k(t_k)$. As $\langle F_k(t_k)\rangle$ is not dependent on time, one has

$$\langle i_k\rangle = \eta_k\langle q_k\rangle\langle F_k\rangle. \tag{13}$$

The quantum-mechanical second-order intensity correlation function is determined by the probability density to have a photon detected at time t and another one at time t'. Therefore the auto-correlation and cross-correlation functions for the currents can be expressed as

$$\langle i_k(t)i_j(t+\tau)\rangle = \sum_{n,m}\langle q_{kn}q_{jm}f(t-t_{kn})f(t-t_{jm}+\tau)\rangle = \iint \mathrm{d}t_k\mathrm{d}t_j\langle q_kq_j\rangle f(t-t_k)f(t-t_j+\tau)\langle F_k(t_k)F_j(t_j)\rangle, \tag{14}$$

respectively. For $k=j$ the equation defines the electron current autocorrelation function for each detector, one registering the intensity of the signal beam and the other registering the intensity of the idler beam, and $\langle F_k(t_k)F_k(t'_k)\rangle$ is the auto-correlation function of the photon flux at detector D_k. For $k\neq j$ the equation defines the cross-correlation function between the electron currents produced by the two different detectors, and $\langle F_k(t_k)F_j(t_j)\rangle$ is the cross-correlation between the photon fluxes incident at the two different detectors.

According to eq. (12), the mean value of the electron current (the analog of eq. (9) in photon-counting regime) is equal to

$$\langle i_1\rangle = \eta_1\langle q_1\rangle\langle F_1\rangle\,. \tag{15}$$

The factor $\langle q_1\rangle$ is the average charge produced in a detection event.

As shown in [76] the current correlation functions have, roughly, a sinc-like behavior in time, with the central peak width equal to the coherence time of PDC $\tau_{\mathrm{coh}} = 1/\Omega_0 \sim 10^{-13}\,\mathrm{s}$. As the resolving time of a real analog detector is finite and, in general, much larger than the SPDC coherence time, any fluctuations in the intensity of light are integrated over τ_p during the detection process. So in the limit $\tau_p \gg \tau_{\mathrm{coh}}$, and for $k=1$ and $j=2$, one has

$$\langle i_1(t)i_1(t+\tau)\rangle = \langle i_1\rangle^2 + \eta_1\langle q_1^2\rangle\mathcal{F}(\tau)\cdot[\langle F_1\rangle + \eta_1\vartheta] \tag{16}$$

and (the analog of eq. (10) in the photon-counting regime)

$$\langle i_1(t)i_2(t+\tau)\rangle = \langle i_1\rangle\langle i_2\rangle + \eta_1\eta_2\langle q_1\rangle\langle q_2\rangle\mathcal{F}(\tau)\cdot[\langle F_1\rangle + \vartheta]\,, \tag{17}$$

where $\mathcal{F}(\tau)\equiv\int dt f(t)f(t+\tau)$ is the convolution of the response function of the detectors. The term ϑ depends on the second power of the parametric gain V, *i.e.* on the mean number of photons per mode of the radiation beam. For the purpose of this review, $\vartheta \simeq V\langle F\rangle$.

The last two equations are the fundamental tools for studying the problem of absolute calibration of analog detectors.

The presence of $\langle F_1 \rangle$ in the autocorrelation function is due to the shot noise contribution and for this reason the quantum efficiency η enters linearly, while in the current cross-correlation function the corresponding term is due to the high quantum correlation between the signal and idler beams of PDC and the quantum efficiency appears quadratically. It is equivalent to the right-hand side of eq. (10) for the counting regime and its presence is the key for absolute calibration.

The term ϑ is important for both auto- and cross-correlation functions only when the number of photons per coherence time is not close to zero and the presence of two or more photons within that time is not negligible. Thus, it can be neglected as long as $V \ll 1$, *i.e.* the mean number of photons per coherence volume is much smaller than one. However, if the duration of the photocurrent pulse is much larger than the coherence time, this assumption does not prevent photodetection being in a strongly analog regime, because a lot of photons can be absorbed during the pulse duration generating the overlapping of pulses.

The term proportional to $\langle i \rangle^2 \propto \langle F \rangle^2$ is due to the presence of more than one photon within a time interval τ_p. For that reason it can be neglected only if $\langle F \rangle \ll \mathcal{F}(\tau)$. Since the pulse $f(t)$ has a height around $1/\tau_p$, $\max[\mathcal{F}(\tau)] = \mathcal{F}(0) \sim 1/\tau_p$. So the condition becomes $\langle F \rangle \tau_p \ll 1$, *i.e.* the number of incident photons during the resolving time of the detector should be much less than one, *i.e.* one should work in a non-overlapping regime. In principle, in this case one could distinguish between different pulses of the current and work in the counting mode.

The usual definition of the quantum efficiency is the ratio between the number of photons detected and the number of photons incident on the detector surface. This definition is completely suitable in the case of counting detectors and is exactly the meaning of η in this paper. But in the case of analog detection, we cannot in principle distinguish between different current pulses. Thus, according to formula (15), we adopt the definition of analog quantum efficiency as $\Gamma \equiv \eta \langle q \rangle = \langle i \rangle / \langle F \rangle$, having the meaning of the electron charge produced per single incident photon, or the ratio between the electron flux and the photon flux. If the charge q produced per photon fluctuates, it increases the current fluctuations. This explains why $\langle q_1^2 \rangle$ appears in eq. (16) instead of $\langle q_1 \rangle^2$. In principle, the most general way to obtain an estimation of η working with the PDC light intensity in the photon-counting regime is dividing the coincidence counting rate (proportional to the cross-correlation function) by the detector counting rate (proportional to the intensity). This method works because in the photon-counting regime the temporal shape of the current pulses and their width is not important; instead, one either registers a single pulse or does not register. This is not the case for analog detection in which the pulse shape $f(t)$ appears in formulas through the factor $\mathcal{F}(\tau)$. In general, we do not know this function, and this makes the absolute calibration of analog detectors more difficult. However, as we are going to show, under some condition it is possible to overcome this drawback.

Let us distinguish between three different regimes: very low intensity (I), middle intensity (II), high intensity (III).

(I) $\langle F\rangle\tau_p \ll 1$ (*i.e.* photocurrent pulses do not overlap on the average). For a detector with a time constant $\tau_p = 10\,\text{ns}$, the corresponding photon flux must be below 10^8 photons/s, which, for a wavelength of 500 nm, means a power of tens of pW.

By neglecting ϑ and terms in $\langle i\rangle^2$, eqs. (16) and (17) become then

$$\langle i_1(t)i_1(t+\tau)\rangle = \eta_1\langle q_1^2\rangle\mathcal{F}(\tau)\langle F_1\rangle, \tag{18}$$

$$\langle i_1(t)i_2(t+\tau)\rangle = \eta_1\eta_2\langle q_1\rangle\langle q_2\rangle\mathcal{F}(\tau)\langle F_1\rangle. \tag{19}$$

The same equations has been found in [59] and the quantum efficiency has been estimated as

$$\Gamma_2 = \eta_2\langle q_2\rangle = \frac{\langle q_1^2\rangle}{\langle q_1\rangle^2}\langle q_1\rangle\frac{\langle i_1(t)i_2(t+\tau)\rangle}{\langle i_1(t)i_1(t+\tau)\rangle}\,. \tag{20}$$

This formula is not satisfying for metrology because of the presence of an unknown parameter related to the statistics of charge fluctuations that has to be estimated in some other way. The problem is avoided by integrating eq. (19) over time τ. This can be done after the acquisition of the profile of the function has been performed. We would like to stress that in this case, the assumption $f_1(t) = f_2(t) = f(t)$ is not necessary. Since $\int \mathrm{d}\tau\mathcal{F}(\tau) = 1$, integrating eq. (19) in τ and dividing it by eq. (15), we obtain

$$\Gamma_2 = \eta_2\langle q_2\rangle = \frac{\int \mathrm{d}\tau\langle i_1(t)i_2(t+\tau)\rangle}{\langle i_1\rangle}\,. \tag{21}$$

As pointed out, another drawback of (20) is that it works only at very low intensity, where no overlapping between pulses happens. In principle, one could distinguish between different pulses and work in the counting mode provided the amplitude q_n of each pulse is large enough to be detected. This, in terms of experiment, means that one has to work in the so-called "charge accumulation mode", in which the electron charge is accumulated until reaching some detectable threshold. In the case of avalanche devices, it corresponds to the regime of direct photon-counting.

(II) $\langle F\rangle\tau_p \geq 1$ but still $V \ll 1$ (*i.e.* photocurrent pulses overlap but the parametric gain and photon flux are still quite low). By considering the same parameters as used in case I, coherence time τ_{coh} of the order of 100 fs, and the requirement that $V \leq 0.001$, this means a photon flux up to 10^{10} photons/s, *i.e.* a power of few nW. Here only the rem ϑ can be neglected and eqs. (16) and (17) become

$$\langle i_1(t)i_1(t+\tau)\rangle = \langle i_1\rangle^2 + \eta_1\langle q_1^2\rangle\mathcal{F}(\tau)\langle F_1\rangle, \tag{22}$$

$$\langle i_1(t)i_2(t+\tau)\rangle = \langle i_1\rangle\langle i_2\rangle + \eta_1\eta_2\langle q_1\rangle\langle q_2\rangle\mathcal{F}(\tau)\langle F_1\rangle. \tag{23}$$

By defining the correlation functions of the current fluctuations as

$$\begin{aligned}\langle\delta i_k(t)\delta i_l(t+\tau)\rangle \equiv{}& \langle i_k(t)i_l(t+\tau)\rangle -\\ &-\langle i_k(t)\rangle\langle i_l(t+\tau)\rangle \quad (k,l=1,2)\,,\end{aligned} \tag{24}$$

a relationship similar to (20) is obtained for analog quantum efficiency estimation:

$$\eta_2\langle q_2\rangle = \frac{\langle q_1^2\rangle}{\langle q_1\rangle^2}\langle q_1\rangle \frac{\langle \delta i_1(t)\delta i_2(t+\tau)\rangle}{\langle \delta i_1(t)\delta i_1(t+\tau)\rangle}\,. \tag{25}$$

Once again, the drawback of this formula is the presence of the unknown parameter $M = \langle q_1^2\rangle/\langle q_1\rangle^2$, related to the statistics of charge fluctuations and that requires additional measurements to be performed. As before, this problem can be overcome by integrating over τ the expression for the cross-correlation, *i.e.* the definition (24) for $k = 1, j = 2$. This corresponds to evaluate the power spectrum of the fluctuations at frequencies much smaller than $1/\tau_p$. In this way, one has

$$\eta_2\langle q_2\rangle = \frac{\int \mathrm{d}\tau \langle \delta i_1(t)\delta i_2(t+\tau)\rangle}{\langle i_1\rangle}\,. \tag{26}$$

Here, $\eta_2\langle q_2\rangle$ is the electron charge per single incident photon, or, the ratio between the current and the photon flux.

This equation shows that the absolute calibration of analog detectors by using SPDC is indeed possible.

(III) $V \geq 1$ (*i.e.* high-intensity regime).

In this regime each term of (16) and (17) is important and no simple general way can be found for the absolute calibration of analog detectors, at least with a CW pump.

It can be shown that in this case one should be able to collect exactly the same number of correlated modes by D_1 and D_2. Since SPDC takes place with a very large spectral and spatial bandwidth, it would require accurate and well-balanced spatial and spectral selection. This could originate systematic errors difficult to be evaluated.

4·3. *Calibration of analog detectors by stimulated SPDC.*

– Calibration for photon fluxes larger than 10^{10} photons/s can be addressed by using a parametric amplifier configuration, where a seed coherent beam of power Φ is injected along one direction to stimulate the emission of the two correlated beams. The photon fluxes can be increased by varying the power Φ of the seed beam without increasing V, *i.e.* going back to the condition $V \ll 1$.

By introducing proper expressions for the fluxes in eq. (13), when the spontaneous emission is negligible, one has

$$\langle i_1\rangle_s = \eta_1\langle q_1\rangle V\Phi\,, \tag{27}$$

$$\langle i_2\rangle_s = \eta_2\langle q_2\rangle (1+V)\Phi\,, \tag{28}$$

where the subscript s reminds that currents are calculated for the stimulated SPDC. Using eq. (14) the correlation function of the current fluctuations can be expressed in a simple form, under the assumption $V \ll 1$, and the quantum efficiency can be evaluated as

$$\eta_2\langle q_2\rangle = \frac{1}{2}\frac{\langle q_1^2\rangle}{\langle q_1\rangle^2}\langle q_1\rangle\frac{\langle\delta i_1(t)\delta i_2(t+\tau)\rangle_s}{\langle\delta i_1(t)\delta i_1(t+\tau)\rangle_s}\,. \tag{29}$$

Alternatively, by integrating over τ the cross-correlation, one has

$$\eta_2\langle q_2\rangle = \frac{1}{2}\frac{\int \mathrm{d}\tau\langle\delta i_1(t)\delta i_2(t+\tau)\rangle_s}{\langle i_1\rangle_s}\,, \tag{30}$$

where the avalanche gain factor disappears.

Using the stimulated PDC has two main advantages. Firstly, there is no upper limit in the photon fluxes. For detectors without internal gain, where the Johnson noise is predominant, good signal-to-noise ratios can be obtained by increasing the signal. Secondly, the stimulated emission has a very narrow bandwidth, both in frequency and in space, and the correlated beams are easier to collect.

The expected uncertainty in evaluating the correlation functions depends on the measurement technique. Here the two current outputs by d_1 and D_2 are combined by an analog multiplier and, then, integrated for a time T much larger than the response tome τ_p. In the approximation of small fluctuations, *i.e.*, $\delta_i(t) \ll \langle i\rangle$, the maximal uncertainty for both auto- and cross-correlation is given by $\Delta \approx \eta^2\langle q\rangle^2\langle F\rangle^{3/2}T^{-1/2}$. For a detector without avalanche gain, the relative uncertainty on the quantum efficiency results to be $\Delta\eta/\eta \approx \langle N_{\tau_p}\rangle^{1/2}(\tau_p/T)^{1/2}$, where N_{τ_p} is the number of photons detected during the detector response time. The uncertainty scles with the square root of the measurement time T. For $T = 1\,\mathrm{s}$, it is lower than one part in 10^{-3}.

The noise contributions from the detector (dark current, noise of the transimpedance amplifier) and background light are supposed to be statistically independent of each other and of the photocurrents produced by SPDC light in the two detectors. Thus, they do not give any contribution to the cross correlation function of the two current fluctuations in the numerator of (25). In the denominator of the same equation, containing the autocorrelation of the current fluctuations, their effect should be measured (for example, rotating by 90^o the polarization of the laser pump, to eliminate the SPDC signal), and subtracted. The statistical uncertainty of this measurement can be kept so small that it does not increase significantly the total relative uncertainty given above.

Concerning the systematic contributions, we do not expect significant changes with respect to the photon-counting regime. They are basically due to the optical losses and have been estimated in some part in 10^{-3} [74-76]. Thus, they should be the dominant contribution to the uncertainty budget, largely exceeding the statistical one.

REFERENCES

[1] Quinn T. J., *Metrologia*, **31** (1994/1995) 515.
[2] Quinn T. J., *Metrologia*, **34** (1997) 61.

[3] Kose V., Siebert B. R. and Wöger W., *Metrologia*, **40** (2003) 146.
[4] Fox N. P., *Metrologia*, **32** (1996) 535.
[5] Fox N. P., *Metrologia*, **37** (2000) 507.
[6] Sapritsky V. I., *Metrologia*, **32** (1995/1996) 441.
[7] Zalewski E. F. and Geist J., *Appl. Opt.*, **19** (1980) 1214.
[8] Zalewski E. F. and Duda C. R., *Appl. Opt.*, **22** (1983) 2867.
[9] Jellison G. E. jr., *Opt. Mater.*, **1** (1992) 41.
[10] Geist J., Migdall A. and Baltes H., *Appl. Opt.*, **27** (1998) 3777.
[11] Kärhä P., Lassila A., Ludvigsen H., Manoochehri F., Fagerlund H. and Ikonen E., *Opt. Eng.*, **34** (1995) 2611.
[12] Houston J. M. and Zalewski E. F., *Optical Radiation Measurements II*, *SPIE*, Vol. **1109** (1989) 268.
[13] Hughes C. G., *Appl. Opt.*, **21** (1982) 2129.
[14] Geist J., Farmer J. D., Martin P. J., Wilkinson F. J. and Collocott S. J., *Appl. Opt.*, **21** (1982) 1130.
[15] Booker R. L. and Geist J., *Appl. Opt.*, **23** (1984) 1940.
[16] Geist J., *Optical Radiation Measurements II*, **1109** (1989) 246.
[17] Geist J., Zalewski E. F. and Schaefer A. R., *Appl. Opt.*, **19** (1980) 3795.
[18] Key P. J., Fox N. P. and Rastello M. L., *Metrologia*, **21** (1985) 81.
[19] Geist J., Gladden W. K. and Zalewski E. F., *J. Opt. Soc. Am.*, **72** (1982) 1068.
[20] Korde R. and Geist J., *Appl. Opt.*, **26** (1987) 5284.
[21] Hoyt C., Miller P. J., Foukal P. and Zalewski E. F., *Optical Radiation Measurements II*, **1109** (1989) 236.
[22] Geist J., *Appl. Opt.*, **18** (1979) 760.
[23] Gentile T. R., Houston J. M. and Cromer C. L., *Appl. Opt.*, **35** (1996) 4392.
[24] Werner L., Fischer J., Johannsen U. and Hartmann J., *Metrologia*, **37** (2000) 279.
[25] Kübarsepp T., Kärhä P. and Ikonen E., *Appl. Opt.*, **39** (2000) 9.
[26] Campos J., Pons A. and Corredera P., *Metrologia*, **40** (2003) S181.
[27] Gran J. and Sudbo A., *Metrologia*, **41** (2004) 204.
[28] Gran J., *Accurate and independent spectral response scale based on silicon trap detectors and spectrally invariant detectors*, in *Series of dissertation* (Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo) 2005, pp. 16-20.
[29] Zalewski E. F. and Hoyt C. C., *Metrologia*, **28** (1991) 203.
[30] Stock K. D., Hofer H., White M. and Fox N. P., *Metrologia*, **37** (2000) 437.
[31] Geist J., Brida G. and Rastello M. L., *Metrologia*, **40** (2003) S132.
[32] Basore P. A., Rover D. T., Thorson G. M., Smith A. W. and Hansen B. R., *PC-1D, Version 2.01*, ISU Extension Software (Iowa State University, Ames, IA) 1998, pp. 294-8658.
[33] Malitson I. H., *J. Opt. Soc. Am.*, **55** (1965) 1205.
[34] Geist J., *Silicon (Si) Revisited*, in *Handbook of Optical Constants of Solids III*, edited by Palik E. W. (Academic Press, Boston) 1998, p. 519.
[35] Auslender M. and Hava S., *Doped n-type Silicon (n-Si)*, in *Handbook of Optical Constants of Solids III*, edited by Palik E. W. (Academic Press, Boston) 1998, p. 155-173.
[36] Geist J., *Planar Silicon Photosensors*, in *Sensor Technology and Devices*, edited by Ristic L. J. (Artech House, Norwood, MA) 1994, p. 317.
[37] Sze S. M., *Physics of Semiconductor Devices, second edition* (John Wiley Sons, NY) 1981, pp. 47-49.
[38] Koehler R., Luther J. L. and Geist J., *Appl. Opt.*, **29** (1990) 3130.
[39] Bloch F., *Phys. Rev. B*, **2** (1970) 109.
[40] Taylor B. N., *Phys. Letters*, **153** (1991) 308.

[41] Taylor B. N. and Witt T. J., *Metrologia*, **26** (1986) 47.

[42] Martin J. E. and Haycocks P. R., *Metrologia*, **35** (1998) 229.

[43] Planck M., *Verhr. Deutsch. Phys. Ges.*, **2** (1900) 202.

[44] Einstein A., *Ann. Phys.*, **17** (1905) 132.

[45] Dirac P. A. M, *Proc. Roy. Soc.*, **114** (1927) 243.

[46] Glauber R. J., *Phys. Rev.*, **130** (1963) 2529.

[47] Mandel L., *Phys. Rev.*, **131** (1963) 2766.

[48] Bell J. S., *Phys. World*, **3** (1990) 33.

[49] Klyshko D. N., *Sov. Phys. Usp.*, **31** (1988) 74.

[50] Ware M. and Migdall A., *J. Mod. Opt.*, **51** (2004) 1549.

[51] Genovese M., Brida G., Gramegna M., Bondani M., Zambra G., Andreoni A., Rossi A. and Paris M. G., *Laser Phys.*, **16** (2006) 1.

[52] Migdall A., *Phys. Today*, **52** (1999) 41.

[53] Migdall A. L., Datla R. U., Sergienko A., Orszak J. S. and Shih Y. H., *Metrologia*, **32** (1995) 479.

[54] Brida G., Degiovanni I. P., Novero C. and Rastello M. L., *Metrologia*, **37** (2000) 625.

[55] Castelletto S., Degiovanni I. P. and Rastello M. L., *Metrologia*, **37** (2000) 613.

[56] Ingerson T. E., Kearney R. G. and Coulter R. L., *Appl. Opt*, **22** (1983) 2013.

[57] Zel'dovich B. Y. and Klyshko D. N., *Sov. Phys. JETP Lett.*, **9** (1969) 69.

[58] Burnham D. C. and Weinberg D. L., *Phys. Rev. Lett.*, **25** (1970) 84.

[59] Klyshko D. N., *Sov. J. Quantum Electron*, **10** (1980) 1112.

[60] Malygin A., Penin A. N. and Sergienko A. V., *Sov. Phys. JETP Lett.*, **33** (1981) 477.

[61] Klyshko D. N. and Penin A. N., *Sov. Phys. Usp.*, **30** (1987) 716.

[62] Brida G., Genovese M. and Novero C., *J. Mod. Opt.*, **47** (2000) 2099.

[63] Brida G., Genovese M. and Gramegna M., *Laser Phys. Lett.*, **3** (2006) 115.

[64] Ginzburg V. M., Keratishvili N. G., Korzhenevich E. L., Lunev G. V., Penin A. N. and Sapritsky V. I., *Opt. Eng.*, **32** (1993) 2911.

[65] Brida G., Chekhova M. V., Genovese M., Gramegna M., Krivitsky L. A. and Kulik S. P., *Phys. Rev. A*, **70** (2004) 032332.

[66] Brida G., Chekhova M., Genovese M., Gramegna M., Krivitsky L. A. and Rastello M. L., *Journ. Opt. Soc. Am. B*, **22** (2005) 488.

[67] Brida G., Chekhova M., Genovese M., Gramegna M., Krivitsky L. A. and Rastello M. L., *IEEE Trans. IM*, **54** (2005) 898.

[68] Sergienko A. V. and Penin A. N., *Sov. Tech. Phys. Lett.*, **12** (1986) 328.

[69] Mertz J., Heidmann A. and Fabre C., *Phys. Rev. A*, **44** (1991) 3229.

[70] Shapiro J. H., Saplakoglu G., Ho S. T., Kumar P., Saleh B. E. and Teich M. C., *J. Opt. Soc. Am. B*, **4** (1987) 1604.

[71] Tapster P. R., Rarity J. G. and Satchell J. S., *Phys. Rev. A*, **37** (1988) 2963.

[72] Brambilla E., Gatti A., Bache M. and Lugiato L. A., *Phys. Rev. A*, **69** (2004) 023802.

[73] Brambilla E., Gatti A., Lugiato L. A. and Kolobov M. I., *Eur. Phys. J. D*, **15** (2001) 127.

[74] Dauler E., Migdall A., Boeuf N., Dalta R. U., Mullerand A. and Sergienko A., *Metrologia*, **35** (1998) 295.

[75] Brida G., Castelletto S., Degiovanni I. P., Novero C. and Rastello M. L., *Metrologia*, **37** (2000) 629.

[76] Brida G., Chekhova M., Genovese M., Penin A. and Ruo Berchera I., *J. Opt. Soc. Am. A*, **23** (2006) 2185.

[77] Soda K., Nishio I. and Wada A., *J. Appl. Phys.*, **47** (1976) 729.

This page intentionally left blank

International School of Physics “Enrico Fermi”

Villa Monastero, Varenna

Course CLXVI

18–28 July 2006

“Metrology and Fundamental Constants”

Directors

Theodore W. HÄNSCH
Max-Planck-Institut für Quantenoptik
Hans-Kopfermann-strasse 1
85748 Garching
Germany
Tel.: ++49 89 32905712
Fax: ++49 89 32905312
rosemarie.lechner@mpq.mpg.de

Sigfrido LESCHIUTTA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919306
Fax: ++39 011 346384
siglesc@libero.it

Andrew J. WALLARD
BIPM
Bureau International des Poids et Mesures
Pavillon de Breteuil
92312 Sèvres Cedex
France
Tel.: ++33 1 45077001
Fax: ++33 1 45342021
f.joly@bipm.org

Scientific Secretary

Maria Luisa RASTELLO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919219
Fax: ++39 011 346384
rastello@inrim.it

Lecturers

Hans BACHMAIR
Physikalisch-Technische Bundesanstalt
Electricity
Abteilung 2
Bundesallee 100
38116 Braunschweig
Germany
Tel.: ++49 531 5922010
Fax: ++49 531 5922015
Hans.Bachmair@ptb.de

Andreas BAUCH
Physikalisch-Technische Bundesanstalt
WG Dissemination of Time
Arbeitsgruppe 4.42
Bundesallee 100
38116 Braunschweig
Germany
Tel.: ++49 531 5924320
Fax: ++49 531 5924479
Andreas.bauch@ptb.de

Kim CARNEIRO
Danish Ltd. of Fundamental Metrology
Building 307
Matematiktorvet
DK-2800 Lyngby
Denmark
Tel.: ++45 45 255867
Fax: ++45 45 931137
kc@dfm.dtu.dk

Joachim FISCHER
Physikalisch-Technische Bundesanstalt
Temperature
Fachbereich 7.4
Abbestr. 2-12
10587 Berlin
Germany
Tel.: ++49 030 34817473
Fax: ++49 030 34817508
joachim.fischer@ptb.de

Nigel FOX
National Physical Laboratory
Quality of Life
Building 64, Hampton Road
Teddington, Middlesex TW11 0LW
UK
Tel.: ++44 20 89436825
Fax: ++44 20 89436458
nigel.fox@npl.co.uk

Beat JECKELMANN
Swiss Federal Office of Metrology
and Accreditation (METAS)
Lindenweg 50
3003 Bern-Wabern
Switzerland
Tel.: ++41 31 3233201
Fax: ++41 31 3233579
beat.jeckelmann@eam.admin.ch

Pierre LEMONDE
Observatoire de Paris
SYRTE - Bâtiment B
61 avenue de l'Observatoire
75014 Paris
France
Tel.: ++33 1 40512324
Fax: ++33 1 43255542
pierre.lemonde@obspm.fr

Martin MILTON
National Physical Laboratory
Quality of Life
Building 64, Hampton Road
Teddington, Middlesex TW11 0LW
UK
Tel.: ++44 20 89436826
Fax: ++44 20 89436458
martin.milton@npl.co.uk

William D. PHILLIPS
National Institute of Standards
and Technology
100 Bureau Drive, Stop 8424
Gaithersburg, MD 20899-8424
USA
Tel.: ++1 301 9756554
william.phillips@nist.gov

François PIQUEMAL
LNE
ZA de Trappes-Élancourt
29, avenue Roger Hennequin
78197 Trappes Cedex
France
Tel.: ++33 1 30692173
Fax: ++33 1 30162841
francois.piquemal@lne.fr

Terry QUINN
Former director of Bureau International des Poids et Mesures
Rue Brancas 92
92310 Sèvres
France
Tel.: ++33 1 46230656
terry.quinn@physics.org,
tquinn@bipm.org

Philippe RICHARD
Swiss Federal Office of Metrology and Accreditation (METAS)
Lindenweg 50
CH-3003 Bern-Wabern
Switzerland
Tel.: ++41 31 3233201
Fax: ++41 31 3233579
philippe.richard@eam.admin.ch

Richard RUSBY
National Physical Laboratory
Engineering and Process Control
Building 95, Hampton Road
Teddington, Middlesex TW11 0LW
UK
Tel.: ++44 20 89437036
Fax: ++44 20 89436458
richard.rusby@npl.co.uk

Thomas UDEM
Department of Laserspectroscopy
Max-Planck-Institut für Quantenoptik
Hans-Kopfermann-Str., 1
Garching 85758
Germany
Tel.: ++49 89 32905282
Thomas.udem@mpq.mpg.de

Seminar Speakers

Elisa Felicitas ARIAS
Bureau International des Poids et Mesures
Time and Frequency Section
Pavillon de Breteuil
F-92312 Sèvres Cedex
France
Tel.: ++33 1 45077076
Fax: ++33 1 45342021
farias@bipm.org

Walter BICH
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977445
Fax: ++39 011 3977437
w.bich@imgc.cnr.it,
w.bich@inrim.it

Marcello CORADINI
European Space Agency HQ
Coordinator, Solar System Missions
ESA Science Directorate
8-10, rue Mario Nikis
75738 Paris Cedex
France
Tel.: ++33 1 53697555
Fax: ++33 1 53697751
marcello.coradini@esa.int

Richard DAVIS
Bureau International des Poids et Mesures
Mass Section
Pavillon de Breteuil
92312 Sèvres Cedex
France
Tel.: ++33 1 45077011
Fax: ++33 1 45342021
rdavis@bipm.org

Roberto M. GAVIOSO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919616
Fax: ++39 011 346384
gavioso@ien.it

Giovanni MANA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977421
Fax: ++39 011 3977426
g.mana@imgc.cnr.it

Students

Paola AMERIO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977471
Fax: ++39 011 3977459
p.amerio@inrim.it

Xavier BAILLARD
SYRTE, Observatoire de Paris
61, avenue de l'Observatoire
75014 Paris
France
Tel.: ++33 1 40512074
Fax: ++33 1 43255542
xavier.baillard@obspm.fr

Deborah BAINES
NPL
Hampton Road, Bldg 5, room 105
TW11 0LW Teddington, Middlesex
Gran Bretagna
Tel.: ++44 020 89436462
Fax: ++44 020 89436529
deborah.baines@npl.co.uk

Zeb BARBER
NIST
325 S. Broadway, Mail Stop 847.00
80305 Boulder (CO)
USA
Tel.: ++1 303 4974112
Fax: ++1 303 4972845
zbarber@boulder.nist.gov

Mark BART
Temperature Standards
New Zealand Measurement Standards
Laboratory
Industrial Research Limited
PO Box 31-310
Lower Hutt
New Zealand
Tel.: ++64 4 9313440
m.bart@irl.cri.nz

Andrea BERNARDI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919352
Fax: ++39 011 3919621
bernardi@inrim.it

Chiara BOVERI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919432
Fax: ++39 011 346384
boveri@inrim.it

Vincenzo CACCIATORE
Dipartimento di Elettronica
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Italy
Tel.: ++39 011 5644110
vincenzo.cacciatore@polito.it

Giovanni CASA
Dipartimento di Scienze Ambientali
Seconda Università di Napoli
Via Vivaldi 43
81100 Caserta
Italy
Tel.: ++39 0823 274628
Fax: ++39 0823 274605
giovanni.casa@unina2.it

Giancarlo CERRETTO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919258
Fax: ++39 011 3919259
cerretto@inrim.it

Frédéric CHAPELET
LNE-SYRTE, Observatoire de Paris
61, avenue de l'Observatoire
75014 Paris
France
Tel.: ++33 1 40512097
Fax: ++33 1 43255542
frederic.chapelet@obspm.fr

Nicola COLUCCELLI
Dipartimento di Fisica
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano
Italy
Tel.: ++39 02 23996160/6161
Fax: ++39 02 23996126
nicola.coluccelli@polimi.it

Christophe CONSEJO
LNE
ZA de Trappes-Élancourt
29 avenue Roger Hennequin
78197 Trappes Cedex
France
Tel.: ++33 1 30692177
Fax: ++33 1 30162841
christophe.consejo@lne.fr

Giancarlo D'AGOSTINO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977419
Fax: ++39 011 3977426
gc.dagostino@inrim.it

Natascia DE LEO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919410
Fax: ++39 011 346384
deleo@inrim.it

Sophie DJORDJEVIC
LNE
ZA de Trappes-Élancourt
29 avenue Roger Hennequin
78197 Trappes Cedex
France
Tel.: ++33 1 30692157
Fax: ++33 1 30162841
sophie.djordjevic@lne.fr

Florin GAROI
National Institute for Laser, Plasma
and Radiation Physics
Atomistilor 409, PO Box MG-36
Magurele 077125, Ilfov
Romania
Tel.: ++40 21 4574243
Fax: ++40 21 4574243
florin.garoi@inflpr.ro

Davide GATTI
Dipartimento di Fisica
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano
Italy
Tel.: ++39 02 23996161/6098
Fax: ++39 02 23996126
d.gatti@polimi.it

Salvatore GIUNTA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977347
Fax: ++39 011 3977348
s.giunta@inrim.it

Felicia GREEN
NPL
Hampton Road
Teddington, Middlesex TW11 0LM
UK
Tel.: ++44 020 89436153
Fax: ++44 020 89436453
felicia.green@npl.co.uk

Christian HOF
Swiss Federal Office of Metrology
and Accreditation (METAS)
Lindenweg 50
3003 Bern-Wabern
Switzerland
Tel.: ++41 31 3234642
Fax: ++41 31 3233210
christian.hof@metas.ch

Riccardo INTROZZI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919303/227
Fax: ++39 011 346384
introzzi@inrim.it

Iuliana Mariana IORDACHE
National Institute for Laser, Plasma
and Radiation Physics
Atomistilor 409, PO Box MG-36
077125 Magurele, Ilfov
Romania
Tel.: ++40 21 4574243
Fax: ++40 21 4574243
iuliana.iordache@inflpr.ro

Aleksandr KRAVCHENKO
Research and Educational Center
"PLASMA"
Office 225
Lenina av. 33
185910 Petrozavodsk
Karelia Republic
Russia
Tel.: ++7 814 2782693
Fax: ++7 814 2711000
aleksandr@sampo.ru

Aleksandar KRMPOT
Institute of Physics
Pregrevica 118
11080 Belgrado
Serbia
Tel.: ++381 11 3160793
Fax: ++381 11 3162190
krmpot@phy.bg.ac.yu

Andrea MALENGO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977447
Fax: ++39 011 3977437
a.malengo@inrim.it

Nicola MALOSSI
Niels Bohr Institute
University of Copenhagen
Universitetsparken 5
2100 Kobenhaven
Denmark
Tel.: ++45 35 320504
Fax: ++45 35 320460
onimoon@fys.ku.dk

Domenico MARI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977359
Fax: ++39 011 3977426
d.mari@inrim.it

Alice MEDA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 338 9081782
a.meda@inrim.it

Anastasia MERISTOUDI
National Hellenic Research Foundation
Institute of Theoretical and Physical
Chemistry
48 Vassileos Costantinou avenue
11635 Athens
Greece
Tel.: ++30 697 7419187
Fax: ++30 210 7273737
amerist@eie.gr

Piercarlo MIGLIETTA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919626
Fax: ++39 011 3919621
piercarlo.miglietta@polito.it,
piercarlo.miglietta@fiat.com

Alberto MURA
Dipartimento di Elettronica
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Italy
Tel.: ++39 011 5644125
Fax: ++39 011 5644099
alberto.mura@polito.it

Luca OBERTO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919321
Fax: ++39 011 3919259
oberto@inrim.it

Francesca PENNECCHI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977445
Fax: ++39 011 3977437
f.pennecchi@inrim.it

Sergio PERERO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919822
sergio.perero@polito.it

Aline PICCATO
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977444
Fax: ++39 011 3977437
a.piccato@inrim.it

Giuseppe PISTONE
Dipartimento di Fisica della Materia
e Tecnologie Fisiche Avanzate
Università di Messina
Salita Sperone, 31
98166 Messina
Italy
Tel.: ++39 090 6765 454
Fax: ++39 090 391382
pistone@unime.it

Emiliano PUDDU
Dipartimento Fisica e Matematica
Università dell'Insubria
Via Valleggio, 11
22100 Como
Italy
Tel.: ++39 031 2386253
Fax: ++39 031 2386119
puddu@uninsubria.it

Frédéric PYTHOUD
Swiss Federal Office of Metrology
and Accreditation (METAS)
Lindenweg 50
3003 Bern-Wabern
Switzerland
Tel.: ++41 31 3233335
Fax: ++41 31 3233210
frederic.pythoud@metas.ch

Danilo QUAGLIOTTI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 73
10135 Torino
Italy
Tel.: ++39 011 3977427
Fax: ++39 011 3977426
d.quagliotti@inrim.it

Maryness SALAZAR
National Metrology Laboratory
Industrial Technology Development
Institute
DOST Compound, General Santos Avenue
Bicutan, Taguig City,
Metro Manila
Filippine
Tel.: ++632 837 2071 ext. 2264
Fax: ++632 837 6150
nhet28@yahoo.com,
misalazar@dost.gov.ph

Edcel John SALUMBIDES
Laser Centre Vrije Unversiteit
Dr. Boelelaan 1081
1081 HV Amsterdam
The Netherlands
Tel.: +31 20 598 7944
Fax: +31 20 598 7999
ejsalum@Nationalvu.nl

Valentina SCHETTINI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919250
Fax: ++39 011 346384
schettin@inrim.it

Félicien SCHOPFER
LNE
ZA de Trappes-Élancourt
29 Avenue Roger Hennequin
78197 Trappes Cedex
France
Tel.: ++33 1 30692169
Fax: ++33 1 30162841
felicien.schopfer@lne.fr

Marco SELLONE
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919321
Fax: ++39 011 3919259
Raccomandato da: sellone@inrim.it

Ilaria SESIA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919258
Fax: ++39 011 3919259
sesia@inrim.it

Mario SICILIANI DE CUMIS
Dipartimento di Fisica
Università di Catania
Via Santa Sofia, 64
95100 Catania
Italy
Tel.: ++39 095 3785396
Fax: ++39 095 3785231
madecumis@ssc.unict.it,
mario.sicilianidecumis@imm.cnr.it

Emanuele TARALLI
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919220
Fax: ++39 011 346384
taralli@inrim.it

Alejandra TONINA
Instituto Nacional de Tecnología Industrial
Centro de Física y Metrología
Av. Gral Paz 5445
Casilla de Correo 157
B1650WAB San Martin
Buenos Aires
Argentina
Tel.: ++54 11 47246200/6300/6400
Fax: ++54 11 47134140
atonina@inti.gov.ar

Bruno Ottavio TRINCHERA
INRiM
Istituto Nazionale di Ricerca Metrologica
Strada delle Cacce 91
10135 Torino
Italy
Tel.: ++39 011 3919432
Fax: ++39 011 346384
trincher@inrim.it

Sándor VÖRÖS
Swiss Federal Office of Metrology
and Accreditation (METAS)
Lindenweg 50
3003 Bern-Wabern
Switzerland
Tel.: ++41 31 3233499
Fax: ++41 31 3233210
sandor.voros@metas.ch

Barney WALTON
NPL
Hampton Road
Teddington, Middlesex TW11 0LW
UK
Tel.: ++44 020 89436674
Fax: ++44 020 89432945
barney.walton@npl.co.uk

Rainer WINKLER
NPL
Hampton Road
Teddington, Middlesex TW11 0LW
UK
Tel.: ++44 020 89436910
fax ++44 020 89436755
rainer.winkler@npl.co.uk

Anne Lisa WOLF
Laser Centre Vrije Unversiteit
Dr. Boelelaan 1081
1081 HV Amsterdam
The Netherlands
Tel.: ++31 20 5987348
Fax: ++31 20 5987999
alwolf@few.vu.nl

Peter WOOLLIAMS
NPL
Hampton Road
TEDDINGTON, Middlesex TW11 0LW
UK
Tel.: ++44 020 89436328
peter.woolliams@npl.co.uk

Cédric ZUMSTEG
Université de Provence
Physique des Interactions ionique
et moléculaires
Centre de St. Jérôme, case C21
13397 MARSEILLE Cedex 20
France
Tel.: ++33 4 91288921
Fax: ++33 4 91288745
cedric.zumsteg@etu.univ-provence.fr

Observers

Luigi CACCIAPUOTI
ESA
Keplerlaan 1, PO Box 299
2200 AG NOORDWIJK
The Netherlands
Tel.: ++31 71 5655516
Fax: ++31 71 5654297
luigi.cacciapuoti@esa.int

Ralf D. GECKELER
PTB
Bundesallee 100
38116 BRAUNSCHWEIG
Germany
Tel.: ++49 531 5924211
Fax: ++49 531 5924218
ralf.geckeler@ptb.de

Jonathan PEARCE
NPL
Hampton Road
TEDDINGTON, Middlesex TW11 0LM
UK
Tel.: ++44 020 89436886
Fax: ++44 020 89436364
jonathan.pearce@npl.co.uk

This page intentionally left blank

PROCEEDINGS OF THE INTERNATIONAL SCHOOL OF PHYSICS "ENRICO FERMI"

Course I (1953)
Questioni relative alla rivelazione delle particelle elementari, con particolare riguardo alla radiazione cosmica
edited by G. Puppi

Course II (1954)
Questioni relative alla rivelazione delle particelle elementari, e alle loro interazioni con particolare riguardo alle particelle artificialmente prodotte ed accelerate
edited by G. Puppi

Course III (1955)
Questioni di struttura nucleare e dei processi nucleari alle basse energie
edited by C. Salvetti

Course IV (1956)
Proprietà magnetiche della materia
edited by L. Giulotto

Course V (1957)
Fisica dello stato solido
edited by F. Fumi

Course VI (1958)
Fisica del plasma e relative applicazioni astrofisiche
edited by G. Righini

Course VII (1958)
Teoria della informazione
edited by E. R. Caianiello

Course VIII (1958)
Problemi matematici della teoria quantistica delle particelle e dei campi
edited by A. Borsellino

Course IX (1958)
Fisica dei pioni
edited by B. Touschek

Course X (1959)
Thermodynamics of Irreversible Processes
edited by S. R. de Groot

Course XI (1959)
Weak Interactions
edited by L. A. Radicati

Course XII (1959)
Solar Radioastronomy
edited by G. Righini

Course XIII (1959)
Physics of Plasma: Experiments and Techniques
edited by H. Alfvén

Course XIV (1960)
Ergodic Theories
edited by P. Caldirola

Course XV (1960)
Nuclear Spectroscopy
edited by G. Racah

Course XVI (1960)
Physicomathematical Aspects of Biology
edited by N. Rashevsky

Course XVII (1960)
Topics of Radiofrequency Spectroscopy
edited by A. Gozzini

Course XVIII (1960)
Physics of Solids (Radiation Damage in Solids)
edited by D. S. Billington

Course XIX (1961)
Cosmic Rays, Solar Particles and Space Research
edited by B. Peters

Course XX (1961)
Evidence for Gravitational Theories
edited by C. Møller

Course XXI (1961)
Liquid Helium
edited by G. Careri

Course XXII (1961)
Semiconductors
edited by R. A. Smith

Course XXIII (1961)
Nuclear Physics
edited by V. F. Weisskopf

Course XXIV (1962)
Space Exploration and the Solar System
edited by B. Rossi

Course XXV (1962)
Advanced Plasma Theory
edited by M. N. Rosenbluth

Course XXVI (1962)
Selected Topics on Elementary Particle Physics
edited by M. Conversi

Course XXVII (1962)
Dispersion and Absorption of Sound by Molecular Processes
edited by D. Sette

Course XXVIII (1962)
Star Evolution
edited by L. Gratton

Course XXIX (1963)
Dispersion Relations and their Connection with Casuality
edited by E. P. Wigner

Course XXX (1963)
Radiation Dosimetry
edited by F. W. Spiers and G. W. Reed

Course XXXI (1963)
Quantum Electronics and Coherent Light
edited by C. H. Townes and P. A. Miles

Course XXXII (1964)
Weak Interactions and High-Energy Neutrino Physics
edited by T. D. Lee

Course XXXIII (1964)
Strong Interactions
edited by L. W. Alvarez

Course XXXIV (1965)
The Optical Properties of Solids
edited by J. Tauc

Course XXXV (1965)
High-Energy Astrophysics
edited by L. Gratton

Course XXXVI (1965)
Many-body Description of Nuclear Structure and Reactions
edited by C. L. Bloch

Course XXXVII (1966)
Theory of Magnetism in Transition Metals
edited by W. Marshall

Course XXXVIII (1966)
Interaction of High-Energy Particles with Nuclei
edited by T. E. O. Ericson

Course XXXIX (1966)
Plasma Astrophysics
edited by P. A. Sturrock

Course XL (1967)
Nuclear Structure and Nuclear Reactions
edited by M. Jean and R. A. Ricci

Course XLI (1967)
Selected Topics in Particle Physics
edited by J. Steinberger

Course XLII (1967)
Quantum Optics
edited by R. J. Glauber

Course XLIII (1968)
Processing of Optical Data by Organisms and by Machines
edited by W. Reichardt

Course XLIV (1968)
Molecular Beams and Reaction Kinetics
edited by Ch. Schlier

Course XLV (1968)
Local Quantum Theory
edited by R. Jost

Course XLVI (1969)
Physics with Intersecting Storage Rings
edited by B. Touschek

Course XLVII (1969)
General Relativity and Cosmology
edited by R. K. Sachs

Course XLVIII (1969)
Physics of High Energy Density
edited by P. Caldirola and H. Knoepfel

Course IL (1970)
Foundations of Quantum Mechanics
edited by B. d'Espagnat

Course L (1970)
Mantle and Core in Planetary Physics
edited by J. Coulomb and M. Caputo

Course LI (1970)
Critical Phenomena
edited by M. S. Green

Course LII (1971)
Atomic Structure and Properties of Solids
edited by E. Burstein

Course LIII (1971)
Developments and Borderlines of Nuclear Physics
edited by H. Morinaga

Course LIV (1971)
Developments in High-Energy Physics
edited by R. R. Gatto

Course LV (1972)
Lattice Dynamics and Intermolecular Forces
edited by S. Califano

Course LVI (1972)
Experimental Gravitation
edited by B. Bertotti

Course LVII (1972)
History of 20th Century Physics
edited by C. Weiner

Course LVIII (1973)
Dynamics Aspects of Surface Physics
edited by F. O. Goodman

Course LIX (1973)
Local Properties at Phase Transitions
edited by K. A. Müller and A. Rigamonti

Course LX (1973)
C*-Algebras and their Applications to Statistical Mechanics and Quantum Field Theory
edited by D. Kastler

Course LXI (1974)
Atomic Structure and Mechanical Properties of Metals
edited by G. Caglioti

Course LXII (1974)
Nuclear Spectroscopy and Nuclear Reactions with Heavy Ions
edited by H. Faraggi and R. A. Ricci

Course LXIII (1974)
New Directions in Physical Acoustics
edited by D. Sette

Course LXIV (1975)
Nonlinear Spectroscopy
edited by N. Bloembergen

Course LXV (1975)
Physics and Astrophysics of Neutron Stars and Black Hole
edited by R. Giacconi and R. Ruffini

Course LXVI (1975)
Health and Medical Physics
edited by J. Baarli

Course LXVII (1976)
Isolated Gravitating Systems in General Relativity
edited by J. Ehlers

Course LXVIII (1976)
Metrology and Fundamental Constants
edited by A. Ferro Milone, P. Giacomo and S. Leschiutta

Course LXIX (1976)
Elementary Modes of Excitation in Nuclei
edited by A. Bohr and R. A. Broglia

Course LXX (1977)
Physics of Magnetic Garnets
edited by A. Paoletti

Course LXXI (1977)
Weak Interactions
edited by M. Baldo Ceolin

Course LXXII (1977)
Problems in the Foundations of Physics
edited by G. Toraldo di Francia

Course LXXIII (1978)
Early Solar System Processes and the Present Solar System
edited by D. Lal

Course LXXIV (1978)
Development of High-Power Lasers and their Applications
edited by C. Pellegrini

Course LXXV (1978)
Intermolecular Spectroscopy and Dynamical Properties of Dense Systems
edited by J. Van Kranendonk

Course LXXVI (1979)
Medical Physics
edited by J. R. Greening

Course LXXVII (1979)
Nuclear Structure and Heavy-Ion Collisions
edited by R. A. Broglia, R. A. Ricci and C. H. Dasso

Course LXXVIII (1979)
Physics of the Earth's Interior
edited by A. M. Dziewonski and E. Boschi

Course LXXIX (1980)
From Nuclei to Particles
edited by A. Molinari

Course LXXX (1980)
Topics in Ocean Physics
edited by A. R. Osborne and P. Malanotte Rizzoli

Course LXXXI (1980)
Theory of Fundamental Interactions
edited by G. Costa and R. R. Gatto

Course LXXXII (1981)
Mechanical and Thermal Behaviour of Metallic Materials
edited by G. Caglioti and A. Ferro Milone

Course LXXXIII (1981)
Positrons in Solids
edited by W. Brandt and A. Dupasquier

Course LXXXIV (1981)
Data Acquisition in High-Energy Physics
edited by G. Bologna and M. Vincelli

Course LXXXV (1982)
Earthquakes: Observation, Theory and Interpretation
edited by H. Kanamori and E. Boschi

Course LXXXVI (1982)
Gamow Cosmology
edited by F. Melchiorri and R. Ruffini

Course LXXXVII (1982)
Nuclear Structure and Heavy-Ion Dynamics
edited by L. Moretto and R. A. Ricci

Course LXXXVIII (1983)
Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics
edited by M. Ghil, R. Benzi and G. Parisi

Course LXXXIX (1983)
Highlights of Condensed Matter Theory
edited by F. Bassani, F. Fumi and M. P. Tosi

Course XC (1983)
Physics of Amphiphiles: Micelles, Vesicles and Microemulsions
edited by V. Degiorgio and M. Corti

Course XCI (1984)
From Nuclei to Stars
edited by A. Molinari and R. A. Ricci

Course XCII (1984)
Elementary Particles
edited by N. Cabibbo

Course XCIII (1984)
Frontiers in Physical Acoustics
edited by D. Sette

Course XCIV (1984)
Theory of Reliability
edited by A. Serra and R. E. Barlow

Course XCV (1985)
Solar-Terrestrial Relationship and the Earth Environment in the Last Millennia
edited by G. Cini Castagnoli

Course XCVI (1985)
Excited-State Spectroscopy in Solids
edited by U. M. Grassano and N. Terzi

Course XCVII (1985)
Molecular-Dynamics Simulations of Statistical-Mechanical Systems
edited by G. Ciccotti and W. G. Hoover

Course XCVIII (1985)
The Evolution of Small Bodies in the Solar System
edited by M. Fulchignoni and Ľ. Kresàk

Course XCIX (1986)
Synergetics and Dynamic Instabilities
edited by G. Caglioti and H. Haken

Course C (1986)
The Physics of NMR Spectroscopy in Biology and Medicine
edited by B. Maraviglia

Course CI (1986)
Evolution of Interstellar Dust and Related Topics
edited by A. Bonetti and J. M. Greenberg

Course CII (1986)
Accelerated Life Testing and Experts' Opinions in Reliability
edited by C. A. Clarotti and D. V. Lindley

Course CIII (1987)
Trends in Nuclear Physics
edited by P. Kienle, R. A. Ricci and A. Rubbino

Course CIV (1987)
Frontiers and Borderlines in Many-Particle Physics
edited by R. A. Broglia and J. R. Schrieffer

Course CV (1987)
Confrontation between Theories and Observations in Cosmology: Present Status and Future Programmes
edited by J. Audouze and F. Melchiorri

Course CVI (1988)
Current Trends in the Physics of Materials
edited by G. F. Chiarotti, F. Fumi and M. Tosi

Course CVII (1988)
The Chemical Physics of Atomic and Molecular Clusters
edited by G. Scoles

Course CVIII (1988)
Photoemission and Absorption Spectroscopy of Solids and Interfaces with Synchrotron Radiation
edited by M. Campagna and R. Rosei

Course CIX (1988)
Nonlinear Topics in Ocean Physics
edited by A. R. Osborne

Course CX (1989)
Metrology at the Frontiers of Physics and Technology
edited by L. Crovini and T. J. Quinn

Course CXI (1989)
Solid-State Astrophysics
edited by E. Bussoletti and G. Strazzulla

Course CXII (1989)
Nuclear Collisions from the Mean-Field into the Fragmentation Regime
edited by C. Detraz and P. Kienle

Course CXIII (1989)
High-Pressure Equation of State: Theory and Applications
edited by S. Eliezer and R. A. Ricci

Course CXIV (1990)
Industrial and Technological Applications of Neutrons
edited by M. Fontana and F. Rustichelli

Course CXV (1990)
The Use of EOS for Studies of Atmospheric Physics
edited by J. C. Gille and G. Visconti

Course CXVI (1990)
Status and Perspectives of Nuclear Energy: Fission and Fusion
edited by R. A. Ricci, C. Salvetti and E. Sindoni

Course CXVII (1991)
Semiconductor Superlattices and Interfaces
edited by A. Stella

Course CXVIII (1991)
Laser Manipulation of Atoms and Ions
edited by E. Arimondo, W. D. Phillips and F. Strumia

Course CXIX (1991)
Quantum Chaos
edited by G. Casati, I. Guarneri and U. Smilansky

Course CXX (1992)
Frontiers in Laser Spectroscopy
edited by T. W. Hänsch and M. Inguscio

Course CXXI (1992)
Perspectives in Many-Particle Physics
edited by R. A. Broglia, J. R. Schrieffer and P. F. Bortignon

Course CXXII (1992)
Galaxy Formation
edited by J. Silk and N. Vittorio

Course CXXIII (1992)
Nuclear Magnetic Double Resonsonance
edited by B. Maraviglia

Course CXXIV (1993)
Diagnostic Tools in Atmospheric Physics
edited by G. Fiocco and G. Visconti

Course CXXV (1993)
Positron Spectroscopy of Solids
edited by A. Dupasquier and A. P. Mills jr.

Course CXXVI (1993)
Nonlinear Optical Materials: Principles and Applications
edited by V. Degiorgio and C. Flytzanis

Course CXXVII (1994)
Quantum Groups and their Applications in Physics
edited by L. Castellani and J. Wess

Course CXXVIII (1994)
Biomedical Applications of Synchrotron Radiation
edited by E. Burattini and A. Balerna

Course CXXIX[1] (1994)
Observation, Prediction and Simulation of Phase Transitions in Complex Fluids
edited by M. Baus, L. F. Rull and J. P. Ryckaert

Course CXXX (1995)
Selected Topics in Nonperturbative QCD
edited by A. Di Giacomo and D. Diakonov

Course CXXXI (1995)
Coherent and Collective Interactions of Particles and Radiation Beams
edited by A. Aspect, W. Barletta and R. Bonifacio

Course CXXXII (1995)
Dark Matter in the Universe
edited by S. Bonometto and J. Primack

Course CXXXIII (1996)
Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models
edited by G. Cini Castagnoli and A. Provenzale

Course CXXXIV (1996)
The Physics of Complex Systems
edited by F. Mallamace and H. E. Stanley

Course CXXXV (1996)
The Physics of Diamond
edited by A. Paoletti and A. Tucciarone

Course CXXXVI (1997)
Models and Phenomenology for Conventional and High-Temperature Superconductivity
edited by G. Iadonisi, J. R. Schrieffer and M. L. Chiofalo

Course CXXXVII (1997)
Heavy Flavour Physics: a Probe of Nature's Grand Design
edited by I. Bigi and L. Moroni

Course CXXXVIII (1997)
Unfolding the Matter of Nuclei
edited by A. Molinari and R. A. Ricci

Course CXXXIX (1998)
Magnetic Resonance and Brain Function: Approaches from Physics
edited by B. Maraviglia

Course CXL (1998)
Bose-Einstein Condensation in Atomic Gases
edited by M. Inguscio, S. Stringari and C. E. Wieman

Course CXLI (1998)
Silicon-Based Microphotonics: from Basics to Applications
edited by O. Bisi, S. U. Campisano, L. Pavesi and F. Priolo

[1] This course belongs to the NATO ASI Series C, Vol. 460 (Kluwer Academic Publishers).

Course CXLII (1999)
Plasmas in the Universe
edited by B. Coppi, A. Ferrari and E. Sindoni

Course CXLIII (1999)
New Directions in Quantum Chaos
edited by G. Casati, I. Guarneri and U. Smilansky

Course CXLIV (2000)
Nanometer Scale Science and Technology
edited by M. Allegrini, N. García and O. Marti

Course CXLV (2000)
Protein Folding, Evolution and Design
edited by R. A. Broglia, E. I. Shakhnovich and G. Tiana

Course CXLVI (2000)
Recent Advances in Metrology and Fundamental Constants
edited by T. J. Quinn, S. Leschiutta and P. Tavella

Course CXLVII (2001)
High Pressure Phenomena
edited by R. J. Hemley, G. L. Chiarotti, M. Bernasconi and L. Ulivi

Course CXLVIII (2001)
Experimental Quantum Computation and Information
edited by F. De Martini and C. Monroe

Course CXLIX (2001)
Organic Nanostructures: Science and Applications
edited by V. M. Agranovich and G. C. La Rocca

Course CL (2002)
Electron and Photon Confinement in Semiconductor Nanostructures
edited by B. Deveaud-Plédran, A. Quattropani and P. Schwendimann

Course CLI (2002)
Quantum Phenomena in Mesoscopic Systems
edited by B. Altshuler, A. Tagliacozzo and V. Tognetti

Course CLII (2002)
Neutrino Physics
edited by E. Bellotti, Y. Declais and P. Strolin

Course CLIII (2002)
From Nuclei and their Constituents to Stars
edited by A. Molinari, L. Riccati, W. M. Alberico and M. Morando

Course CLIV (2003)
Physics Methods in Archaeometry
edited by M. Martini, M. Milazzo and M. Piacentini

Course CLV (2003)
The Physics of Complex Systems (New Advances and Perspectives)
edited by F. Mallamace and H. E. Stanley

Course CLVI (2003)
Research on Physics Education
edited by E.F. Redish and M. Vicentini

Course CLVII (2003)
The Electron Liquid Model in Condensed Matter Physics
edited by G. F. Giuliani and G. Vignale

Course CLVIII (2004)
Hadron Physics
edited by T. Bressani, U. Wiedner and A. Filippi

Course CLIX (2004)
Background Microwave Radiation and Intracluster Cosmology
edited by F. Melchiorri and Y. Rephaeli

Course CLX (2004)
From Nanostructures to Nanosensing Applications
edited by A. D'Amico, G. Balestrino and A. Paoletti

Course CLXI (2005)
Polarons in Bulk Materials and Systems with Reduced Dimensionality
edited by G. Iadonisi, J. Ranninger and G. De Filippis

Course CLXII (2005)
Quantum Computers, Algorithms and Chaos
edited by G. Casati, D. L. Shepelyansky, P. Zoller and G. Benenti

Course CLXIII (2005)
CP Violation: From Quarks to Leptons
edited by M. Giorgi, I. Mannelli, A. I. Sanda, F. Costantini and M. S. Sozzi

Course CLXIV (2006)
Ultra-Cold Fermi Gases
edited by M. Inguscio, W. Ketterle and C. Salomon

Course CLXV (2006)
Protein Folding and Drug Design
edited by R. A. Broglia, L. Serrano and G. Tiana

Course CLXVI (2006)
Metrology and Fundamental Constants
edited by T. W. Hänsch, S. Leschiutta, A. J. Wallard and M. L. Rastello