Deep-Submicron CMOS ICs

Deep-Submicron CMOS ICs

From Basics to ASICs

Harry Veendrick

Kluwer academic publishers

Deep-Submicron CMOS ICs

Author:
Ir. H.J.M. Veendrick
Philips Research Laboratories
Prof. Holstlaan 4
NL-5656 AA Eindhoven
The Netherlands
E-mail: harry.veendrick@philips.com

Cover:
Design: Helfrich & Slotemaker, Deventer
Photograph: Philips Semiconductors

Typesetting and layout: Dré van den Elshout
Illustrations: Henny Alblas

First English edition, 1998
Second English edition, 2000

The first edition represents a thoroughly revised, updated and more comprehensive version of the previous book entitled `MOS ICs: from Basics to ASICs' by the same author. Originally published in the Dutch language (Delta Press BV, 1990), in 1992 a revised and translated English edition of this book was jointly published by:
– VCH Verlagsgesellschaft mbH (Weinheim, Federal Republic of Germany)
– VCH Publishers Inc. (NY, USA).

The second edition includes completely updated material, particularly related to technology progress and roadmap implications.

ISBN 90 440 01116
NUGI 832

© 2000 Ten Hagen en Stam, Deventer, The Netherlands
Kluwer academic publishers, Dordrecht, The Netherlands / London, U.K./ Boston, U.S.A.

All rights reserved. No part of this book may be reproduced, stored in a database or retrieval system, or published, in any form or in any way, electronically, mechanically, by print, photoprint, microfilm or any other means without prior written permission from the publisher. Information published in this work, in any form, may be subject to patent rights and is intended for study purposes and private use only.

Although this book was produced with the utmost care, neither the authors and editors nor the publisher can guarantee that the information contained therein is free from errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently contain inaccuracies.

Foreword

Deep-submicron technology will soon enable us to design systems of unprecedented complexity on a single chip. At the same time more and more physical phenomena will affect the performance, reliability and energy consumption of these CMOS circuits. Therefore, system-on-a-chip design is essentially team work, requiring a close dialogue between system designers, software engineers, chip architects, intellectual property providers and process engineers. This requires a common understanding of the CMOS medium, its terminology, its future opportunities and possible pitfalls.

This second edition provides a completely updated, but comprehensive view of all aspects of CMOS ASIC design. Starting from the basics of MOS devices and technologies, circuits and subsystems, it leads the reader into the novel intricacies of deep-submicron systems.

It contains a systematic view on how to maintain low-power dissipation, how to master signal integrity in spite of lower voltages, higher current peaks and larger cross-talk in the interconnections. It shows how to master clock-skew problems and pays attention to packaging, testing and debugging of deep-submicron chips. Finally, the author shares his thoughts on the future of CMOS up to the year 2010.

This book is the first complete comprehensive book that covers all aspects of designing well functioning systems on deep-submicron silicon. It is the reflection of the author's own research in this domain and also of over 20 years experience in interactive teaching of CMOS design to Philips system and IC designers and test and process engineers alike. It provides context and perspective to both sides. I strongly recommend this book to all engineers involved in the design, test and manufacture of future systems-on-silicon as well as to engineering undergraduates who want to understand the basics that make electronics systems work.

Leuven, Summer 2000

Hugo De Man,
Professor K.U. Leuven,
Senior Research Fellow IMEC,
Leuven,
Belgium

Preface to First and Second Edition

An integrated circuit (IC) is a piece of semiconductor material, on which a number of electronic components are interconnected. These interconnected 'chip' components implement a specific function. The semiconductor material is usually silicon but alternatives include gallium arsenide.

ICs are essential in most modern electronic products. The first IC was created by Jack Kilby in 1959. Photographs of this device and the inventor are shown in figure 3. Figure 1 illustrates the subsequent progress in IC complexity. This figure shows the numbers of components for advanced ICs and the year in which these ICs were first presented. This quadrupling in complexity every three years was predicted by Moore's law (INTEL 1964), which is still valid today.

Figures 4 to 7 illustrate the evolution in IC technology. Figure 4 shows a discrete BC107 transistor. The digital filter shown in figure 5 comprises a few thousand transistors while the Digital Audio Broadcasting (DAB) chip in figure 6 contains more than six million transistors. Figure 7 shows a nine million transistors Complex Programmable Logic Device (CPLD).

Figure 2 shows the trends in IC manufacturing technologies from 1983 to 2002. About 80% of all ICs are now manufactured in MOS processes. These processes facilitate the integration of several tens of millions of components on a chip area of 1 cm^2. They are used for the manufacture of memory and so-called VLSI (Very Large Scale Integration) chips. Figure 8 illustrates the magnitude of the associated miniaturisation. The resulting IC details are even smaller than those of an insect.

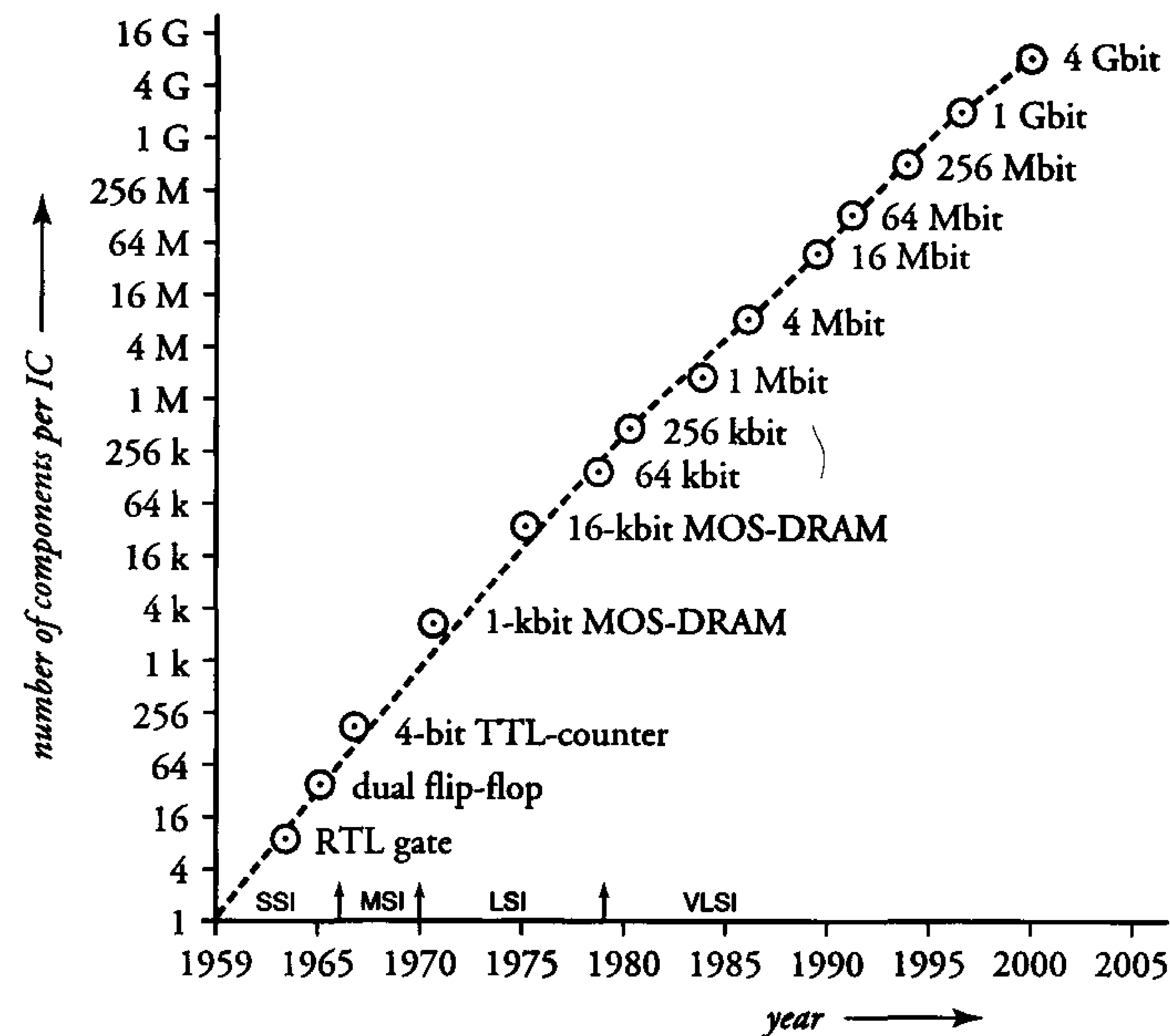


Figure 1: *Growth in the number of components per IC*

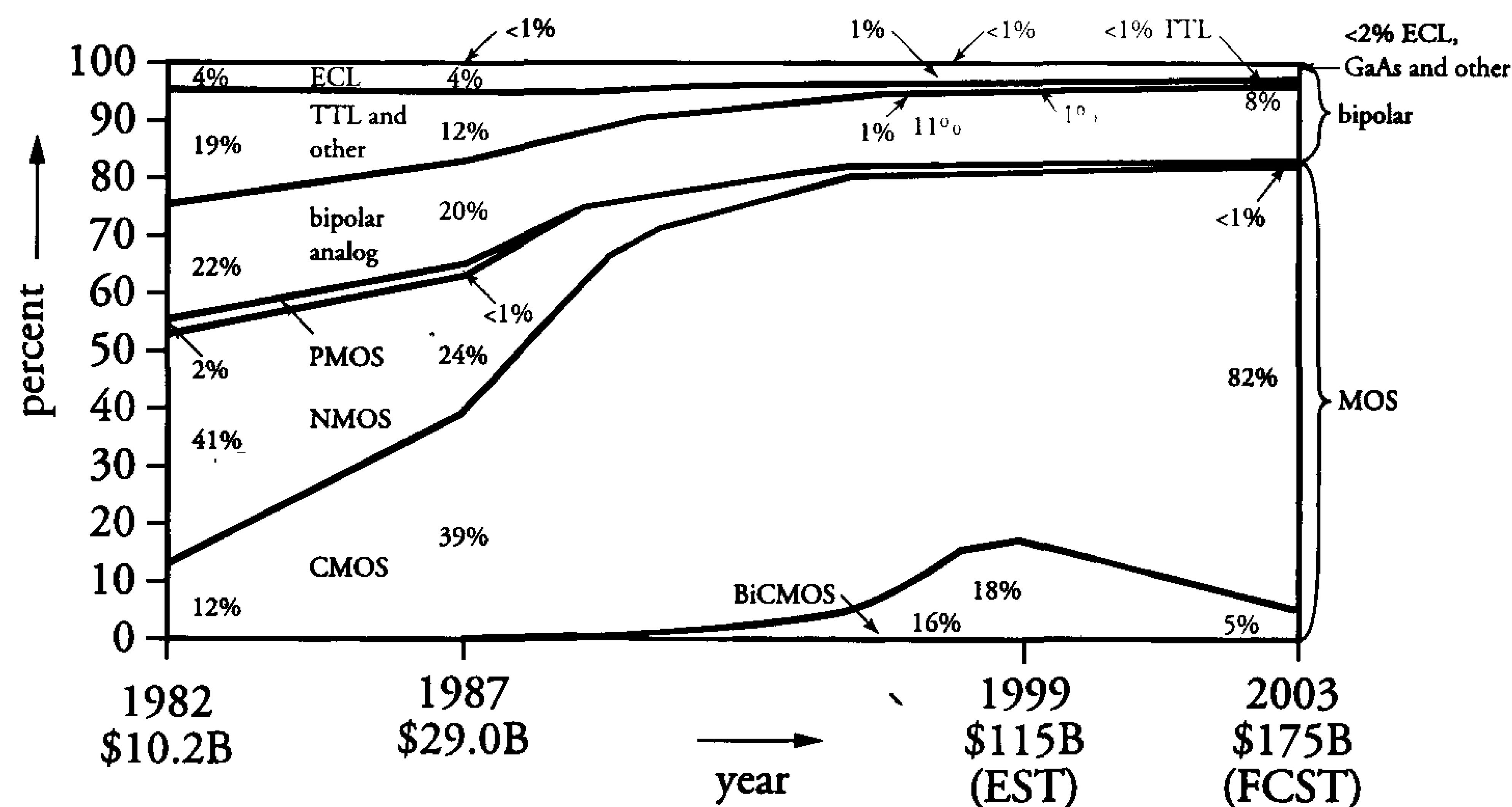


Figure 2: *Technology trends from 1983 to 2003 (source: ICE)*

This book provides an insight into all aspects associated with CMOS ICs. The topics presented include relevant fundamental physics. Technology, design and implementation aspects are also explained and applications are discussed. CAD tools used for the realisation of ICs are described while current and expected developments also receive considerable attention.

The contents of this book are based on the CMOS section of an industry-oriented course entitled 'An introduction to IC techniques'. The course has been given for more than two decades within PHILIPS. Continuous revision and expansion of the course material ensures that this book is highly relevant to the IC industry. The level of the discussions makes this book a suitable introduction for designers, technologists, CAD developers, test engineers, reliability engineers, technical-commercial personnel and IC applicants. The text is also suitable for both graduates and undergraduates in related engineering courses.

Considerable effort has been made to enhance the readability of this book and only essential formulae are included. The large number of diagrams and photographs should reinforce the explanations. The design and application examples are mainly digital. This reflects the fact that more than 95% of all modern CMOS ICs are digital circuits. However, the material presented will also provide the analogue designer with a basic understanding of the physics, manufacture and operation of deep-submicron CMOS circuits. The chapters are summarised below.

Chapter 1 contains detailed discussions of the basic principles and fundamental physics of the MOS transistor. The derivation of simple current-voltage equations for MOS devices and the explanation of their characteristics illustrates the relationship between process parameters and circuit performance.

The continuous reduction of transistor dimensions leads to increased deviation between the performance predicted by the simple MOS formulae and actual transistor behaviour. The effects of temperature and geometry on this behaviour are explained in chapter 2. In addition to their influence on transistor and circuit performance, these effects can also reduce device lifetime and reliability.

The various technologies for the manufacture of CMOS ICs are examined in chapter 3. An explanation of the most important associated photolithographic and processing steps is provided. This precedes a discussion of an advanced deep-submicron technology for the manufacture of modern VLSI circuits.

The design of CMOS circuits is treated in chapter 4. An introduction to the performance aspects of nMOS circuits provides an extremely

useful background for the explanation of the CMOS design and layout procedures.

MOS technologies and their derivatives are used to realise the special devices discussed in chapter 5. Charge-coupled devices (CCDs) and MOS power transistors are among the special devices. Chapter 5 concludes the presentation of the fundamental concepts behind BICMOS circuit operation.

Memories currently represent almost 30 to 40% of the IC market. This share is expected to continue to increase. The majority of available memory types are therefore examined in chapter 6. The basic structures and the operating principles of the various types are explained. In addition, the relationships between their respective properties and application areas is made clear.

Developments in IC technology now facilitate the integration of complete systems on a chip, which contain several tens of millions of transistors. The various IC design and realisation techniques used for these VLSI ICs are presented in chapter 7. The advantages and disadvantages of the techniques and the associated CAD tools are examined. Various modern technologies are used to realise a separate class of VLSI ICs, which are specified by applicants rather than manufacturers. These application-specific ICs (ASICs) are examined in this chapter as well. Market aspects and motives for their use are also discussed.

As a result of the continuous increase of power consumption, the maximum level that can be sustained by cheap plastic packages has been reached. Therefore, all CMOS designers must have a 'low-power attitude'. Chapter 8 presents a complete overview of low-power options for CMOS technologies, as well as for the different levels of design hierarchy.

Increased VLSI design complexities, combined with higher frequencies create a higher sensitivity to physical effects. These effects have started to dominate the reliability and signal integrity of deep-submicron CMOS ICs. Chapter 9 discusses these effects and the design measures to be taken to maintain both reliability and signal integrity at a sufficiently high level.

Finally, testing, debugging and packaging are important factors that contribute to the ultimate costs of an IC. Chapter 10 presents an overview of the state-of-the-art techniques that support testing, debugging and failure analysis. It also includes a rather detailed summary on available packaging technologies and gives an insight into their future trends. Essential factors related to IC production are also examined; these factors include quality and reliability.

The continuous reduction of transistor dimensions associated with successive process generations is the subject of the final chapter (chapter 11). This scaling has various consequences for transistor behaviour and IC performance. The resulting increase of physical effects and the associated effects on reliability and signal integrity are the main focus of attention. The expected consequences of further miniaturisation are described. This provides an insight into the challenges facing the IC industry in the race towards sub-100 nm devices.

Not all data in this book is completely sprout from my mind. A lot of books and papers contributed to make the presented material state-of-the-art. Considerable effort has been made to make the reference list complete and correct. I apologize for possible imperfections.

Acknowledgements

I wish to express my gratitude to all those who contributed to the realisation of this book; it is impossible to include all their names. Their contributions included fruitful discussions, relevant texts and manuscript reviews.

I would especially like to thank Dick Klaassen for reviewing chapter 2 and André Montree, Pierre Woerlee, Rob Wolters, Rob Verhaar for the review of technology topics, and Casper Juffermans for the lithography topics in chapter 3. I also wish to sincerely thank Michel de Langen for his contribution and review of the section on packaging, Jan Stuyt and Jef van Meerbergen for the review of the VLSI chapter, Roger Cuppens, Kees van der Sanden and Roelof Salters for their review of the memory chapter, Paul Simon for the yield section and Erik-Jan Marinissen for his contribution to the test topics. And, I want to thank Ed Huijbregts for the review of the final chapter on scaling aspects. Last but not least, I would like to thank Harm Peters for drawing the CMOS layouts and Andrew Robertson for providing me with beautiful, full colour photographic material.

I am very grateful to all those who attended the course, because their feedback on educational aspects, their corrections and constructive criticism contributed to the quality and completeness of this book.

In addition, I want to thank Philips, in general, for the co-operation I was afforded, Henny Alblas and Ron Salfrais, for the drawing material and the correctness of the English text, respectively, and the Centre for

Technology Training (CTT) for sponsoring this work.

Finally, I wish to thank Dré van den Elshout for his conscientious editing and type-setting work. His efforts to ensure high quality should not go unnoticed by the reader.

However, the most important appreciation and gratitude must go to my family, again, for their years of exceptional tolerance, patience and understanding. The year 1998 was particularly demanding, both personally and professionally. It must seem to them as though I will never stop writing and revising books.

Eindhoven, October 1998 Harry J.M. Veendrick

This second edition contains some corrections and is completely updated with respect to the previous one. In the one-and-a-half years of its existance, the first edition has already been used in more than ten in-house courses. Several typing errors and the like, which showed up during these courses, have been corrected. Moreover, most of the chapters have been updated with state-of-the-art material. Numbers that describe trends and roadmaps have been updated as well, to let the contents of this book be valuable for at least another five years.

Eindhoven, Summer 2000 Harry J.M. Veendrick

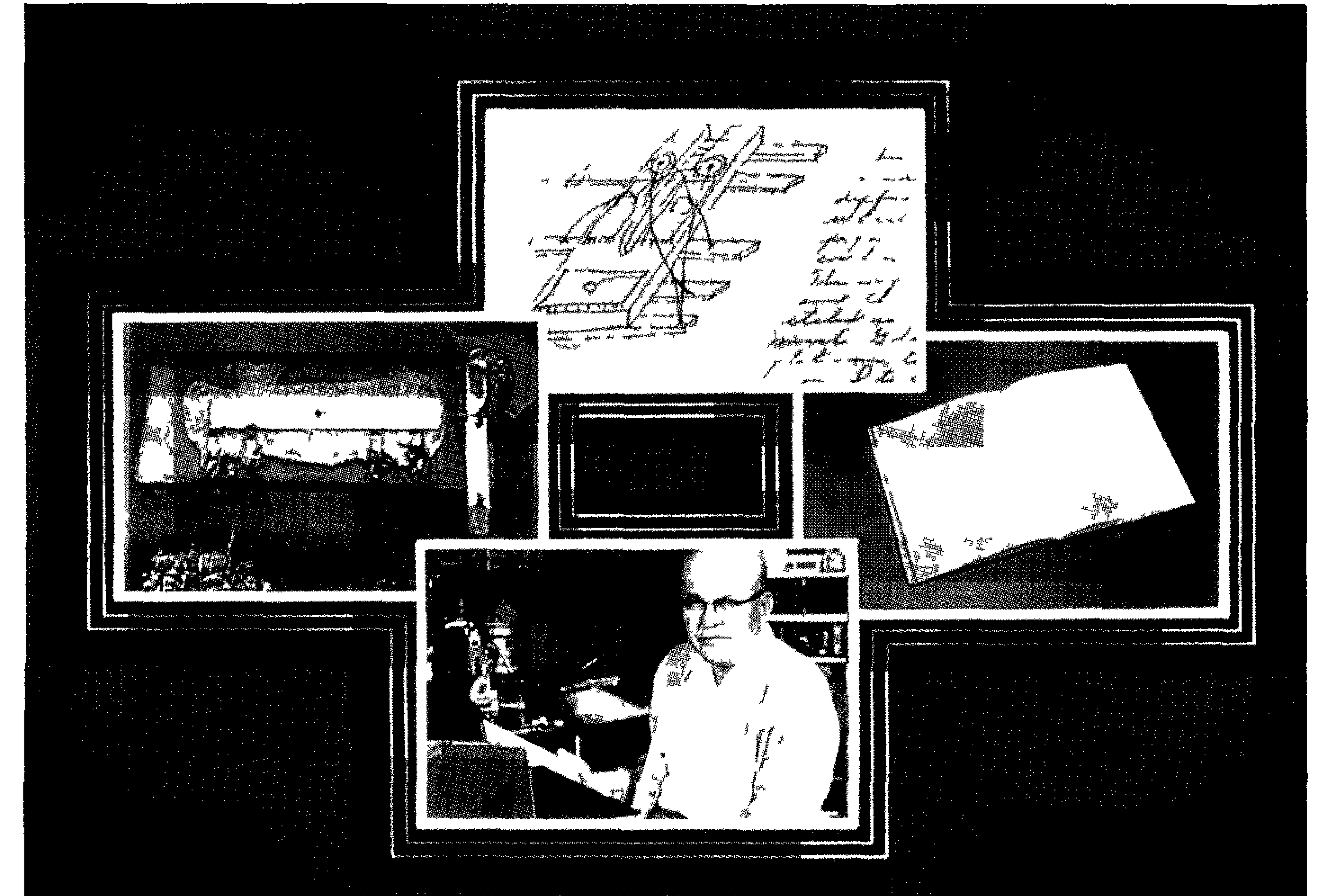


Figure 3: *The development of the first IC: in 1958 Jack Kilby demonstrated the feasibility of resistors and capacitors, in addition to transistors, based on semiconductor technology. Kilby, an employee of Texas Instruments, submitted the patent request entitled 'Miniaturized Electronic Circuits' in 1959. His request was honoured. Recognition by a number of Japanese companies in 1990 means that Texas Instruments is still benefiting from Kilby's patent (photo: Texas Instruments / Koning & Hartman).*

Figure 4: *A single BC107 bipolar transistor (photo: PHILIPS)*

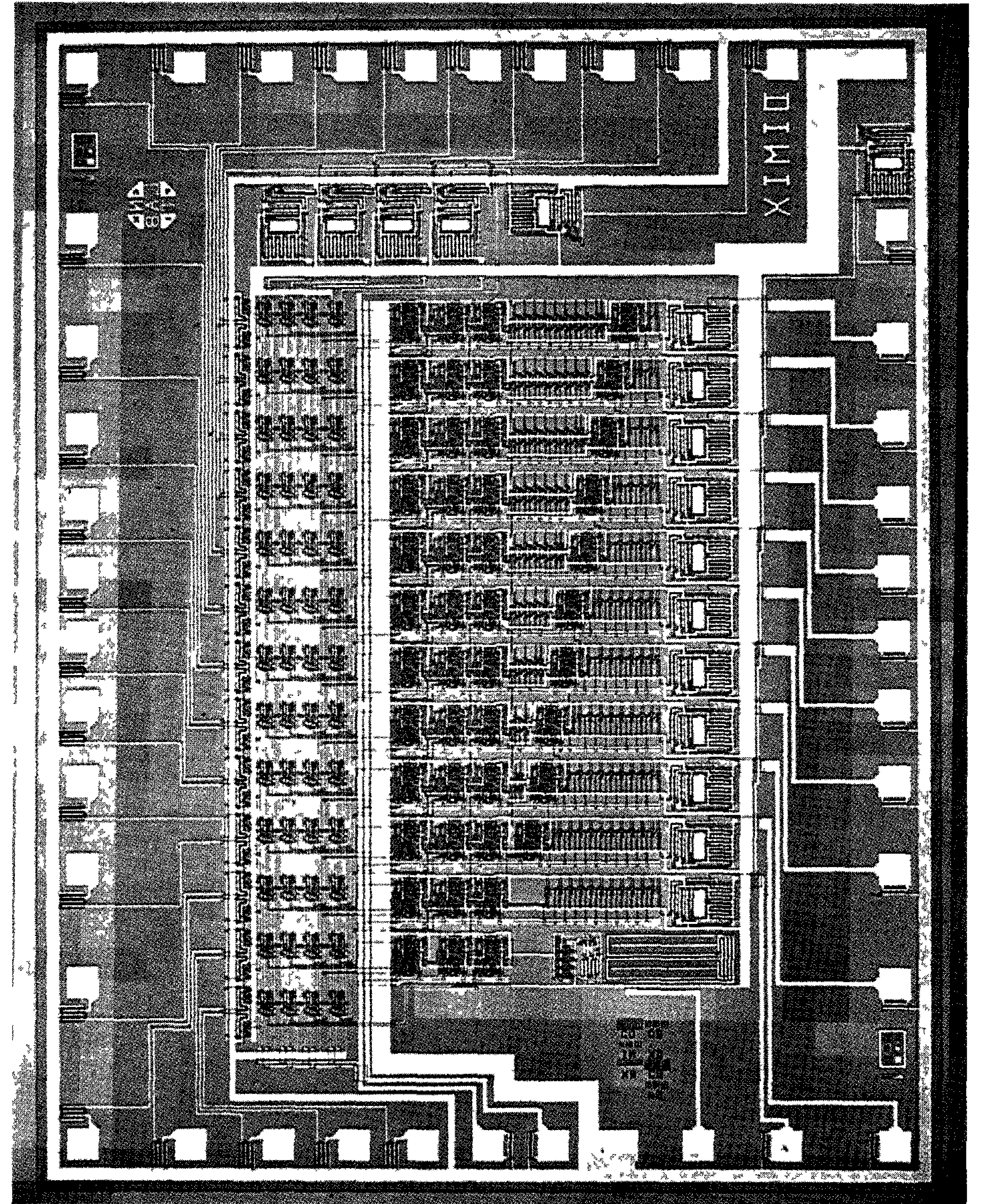


Figure 5: *A digital filter which comprises a few thousand transistors (photo: PHILIPS)*

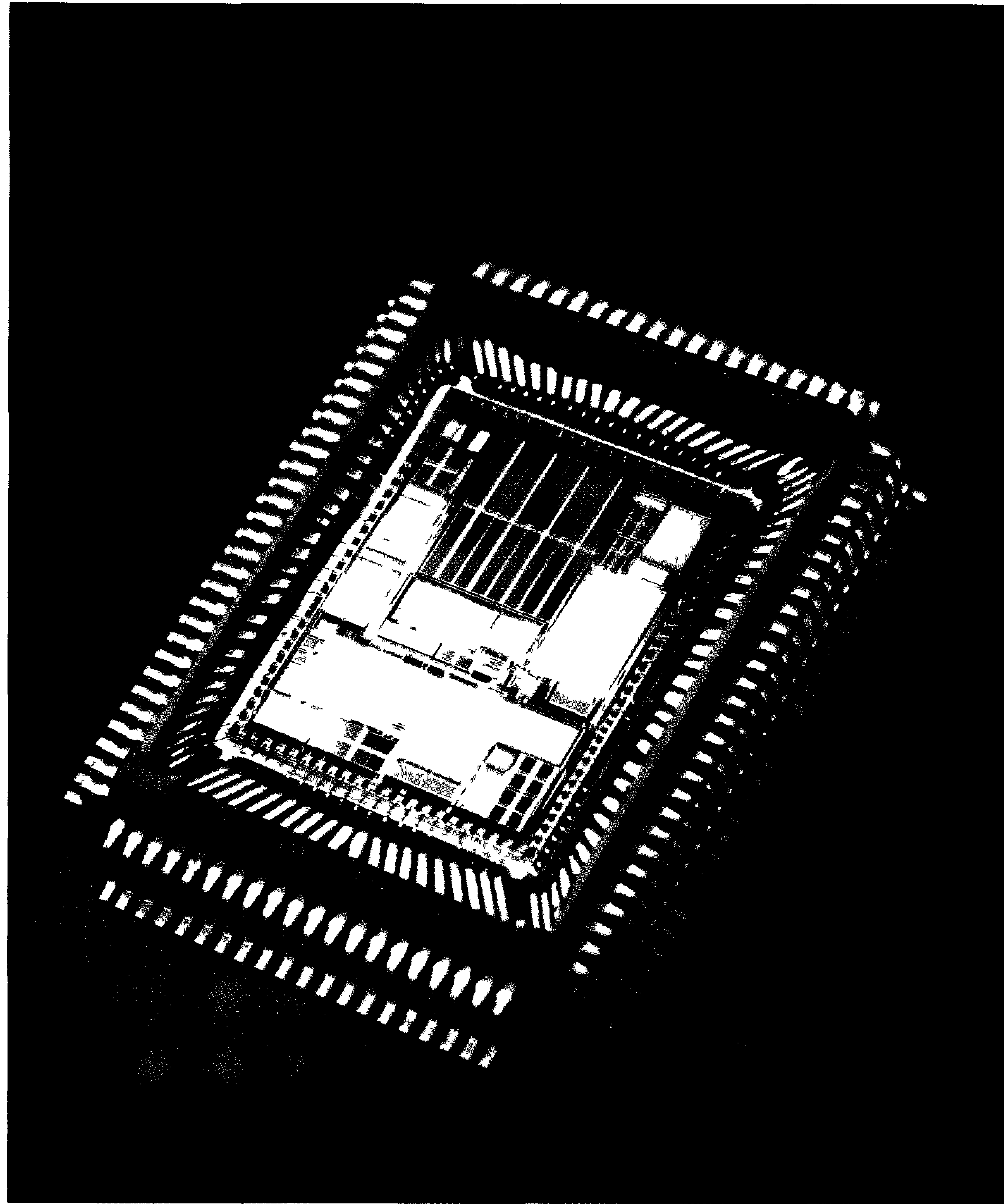


Figure 6: *A Digital Audio Broadcasting (DAB) chip, which comprises more than six million transistors (photo: PHILIPS)*

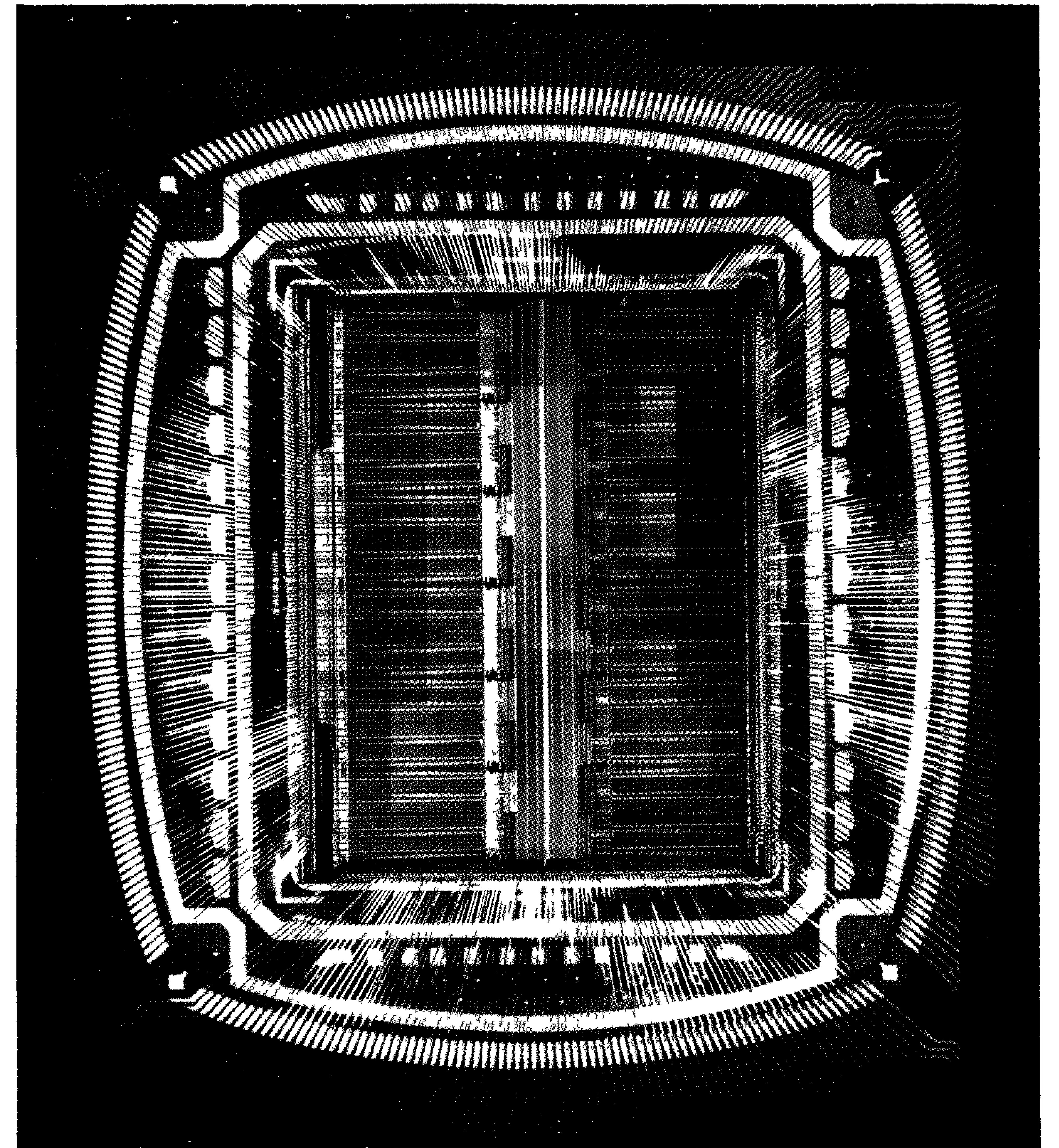


Figure 7: *A Complex Programmable Logic Device, which comprises about nine million transistors (photo: PHILIPS)*

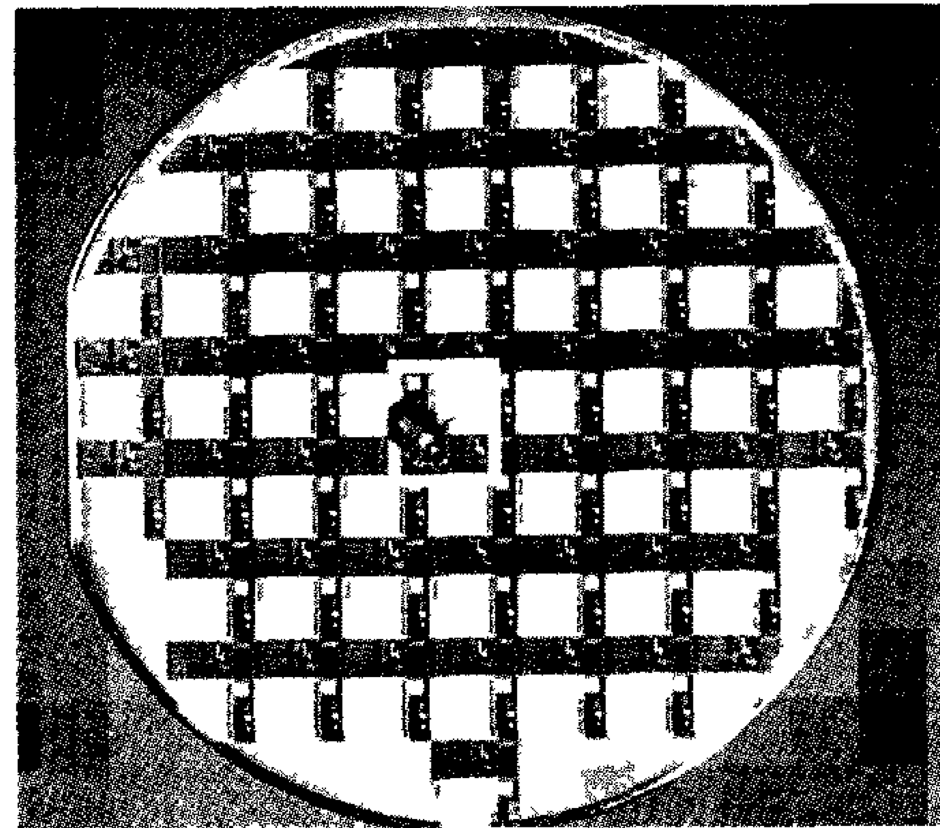
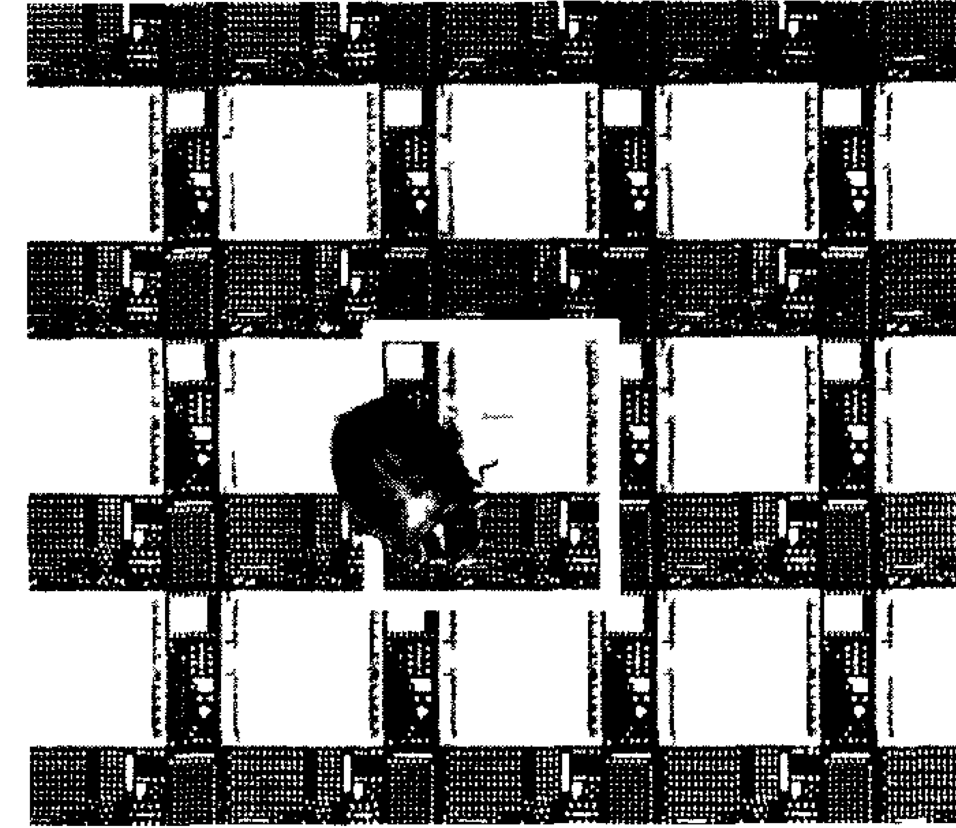
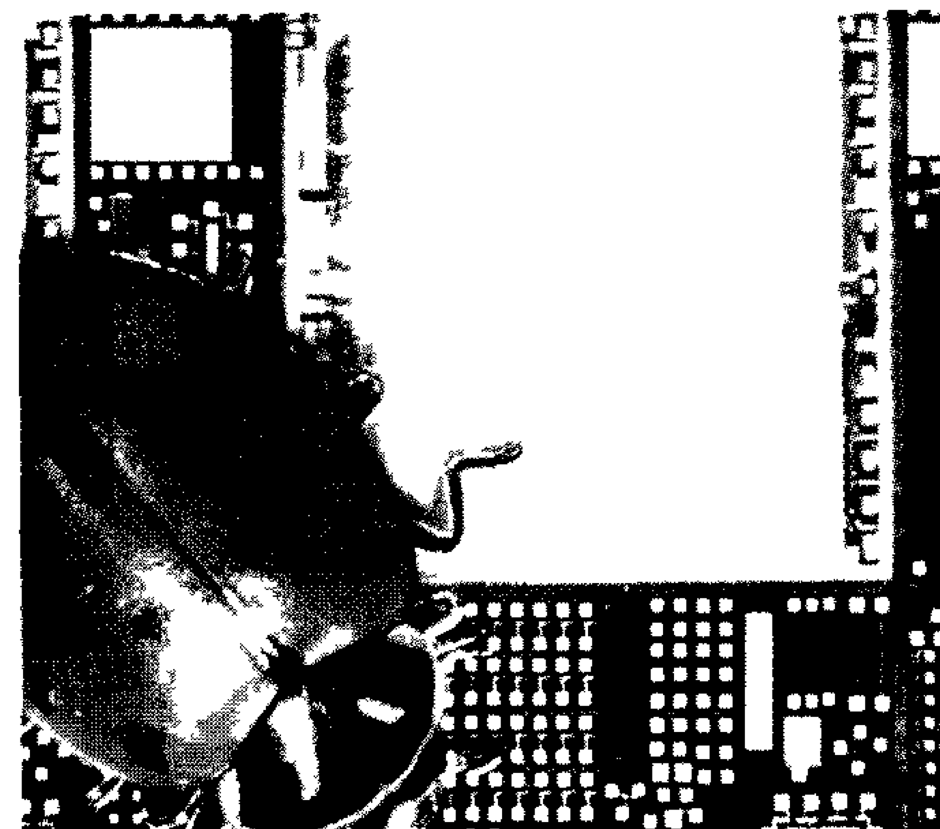
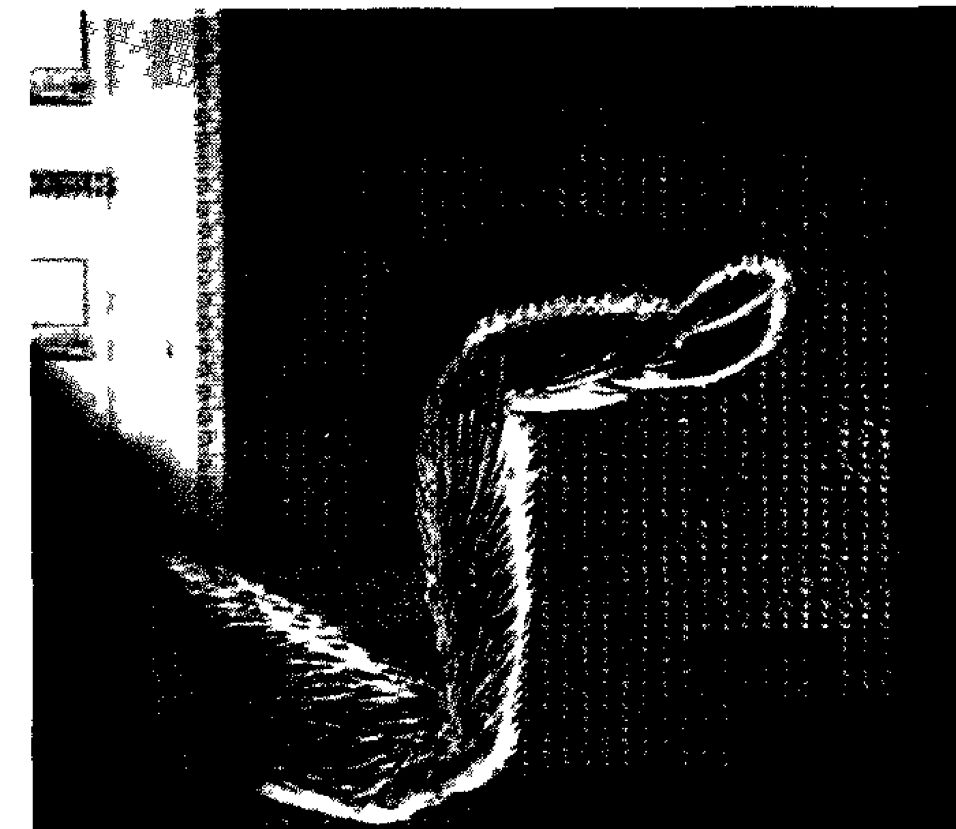
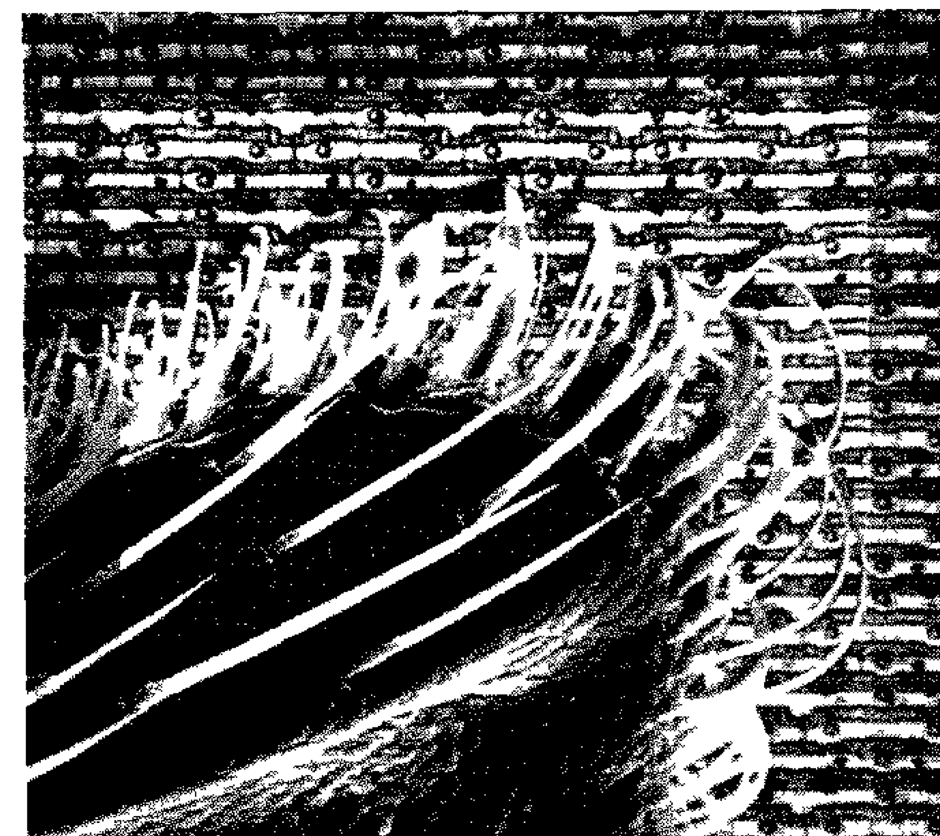
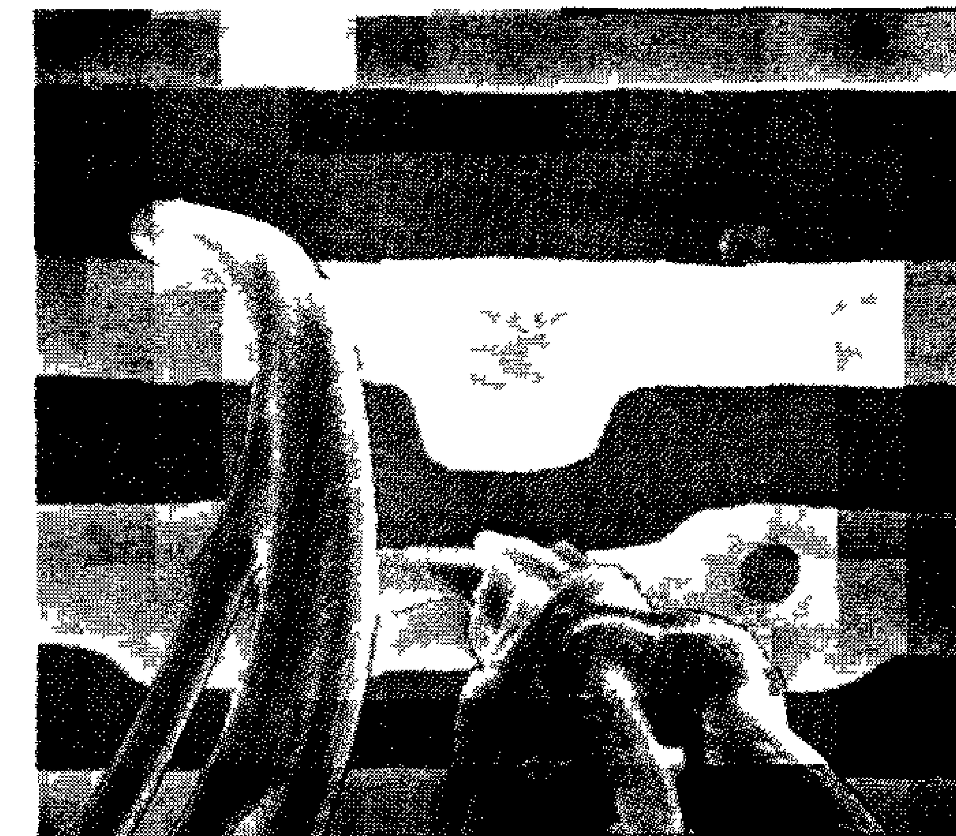







Figure 8: *From a wafer to on-chip connections: the details are comparable with those of an insect (photo: Texas Instruments/Koning & Hartman)*

Overview of symbols

Symbol	Description
α	channel-shortening factor or clustering factor
A	area
A	aspect ratio
a	activity factor
β	MOS transistor gain factor
$\beta_{\square}$	gain factor for MOS transistor with square channel
β_{n}	nMOS transistor gain factor
β_{p}	pMOS transistor gain factor
β_{total}	equivalent gain factor for a combination of transistors
BV	breakdown voltage
C	capacitance
C_{b}	bitline capacitance
C_{d}	depletion layer capacitance
C_{db}	drain-substrate capacitance
C_{g}	gate capacitance
C_{gb}	gate-substrate capacitance
C_{gd}	gate-drain capacitance
C_{gs}	gate-source capacitance
C_{gdo}	voltage-independent gate-drain capacitance
C_{gso}	voltage-independent gate-source capacitance
C_{par}	parasitic capacitance
C_{min}	minimum capacitance
C_{s}'	scaled capacitance
C_{ox}	oxide capacitance
C_{s}	silicon surface-interior capacitance
C_{sb}	source-substrate (source-bulk) voltage
C_{t}	total capacitance
ΔL	difference between drawn and effective channel length
ΔV_{T}	threshold voltage variation
D_0	defect density for uniformly distributed errors (dust particles)
D_{l}	threshold-voltage channel-length dependence factor

D_w threshold-voltage channel-width dependence factor
ϵ dielectric constant
ϵ_0 absolute permeability
ϵ_{ox} relative permeability of oxide
ϵ_r relative permeability
ϵ_{si} relative permeability of silicon
E electric field strength
E_c conduction band energy level
E_f Fermi level
E_i intrinsic (Fermi) level
E_{mx} maximum horizontal electric field strength
E_v valence band energy level
E_x horizontal electric field strength
E_{x_c} critical horizontal field strength
E_z vertical electric field strength
ϕ electric potential
ϕ_f Fermi potential
ϕ_s surface potential of silicon w.r.t. the substrate interior
ϕ_{MS} contact potential between gate and substrate
f clock frequency
f_{max} maximum clock frequency
γ factor which expresses relationship between drain-source voltage and threshold-voltage variation
g_m transconductance
I current
I_b substrate current
I_{ds} drain-source current
I_{ds0} characteristic sub-threshold current for gate-substrate voltage of 0 V
I_{dsD} driver transistor drain-source current
I_{dsL} load transistor drain-source current
I_{dssat} saturated transistor drain-source current
I_{dssub} sub-threshold drain-source current
I_{max} maximum current
I_R current through resistance
$i(t)$ time-dependent current
j current densisty
k Boltzman's constant
K K-factor; expresses relationship between source-substrate voltage and threshold voltage
K amplification factor
λ wavelength of light
L effective transistor channel length and inductance
L_{ref} effective channel length of reference transistor
M yield model parameter
μ_0 substrate carrier mobility
μ_n channel electron mobility
μ_p channel hole mobility
N_A substrate doping concentration
N.A. numeric aperture
ρ charge density
P power dissipation
P_{dyn} dynamic power dissipation
P_{stat} static power dissipation
p voltage scaling factor
Q charge
q elementary charge of a single electron
Q_d depletion layer charge
Q_g gate charge
Q_m total mobile charge in the inversion layer
Q_n mobile charge per unit area in the channel
Q_{ox} oxide charge
Q_s total charge in the semiconductor
R resistance
R_L load resistance
R_{out} output resistance or channel resistance
R_{therm} thermal resistance of a package
r tapering factor
s scale factor
τ delay time
τ_f fall time
τ_r rise time
τ_R dielectric relaxation time
T clock period
T_{min} minimum clock period
$Temp$ temperature
T_{lf} transistor lifetime
t time
t_{cond} conductor thickness

t_{is} isolator thickness
t_{ox} gate-oxide thickness
U computing power
v carrier velocity
v_{sat} carrier saturation velocity
V voltage
V_B breakdown voltage
V_r scaled voltage
V_0 depletion layer voltage
V_{bb} substrate voltage
V_{dd} supply voltage
V_c voltage at silicon surface
V_{ds} drain-source voltage
V_{dssat} drain-source voltage of saturated transistor
V_E Early voltage
V_{fb} flat-band voltage
V_g gate voltage
V_{gg} extra supply voltage
V_{gs} gate-source voltage
V_{gs_L} load transistor gate-source voltage
V_H high voltage level
V_{in} input voltage
V_j junction voltage
V_L low voltage level
V_{PT} transistor punch-through voltage
V_{sb} source-substrate (back-bias) voltage
V_{ss} ground voltage
V_{ws} well-source voltage
V_T threshold voltage
V_{T_D} driver transistor threshold voltage
$V_{T_{dep}}$ depletion transistor threshold voltage
$V_{T_{enh}}$ enhancement transistor threshold voltage
V_{T_L} load transistor threshold voltage
V_{T_n} nMOS transistor threshold voltage
V_{T_p} pMOS transistor threshold voltage
V_{Tpar} parasitic transistor threshold voltage
V_{out} output voltage
$V(x)$ potential at position x
V_x process-dependent threshold voltage term
V_{X_L} process-dependent threshold voltage term for load transistor
V_{X_D} process-dependent threshold voltage term for driver transistor
W transistor channel width
W_n nMOS transistor channel width
W_p pMOS transistor channel width
W_{ref} reference transistor channel width
$\frac{W}{L}$ transistor aspect ratio
$\left(\frac{W}{L}\right)_n$ nMOS transistor aspect ratio
$\left(\frac{W}{L}\right)_p$ pMOS transistor aspect ratio
x distance w.r.t. specific reference point
Y yield
Z_i input impedance

List of physical constants

$\epsilon_0 = 8.85 \times 10^{-14}$ F/cm
$\epsilon_{ox} = 4$ for silicon dioxide
$\epsilon_{si} = 11.7$
$\phi_f = 0.32$ V for silicon substrate
$k = 1.4 \times 10^{-23}$ Joule/K
$q = 1.6 \times 10^{-19}$ Coulomb

Contents

Foreword v

Preface vi

Overview of symbols xviii

List of physical constants xxiii

1 Basic Principles 1
- 1.1 Introduction 1
- 1.2 The field-effect principle 1
- 1.3 The inversion-layer MOS transistor 4
  - 1.3.1 The Metal-Oxide-Semiconductor (MOS) capacitor 10
  - 1.3.2 The inversion-layer MOS transistor 14
- 1.4 Derivation of simple MOS formulae 22
- 1.5 The back-bias effect (back-gate effect, body effect) 26
- 1.6 Factors which characterise the behaviour of the MOS transistor 29
- 1.7 Different types of MOS transistors 30
- 1.8 Parasitic MOS transistors 32
- 1.9 MOS transistor symbols 34
- 1.10 Capacitances in MOS structures 36
- 1.11 Conclusions 46
- 1.12 References 47
- 1.13 Exercises 48

2 Physical and geometrical effects on the behaviour of the MOS transistor 53
- 2.1 Introduction 53
- 2.2 The zero field mobility 54
- 2.3 Carrier mobility degradation 55
  - 2.3.1 Temperature-dependent carrier mobility reduction 55
  - 2.3.2 Vertical and lateral field carrier mobility degradation 55
- 2.4 Channel length modulation and static drain feedback 57
  - 2.4.1 Channel length modulation 58
  - 2.4.2 Static drain feedback 59
  - 2.4.3 The Early voltage 60
- 2.5 Small-channel effects 61
  - 2.5.1 Short-channel effect 61
  - 2.5.2 Narrow-channel effect 63
  - 2.5.3 Modelling small-channel effects 65
- 2.6 Punch-through 66
- 2.7 Hot-carrier effect 67
  - 2.7.1 Introduction 67
  - 2.7.2 The electric field in MOS transistors 67
  - 2.7.3 Impact ionisation 69
  - 2.7.4 Hot-carrier degradation 69
  - 2.7.5 Reducing the maximum electric field in a MOS transistor 70
- 2.8 Weak-inversion behaviour of the MOS transistor 73
- 2.9 Conclusions 76
- 2.10 References 77
- 2.11 Exercises 79

3 Manufacture of MOS devices 81
- 3.1 Introduction 81
- 3.2 Lithography in MOS processes 82
- 3.3 Etching 89
- 3.4 Thermal oxidation 91
- 3.5 Deposition 94
- 3.6 Diffusion and ion implantation 97
- 3.7 Planarisation 100
- 3.8 Basic MOS technologies 107
  - 3.8.1 The basic silicon-gate nMOS process 107
  - 3.8.2 The basic Complementary MOS (CMOS) process 112
  - 3.8.3 An advanced deep-submicron CMOS process 114
  - 3.8.4 Silicon-on-insulator CMOS (SOI-CMOS) process 122
- 3.9 Conclusions 125

3.10 References 126
3.11 Exercises 128

4 CMOS circuits **129**
4.1 Introduction 129
4.2 The basic nMOS inverter 130
4.2.1 Introduction 130
4.2.2 The DC behaviour 132
4.2.3 The transient response 140
4.2.4 Transforming a logic function into an nMOS transistor circuit 146
4.3 Electrical design of CMOS circuits 149
4.3.1 Introduction 149
4.3.2 The CMOS inverter 151
4.4 Digital CMOS circuits 167
4.4.1 Introduction 167
4.4.2 Static CMOS circuits 167
4.4.3 Clocked static CMOS circuits 173
4.4.4 Dynamic CMOS circuits 176
4.4.5 Other types of CMOS circuit 182
4.4.6 Choosing a CMOS implementation 183
4.4.7 Clocking strategies 184
4.5 CMOS input and output (I/O) circuits 184
4.5.1 CMOS input circuits 185
4.5.2 CMOS output buffers (drivers) 186
4.6 The layout process 187
4.6.1 Introduction 187
4.6.2 Layout design rules 188
4.6.3 Stick diagram 192
4.6.4 Example of the layout procedure 195
4.6.5 Guidelines for layout design 199
4.7 Conclusions 201
4.8 References 202
4.9 Exercises 204

5 Special circuits, devices and technologies **209**
5.1 Introduction 209
5.2 CCD and CMOS image sensors 210
5.2.1 Introduction 210
5.2.2 Basic CCD operation 210
5.2.3 CMOS image sensors 214
5.3 Power MOSFET transistors 217
5.3.1 Introduction 217
5.3.2 Technology and operation 219
5.3.3 Applications 220
5.4 BICMOS digital circuits 220
5.4.1 Introduction 220
5.4.2 BICMOS technology 221
5.4.3 BICMOS characteristics 222
5.4.4 BICMOS circuit performance 223
5.4.5 Future expectations and market trends 226
5.5 Conclusions 227
5.6 References 228
5.7 Exercises 230

6 Memories **231**
6.1 Introduction 231
6.2 Serial memories 234
6.3 Random-access memories (RAM) 235
6.3.1 Introduction 235
6.3.2 Static RAMs (SRAM) 235
6.3.3 Dynamic RAMs (DRAM) 242
6.3.4 High-performance DRAMs 247
6.3.5 Error sensitivity 251
6.3.6 Redundancy 252
6.4 Non-volatile memories 252
6.4.1 Introduction 252
6.4.2 Ferroelectric RAM (FRAM) 252
6.4.3 Read-Only Memories (ROM) 254
6.4.4 Programmable Read-Only Memories 258
6.4.5 EEPROMs and flash memories 260
6.4.6 Non-volatile RAM (NVRAM) 263
6.4.7 BRAM (battery RAM) 263
6.5 Embedded memories 264
6.6 Classification of the various memories 267
6.7 Conclusions 268
6.8 References 269
6.9 Exercises 271

7 Very Large Scale Integration (VLSI) and ASICs 273
7.1 Introduction . . . 273
7.2 Digital ICs . . . 275
7.3 Abstraction levels for VLSI . . . 279
7.3.1 Introduction . . . 279
7.3.2 System level . . . 281
7.3.3 Functional level . . . 282
7.3.4 RTL level . . . 283
7.3.5 Logic-gate level . . . 286
7.3.6 Transistor level . . . 287
7.3.7 Layout level . . . 288
7.3.8 Conclusions . . . 288
7.4 Digital VLSI design . . . 291
7.4.1 Introduction . . . 291
7.4.2 The design flow . . . 291
7.4.3 Example of synthesis from VHDL description to layout . . . 294
7.5 The use of ASICs . . . 300
7.6 Silicon realisation of VLSI and ASICs . . . 302
7.6.1 Introduction . . . 302
7.6.2 Handcrafted layout implementation . . . 303
7.6.3 Bit-slice layout implementation . . . 305
7.6.4 ROM, PAL and PLA layout implementations . . . 306
7.6.5 Cell-based layout implementation . . . 310
7.6.6 Gate array layout implementation . . . 312
7.6.7 Programmable Logic Devices (PLDs) . . . 315
7.6.8 Hierarchical design approach . . . 326
7.6.9 The choice of a layout implementation form . . . 328
7.7 Conclusions . . . 331
7.8 References . . . 332
7.9 Exercises . . . 335

8 Low power, a hot topic in IC design 337
8.1 Introduction . . . 337
8.2 Sources of CMOS power consumption . . . 338
8.3 Technology options for low power . . . 340
8.3.1 Reduction of P_{leak} by technological measures . . . 340
8.3.2 Reduction of P_{dyn} by technology measures . . . 343
8.3.3 Reduction of P_{dyn} by reduced-voltage processes . . . 344
8.4 Design options for low power . . . 345
8.4.1 Reduction of P_{short} by design measures . . . 345
8.4.2 Reduction/elimination of P_{stat} by design measures 347
8.4.3 Reduction of P_{dyn} by design measures . . . 347
8.5 Computing power versus chip power, a scaling perspective 378
8.6 Conclusions . . . 381
8.7 References . . . 382
8.8 Exercises . . . 384

9 Circuit reliability and signal integrity in deep-submicron designs 385
9.1 Introduction . . . 385
9.2 Design for reliability . . . 386
9.2.1 Introduction . . . 386
9.2.2 Latch-up in CMOS circuits . . . 386
9.2.3 Electrostatic discharge (ESD) and its protection . 390
9.2.4 Electromigration . . . 396
9.2.5 Hot-carrier degradation . . . 398
9.3 Design for signal integrity . . . 398
9.3.1 Introduction . . . 398
9.3.2 Clock distribution and critical timing issues . . . 398
9.3.3 Clock generation and synchronisation in different (clock) domains on a chip . . . 408
9.3.4 Phenomena related to large current fluctuations . 412
9.3.5 The influence of the interconnection (metallisation and dielectrics) . . . 426
9.3.6 Design organisation . . . 436
9.4 Conclusions . . . 438
9.5 References . . . 438
9.6 Exercises . . . 440

10 Testing, debugging, yield and packaging 441
10.1 Introduction . . . 441
10.2 Testing . . . 442
10.3 Yield . . . 448
10.4 Packaging . . . 453
10.4.1 Introduction . . . 453
10.4.2 Package categories . . . 454
10.4.3 Die attachment and bonding techniques . . . 458
10.4.4 Advances in IC packaging technology . . . 466
10.5 Quality and reliability of packaged dies . . . 475

10.5.1 Quality 475
10.5.2 Reliability 475
10.6 Potential first silicon problems 477
10.6.1 Problems with testing 477
10.6.2 Problems caused by marginal or out-of-specification processing 479
10.6.3 Problems caused by marginal design 481
10.7 First-silicon debug and failure analysis 482
10.7.1 Introduction 482
10.7.2 I_{ddq} testing 482
10.7.3 Diagnosis via Shmoo plots 483
10.7.4 Diagnosis via picoprobing 486
10.7.5 Diagnosis with liquid crystal techniques 490
10.7.6 Diagnosis by photon emission microscopy (PEM) . 492
10.7.7 Diagnosis by electron beam techniques 493
10.7.8 Alternative failure analysis techniques 494
10.7.9 Repair 495
10.7.10 Design for Failure Analysis and Design for Debug . 496
10.8 Conclusions 497
10.9 References 498
10.10 Exercises 502

11 Effects of scaling on MOS IC design and consequences for the roadmap 503
11.1 Introduction 503
11.2 Transistor scaling effects 505
11.3 Interconnection scaling effects 506
11.4 Scaling consequences for overall IC design 510
11.4.1 Scaling consequences for overall chip performance . 510
11.4.2 Scaling consequences for overall design reliability . 512
11.4.3 Scaling consequences for overall signal integrity . . 514
11.5 Potential limitations of the pace of scaling 519
11.6 Conclusions 523
11.7 References 524
11.8 Exercises 525

Index 526

Chapter 1

Basic Principles

1.1 Introduction

The majority of current *VLSI* (Very Large Scale Integration) circuits are manufactured in CMOS technologies. Familiar examples are memories (64 Mbit, 256 Mbit and 1 Gbit), microprocessors and signal processors. A good fundamental treatment of basic MOS devices is therefore essential for an understanding of the design and manufacture of modern *VLSI* circuits. This chapter describes the operation and characteristics of MOS devices. The material requirements for their realisation are discussed and equations that predict their behaviour are derived.

The acronym MOS represents the Metal, Oxide and Semiconductor materials used to realise early versions of the MOS transistor. The fundamental basis for the operation of MOS transistors is the field-effect principle. This principle is quite old, with related publications first appearing in the nineteen-thirties. These include a patent application filed by J.E. Lilienfeld in Canada and the USA in 1930 and one filed by O. Heil, independently of Lilienfeld, in England in 1935. At that time, however, insufficient knowledge of material properties resulted in devices which were unfit for use. The rapid development of electronic valves probably also hindered the development of the MOS transistor by largely fulfilling the transistor's envisaged role.

1.2 The field-effect principle

The field-effect principle is explained with the aid of figure 1.1. This figure shows a rectangular conductor, called a channel, with length L, width W and thickness t_{cond}. The free electrons present in the channel are the mobile charge carriers. There are n electrons per m^3 and the charge q per electron equals -1.602×10^{-19} C(coulomb). The application of a horizontal electric field of magnitude E to the channel causes the electrons to acquire an average velocity $v = -\mu_n \cdot E$. The electron mobility μ_n is positive. The direction of v therefore opposes the direction of E. The resulting current density j is the product of the average electron velocity and the mobile charge density ρ:

$$j = \rho \cdot v = -n \cdot q \cdot \mu_n \cdot E \tag{1.1}$$

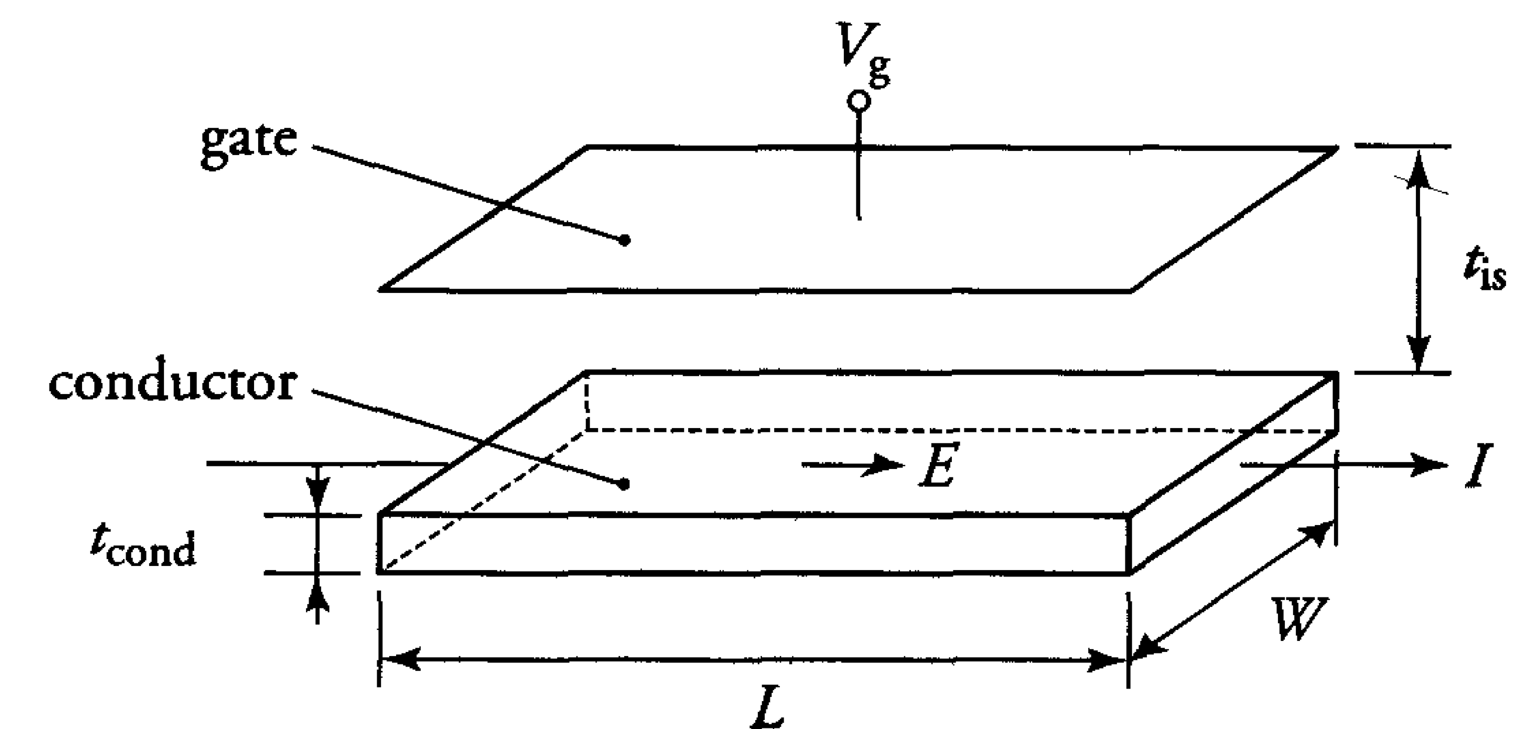


Figure 1.1: *The field-effect principle*

A gate electrode situated above the channel is separated from it by an insulator of thickness t_{is}. A change in the gate voltage V_g influences the charge density ρ in the channel. The current density j is therefore determined by V_g.

Example:
Suppose the insulator is silicon dioxide (SiO_2) with a thickness of 5 nm ($t_{is} = 5 \times 10^{-9}$m). The gate capacitance will then be about 7 mF/m^2. The total gate capacitance C_g is therefore expressed as follows:

$$C_g = 7 \times 10^{-3} \cdot W \cdot L \quad [\text{F}]$$

A change in gate charge $\Delta Q_g = -C_g \cdot \Delta V_g$ causes the following change in channel charge:

$$+C_g \cdot \Delta V_g = 7 \times 10^{-3} \cdot W \cdot L \cdot \Delta V_g \;=\; W \cdot L \cdot t_{\text{cond}} \cdot \Delta\rho$$

Thus:

$$\Delta\rho = \frac{7 \times 10^{-3} \cdot \Delta V_g}{t_{cond}} \; \mathrm{C/m^3}$$

and:

$$|\Delta n| = |\frac{\Delta\rho}{q}| = \frac{4.4 \times 10^{16} \cdot \Delta V_g}{t_{cond}} \; \mathrm{electrons/m^3}$$

If a 1 V change in gate voltage is to cause a tenfold increase in current density j, then the following must apply:

$$\frac{\Delta j}{j} = \frac{\Delta\rho}{\rho} = \frac{\Delta n}{n} = \frac{4.4 \times 10^{16}}{t_{cond} \cdot n} = 10$$

$$\Rightarrow t_{cond} = \frac{4.4 \times 10^{15}}{n}$$

Examination of two materials reveals the implications of this expression for t_{cond}:

Case a The channel material is copper.
This has $n \approx 10^{28}$ electrons/m^3 and hence $t_{cond} \approx 4.4 \times 10^{-13}$ m. The required channel thickness is thus less than the size of one atom ($\approx 3 \times 10^{-10}$ m). This is impossible to realise and its excessive number of free carriers renders copper unsuitable as channel material.

Case b The channel material is 5Ωcm n-type silicon.
This has $n \approx 10^{21}$ electrons/m^3 and hence $t_{cond} \approx 4.4\,\mu$m.

The *transconductance* g_m of a MOS transistor is the ratio of a change in channel (drain) current to the corresponding change in gate voltage:

$$g_m = \frac{\Delta I}{\Delta V_g}$$

$$\text{However} \quad \frac{\Delta I}{I} = \frac{\Delta j}{j}$$

$$\text{Therefore} \quad g_m = \frac{I}{\Delta V_g} \cdot \frac{\Delta j}{j}$$

If $I = j \cdot W \cdot t_{cond} = 1$ mA, $\Delta j / j = 10$ and $\Delta V_g = 1$ V then:

$$g_m = 10 \; \mathrm{mA/V}$$

In this case, a transconductance of 10 mA/V requires a channel thickness of $t_{cond} = 4.4\,\mu$m. Modern IC technologies allow the realisation of much thinner channels.

From the above example, it is clear that field-effect devices can only be realised with semiconductor materials. Aware of this fact, Lilienfeld used copper sulphide as a semiconductor in 1930. Germanium was used during the early fifties. Until 1960, however, usable MOS transistors could not be manufactured. Unlike the transistor channel, which comprised a manufactured thin layer, the channel in these *inversion-layer transistors* is a thin conductive layer, which is realised electrically. The breakthrough for the fast development of MOS transistors came with advances in planar silicon technology and the accompanying research into the physical phenomena in the semiconductor surface.

Generally, circuits are integrated in silicon because widely-accepted military specifications can be met with this material. These specifications require products to function correctly at a maximum operating temperature of 125 °C. The maximum operating temperature of germanium is only 70 °C, while that of silicon is 150 °C. A comparison of a few other germanium (Ge) and silicon (Si) material constants is presented below:

Material constant	Germanium	Silicon
Melting point [°C]	937	1415
Breakdown field [V/μm]	8	30
Relative expansion coeff. [°C]$^{-1}$	5.8×10^{-6}	2.5×10^{-6}
ϵ_r	16.8	11.7
Max. operating temp. [°C]	70	150

1.3 The inversion-layer MOS transistor

A schematic drawing of the inversion-layer nMOS transistor, or simply 'nMOSt', is shown in figure 1.2, which is used to explain its structure and operation. The two n$^+$ areas in the p-type substrate are called the *source* and *drain*. The *gate* electrode is situated above the p area between them. This electrode is either a metal plate, e.g. aluminium or molybdenum, or a heavily doped and thus low-ohmic polycrystalline silicon layer. Normally, the source and drain areas are also heavily doped to minimise series resistance. The resistance R of a 10 μm long and 2 μm

wide track is $\frac{10}{2} \cdot R_{\square}$, where $R_{\square}$ is the sheet resistance of the track material. The sheet resistance of the source and drain areas usually ranges from 3 to 100 $\Omega/\square$. The dope concentration in the p-type substrate is approximately $10^{20} - 10^{22}$ atoms per m^3, while the channel dope (by threshold adjustment implantation, etc.) is between $10^{23} - 10^{24}$ atoms per m^3. A p-channel transistor differs from the above n-channel type in that it contains a p^+ source and drain in an n-type substrate.

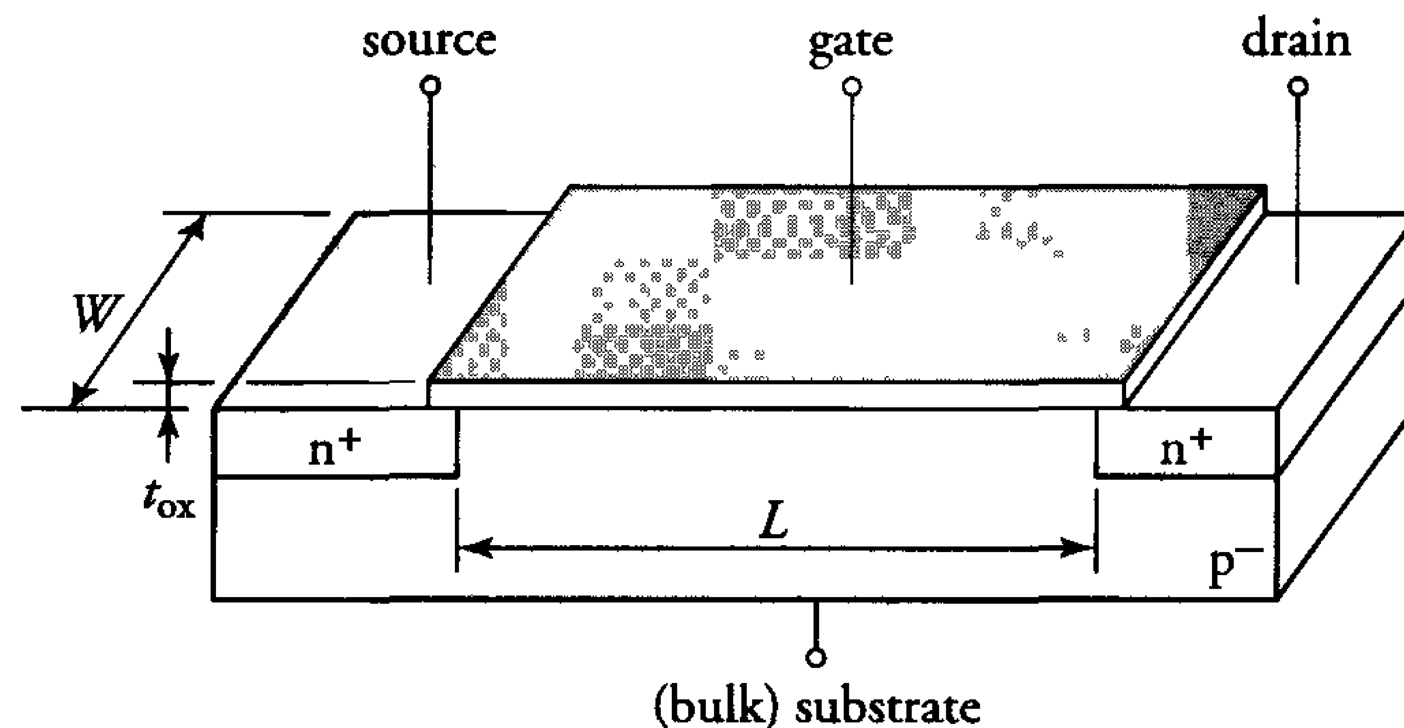


Figure 1.2: *Cross-section of an inversion-layer nMOS transistor*

Characteristic parameters of a MOS transistor are indicated in figure 1.2. These include the width W and length L of the channel and the thickness t_{ox} of the insulating oxide which separates the gate and channel. In modern CMOS VLSI circuits, the minimum values of W and L range from 0.12 μm to 0.5 μm and $t_{ox} \approx 2$ nm - 10 nm. Continuous development will reduce these values in the future. The depth of the source and drain junctions varies from 0.05 μm to 0.3 μm.

The energy band theory and its application to the MOS transistor are briefly summarised below. An understanding of this summary is a pre-requisite for a detailed discussion of the behaviour of the MOS transistor.

The structure of a free silicon atom is shown in figure 1.3. This atom comprises a nucleus, an inner shell and an outer shell. The nucleus contains 14 protons and 14 neutrons while the shells contain 14 electrons. Ten of the electrons are in the inner shell and four are in the outer shell. The positive charge of the protons and the negative charge of the electrons compensate each other to produce an atom with a net neutral charge.

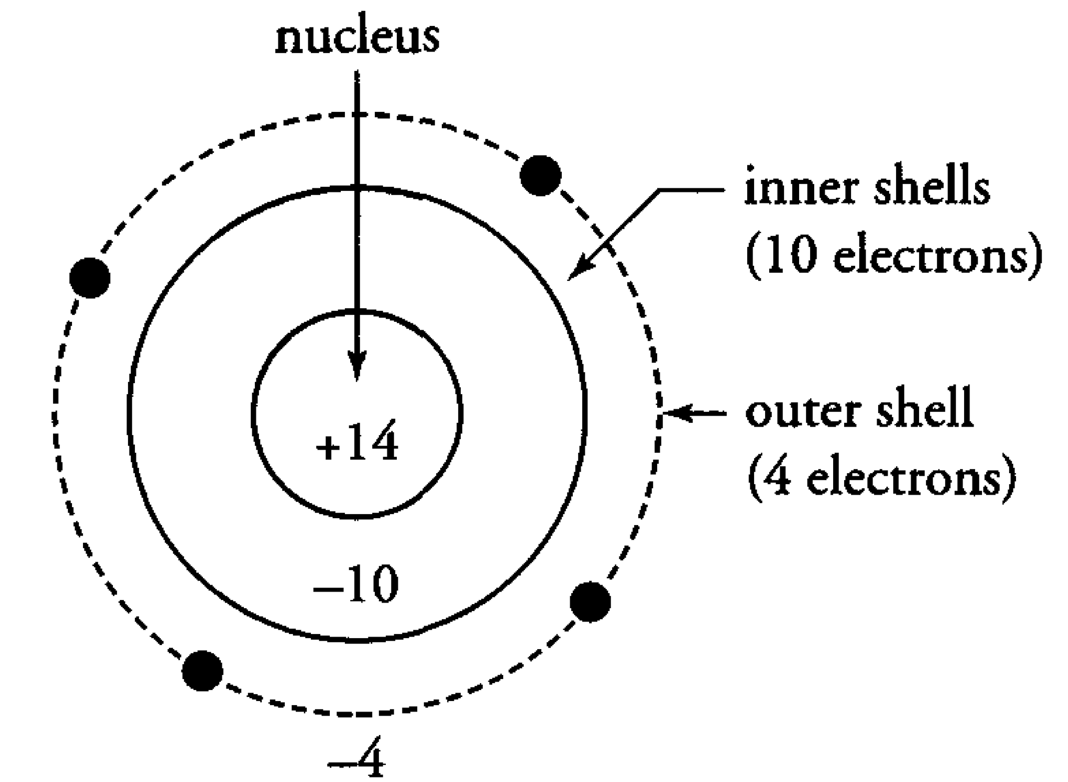


Figure 1.3: *The structure of a free silicon atom*

The electrons in an atom may possess certain energy levels. These energy levels are grouped into energy bands, which are separated by energy gaps. An energy gap represents impossible levels of electron energy. The energy bands that apply to the electrons in an atom's outer shell are *valence* and *conduction* bands. Figure 1.4 shows these bands and the energy gap for a typical solid material. The valence electrons determine the physical and chemical properties of a material.

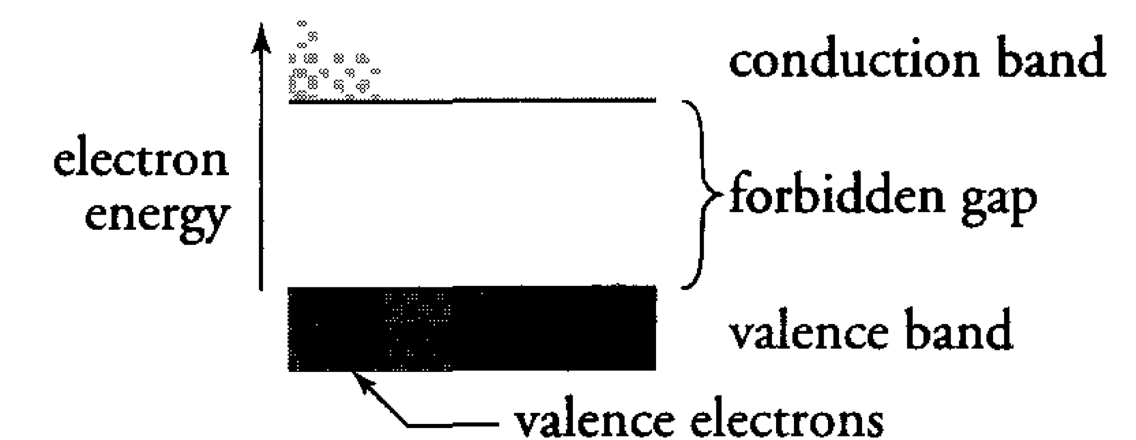


Figure 1.4: *Schematic representation of electron energy bands in a typical solid material*

The four electrons in the outer shell of a silicon atom are in the material's valence band. Figure 1.5 shows the bonds that these electrons form with neighbouring atoms to yield a silicon crystal.

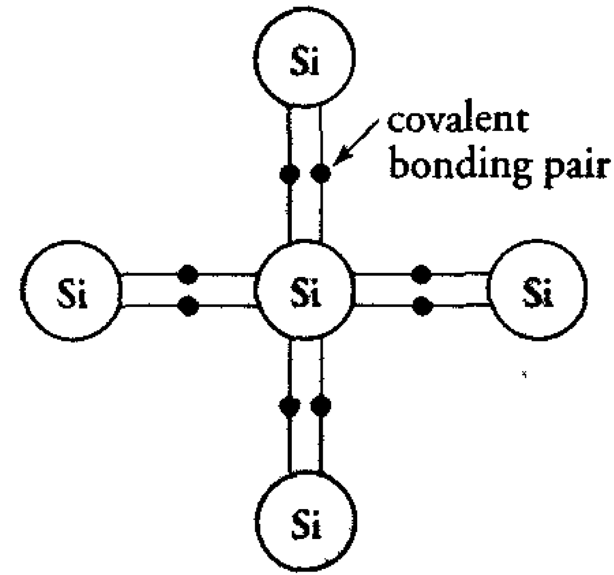


Figure 1.5: *Silicon crystal*

The electrons in a conductor can easily go from the valence band to the conduction band. Therefore, the conduction and valence bands in a conductor partly overlap, as shown in figure 1.6a. In an insulator, however, none of the valence electrons can reach the conduction band. Figure 1.6b shows the large band gap generally associated with insulators. A semiconductor lies somewhere between a conductor and an insulator. The associated small band gap is shown in figure 1.6c. Valence electrons may acquire sufficient thermal energy to reach the conduction band and therefore leave an equal number of positively-charged ions, or 'holes', in the valence band. This produces a limited conduction mechanism in semiconductors.

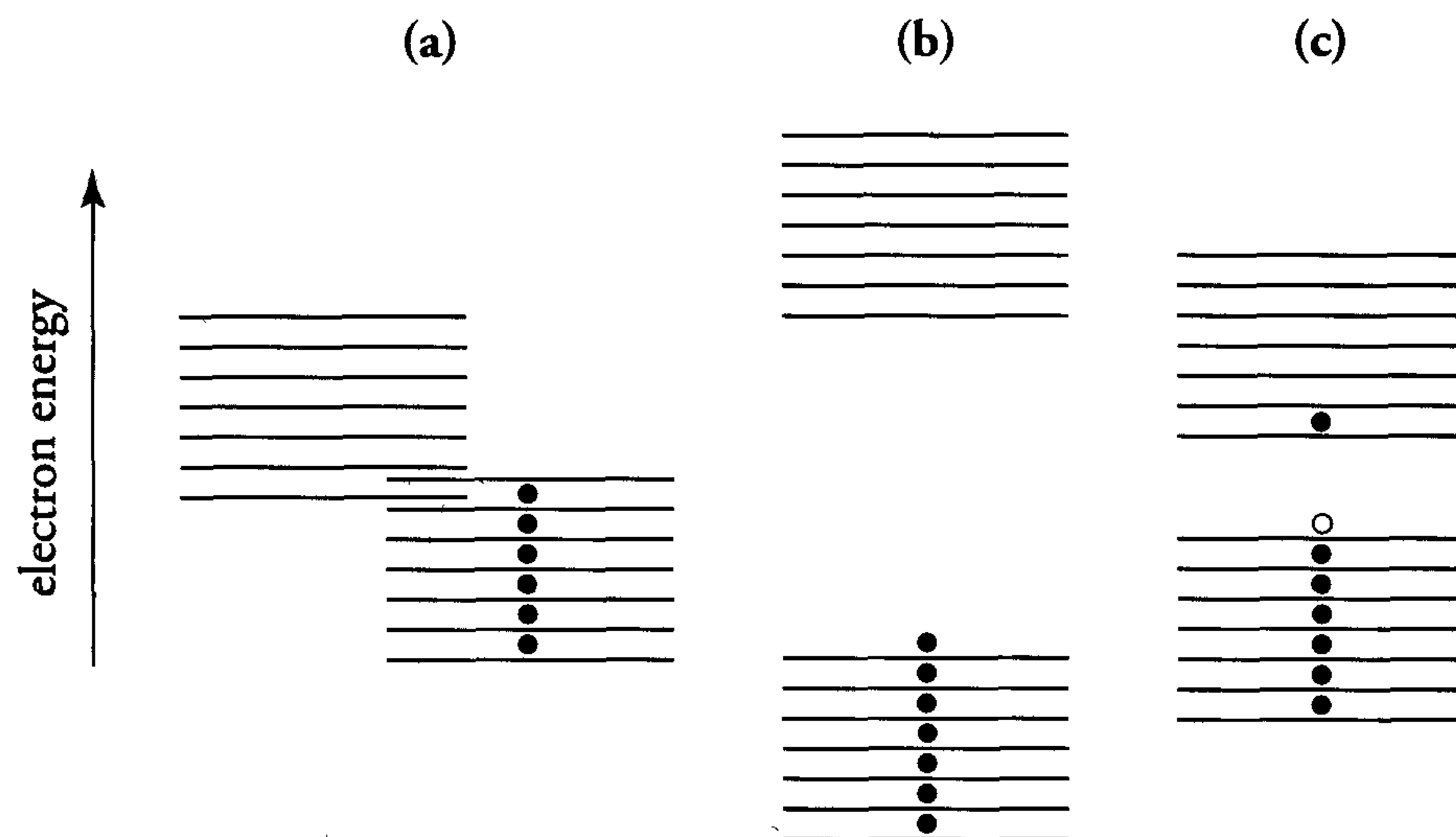


Figure 1.6: *Energy bands of a conductor, an insulator and an intrinsic semiconductor*

Semiconductor materials are located in group IV of this system. The introduction of an element from group III or V in a semiconductor crystal produces an 'acceptor' or a 'donor' atom. This *semiconductor doping process* dramatically changes the crystal properties. The following table shows the location of semiconductor materials in the periodic system of elements.

Group		
III (Acceptors)	IV	V (Donors)
Boron	Carbon	Nitrogen
Aluminium	Silicon	Phosphorous
Gallium	Germanium	Arsenic
Indium	Stannic (tin)	Stibnite

The presence of a group III atom in a silicon crystal lattice is considered first. The situation for boron is illustrated in figure 1.7a. Boron has one electron less than silicon and cannot therefore provide an electron required for a bond with one of its four neighbouring silicon atoms. The hole in the resulting p-type semiconductor is a willing 'acceptor' for an electron from an alternative source. This hole can be removed relatively easily with the ionisation energy of approximately 0.045 eV shown in the energy band diagram of figure 1.7a.

Similar reasoning applies when a group V atom, such as phosphorus, is present in the silicon lattice. This situation is illustrated in figure 1.7c. The extra electron in the phosphorus atom cannot be accommodated in the regular bonding structure of the silicon lattice. It is therefore easy to remove this 'donor' electron in the resulting n-type semiconductor. The mere 0.037 eV ionisation energy required is much lower than the 1.11 eV band gap energy of silicon. Figure 1.7b shows the energy band diagram of an intrinsic silicon lattice, which contains no donor or acceptor 'impurity' atoms.

The energy level indicated by E_f in figure 1.7 is called the *Fermi level.* An electron with this energy has an equal probability of location in the valence band and the conduction band. This probability is exactly 0.5.

The Fermi level of an intrinsic semiconductor is often referred to as the intrinsic Fermi level E_i. The Fermi level E_f in a p-type semiconductor is situated near the valence band E_v, while it is close to the conduction band E_c in an n-type semiconductor. The above theory concerning the different types of semiconductors and their respective energy

band diagrams will now be used to explain the behaviour of the MOS transistor. This explanation is preceded by a description of the structure and operation of the MOS capacitor.

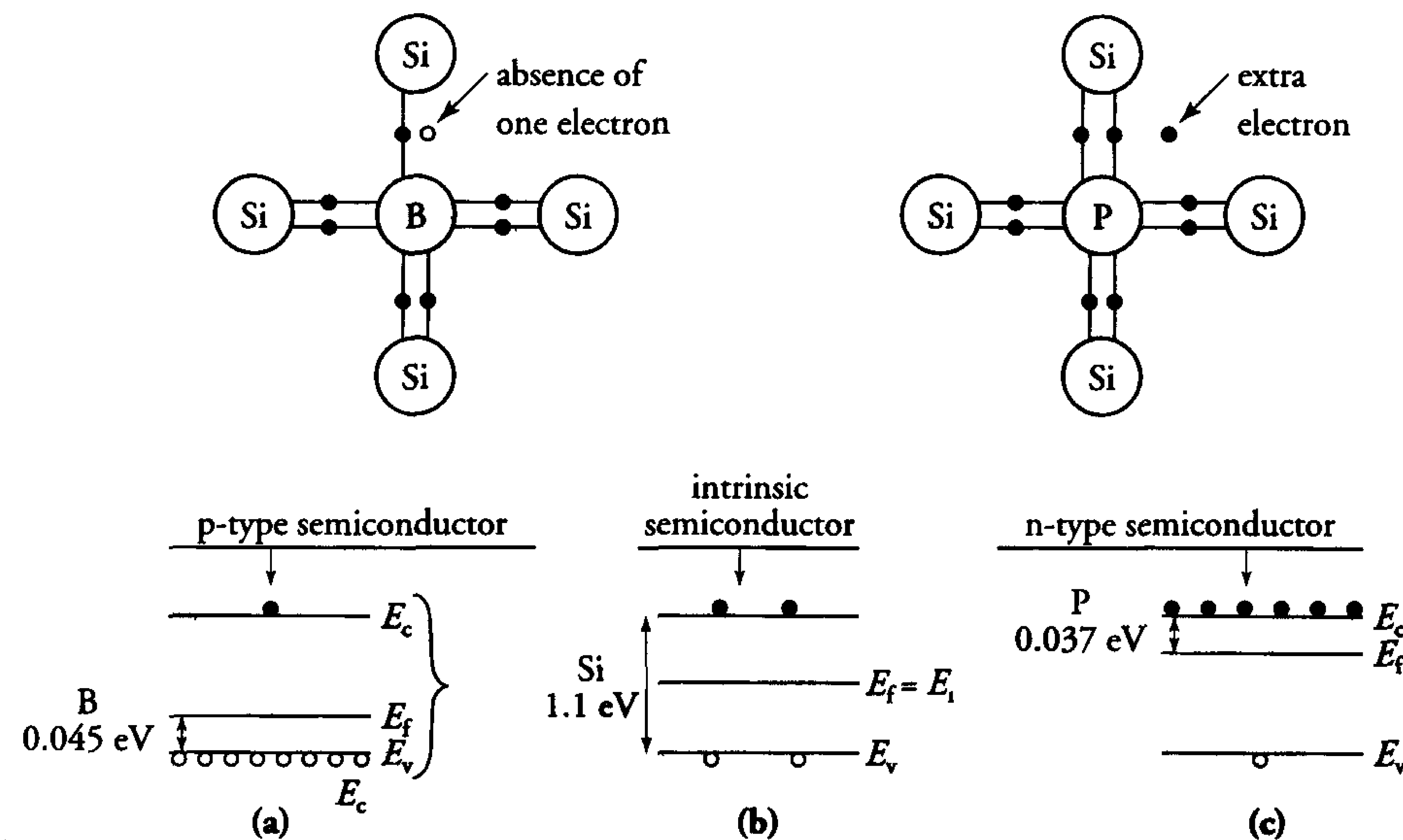


Figure 1.7: *Energy band diagrams for p-type, intrinsic and n-type semiconductor materials*

1.3.1 The Metal-Oxide-Semiconductor (MOS) capacitor

Figure 1.8 shows a cross-section of a basic MOS capacitor. This structure is identical to a MOS transistor except that the source and drain diffusion regions are omitted.

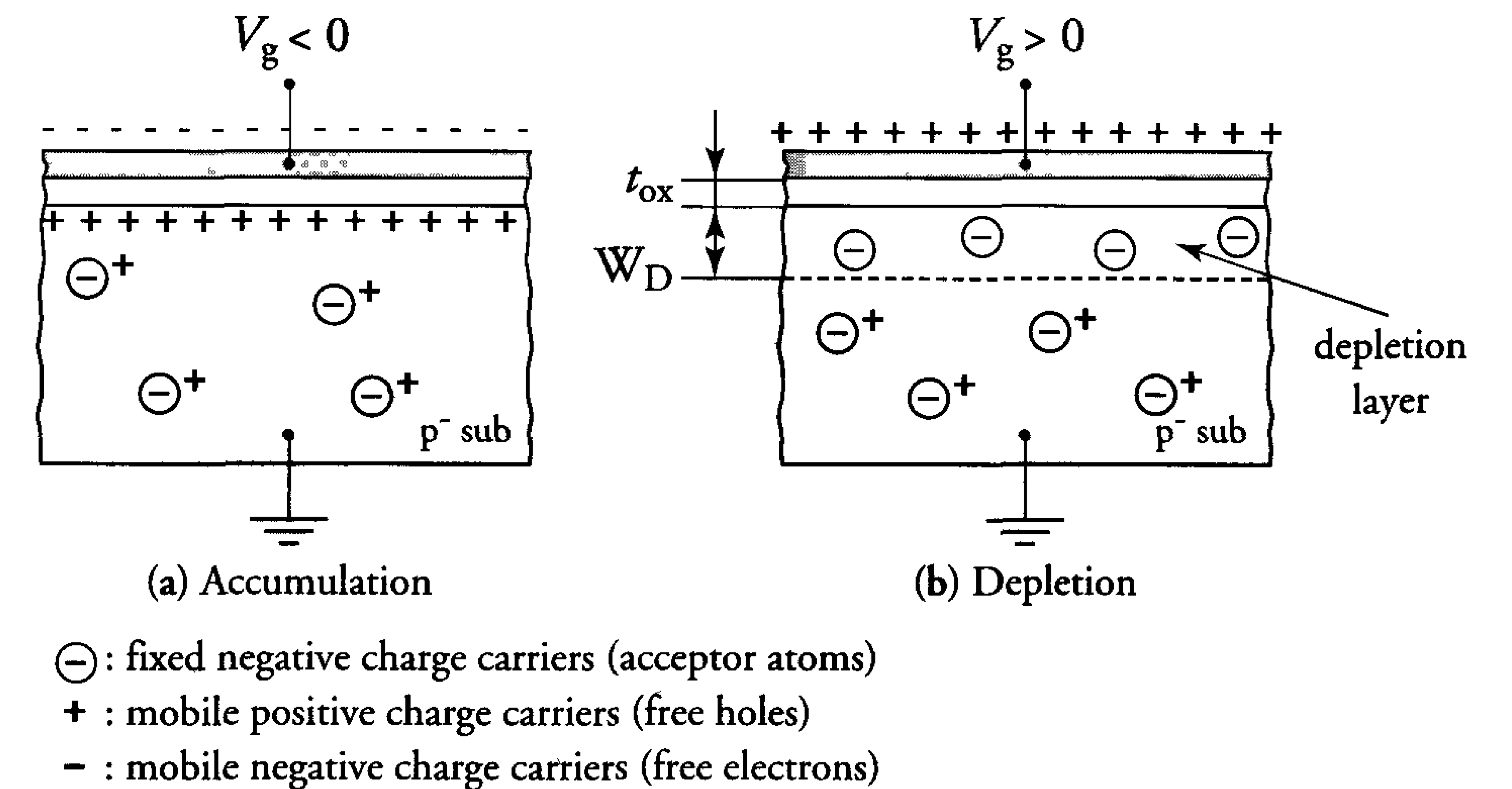


Figure 1.8: *Cross-section of a MOS structure without source and drain areas. There is a capacitance between the gate and substrate.*

The p-type substrate is made with an acceptor dope material, e.g. boron. The substrate is assumed to behave as a normal conductor and contains many free holes. The situation which occurs when the p-type substrate is grounded and a negative voltage is applied to the gate electrode is shown in figure 1.8a. The negative charge on the aluminium gate is compensated by an equal but positive charge in the substrate. This is accomplished by positively charged holes which accumulate at the Si-SiO_2 interface. These holes are the majority charge carriers in the substrate. This *'accumulation'* process continues until the positive charge at the substrate surface equals the negative charge on the gate electrode. Extra holes are supplied through the ground contact to the substrate. The resulting accumulation capacitor can be viewed as an ideal parallel plate capacitor.

A different situation occurs when the potential on the gate electrode is made positive with respect to the grounded substrate. This situation is shown in the cross-section of figure 1.8b. The positive charge which is

present on the aluminium gate must be counter-balanced by a negative charge at the Si-SiO_2 interface in the substrate. Free positively-charged holes are pushed away from the substrate surface to yield a negatively-charged depletion layer. This *'depletion'* process stops when the negative charge of the depletion layer equals the positive charge on the aluminium gate electrode. Clearly, the width of the depletion layer in the equilibrium situation is proportional to the gate voltage. It is important to realise that a depletion layer only contains a fixed charge, i.e. ions fixed in the solid state lattice, and no mobile charge carriers.

Various energy band diagrams are used to explain the behaviour of the inversion layer MOS transistor. To provide a better understanding of these diagrams, Poisson's law is first applied to the different regions of the MOS capacitor. These regions include the aluminium gate, the SiO_2 insulator, the depletion layer in silicon and the p-type silicon substrate. Poisson's law is used to investigate the charge distribution $Q(z)$, the electric field $E(z)$ and the electric potential $\phi(z)$ in these regions as a function of the distance z from the Si-SiO_2 interface.

In its one dimensional form, Poisson's law is formulated as follows:

$$\frac{d^2\phi(\mathrm{z})}{d\mathrm{z}^2} = -\frac{\rho}{\epsilon} \qquad (1.2)$$

where $\phi(\mathrm{z})$ = electrical potential at position z;
z = distance from the $Si - SiO_2$ interface;
ρ = space charge;
ϵ = dielectric constant.

The situation in which no space charge is present is considered first. This is almost true in the SiO_2 insulator, in which case $\rho = 0$. Integration of formula (1.2) once gives the electric field:

$$E(z) = C_1, \qquad C_1 = \text{integration constant.}$$

Integration of formula (1.2) twice gives the electric potential in SiO_2:

$$\phi(z) = C_1 \cdot z + C_2$$

The electric field in the insulator is thus constant and the electric potential is a linear function of the distance z from the Si-SiO_2 interface.

Next, the situation in which a constant space charge is present is considered. This is assumed to be true in the depletion layer, whose width is W_D. In this case:

$$\rho = -q \cdot N_A$$

where q = the charge of an electron
and N_A = the total number of fixed ions in the depletion layer of width W_D.

Integrating formula (1.2) once gives the electric field:

$$E(z) = \frac{q \cdot N_A}{\epsilon} \cdot z + C_1$$

Integrating formula (1.2) twice gives the electric potential in the depletion layer:

$$\phi(z) = \frac{q \cdot N_A}{2\epsilon} \cdot z^2 + C_1 \cdot z + C_2$$

Therefore, the electric field in a depletion layer with constant space charge is a linear function of z, while the electric potential is a square function of z. The space charge in a depletion layer is only constant when the dope of the substrate has a constant value at all distances z from the Si-SiO_2 interface. In practice, the space-charge profile is related to the dope profile which exists in the substrate.

The aluminium gate and the substrate region outside the depletion layer are assumed to behave as ideal conductors. The electric potentials in these regions are therefore constant and their electric fields are zero.

The above results of the application of Poisson's law to the MOS capacitor are illustrated in figure 1.9. Discontinuities in the diagrams are caused by differences between the dielectric constant of silicon and silicon dioxide. The electric charge, the electric field and potential are zero in the grounded substrate outside the depletion region. The observation that the electric potential is a square function of z in the depletion layer is particularly important.

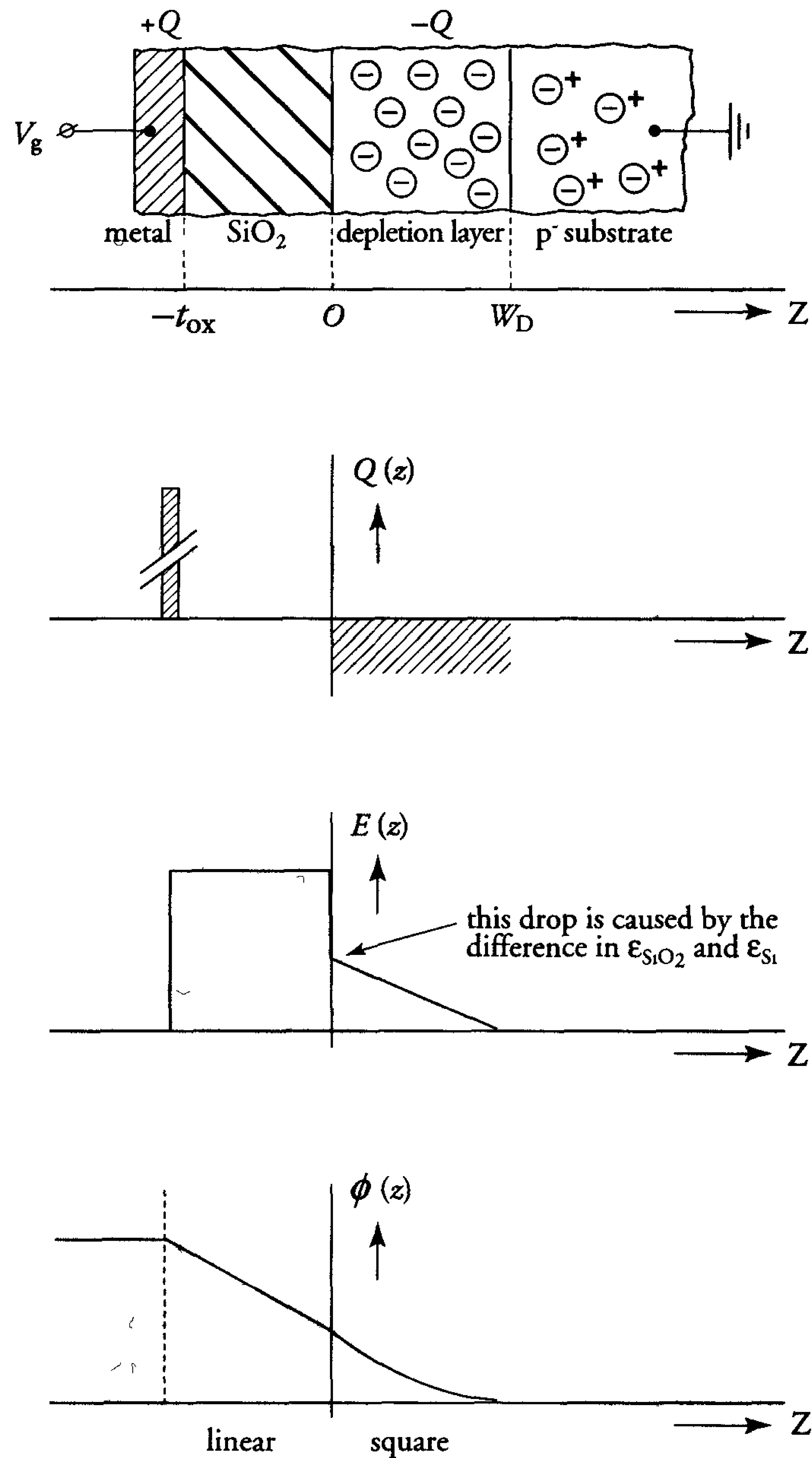


Figure 1.9: *The sections of a MOS capacitor and the associated charge distribution $Q(z)$, electric field $E(z)$ and electric potential $\phi(z)$*

1.3.2 The inversion-layer MOS transistor

Figure 1.10 shows a cross-section of an nMOS transistor with 0 V on all of its terminals. The figure also contains the associated energy band diagram.

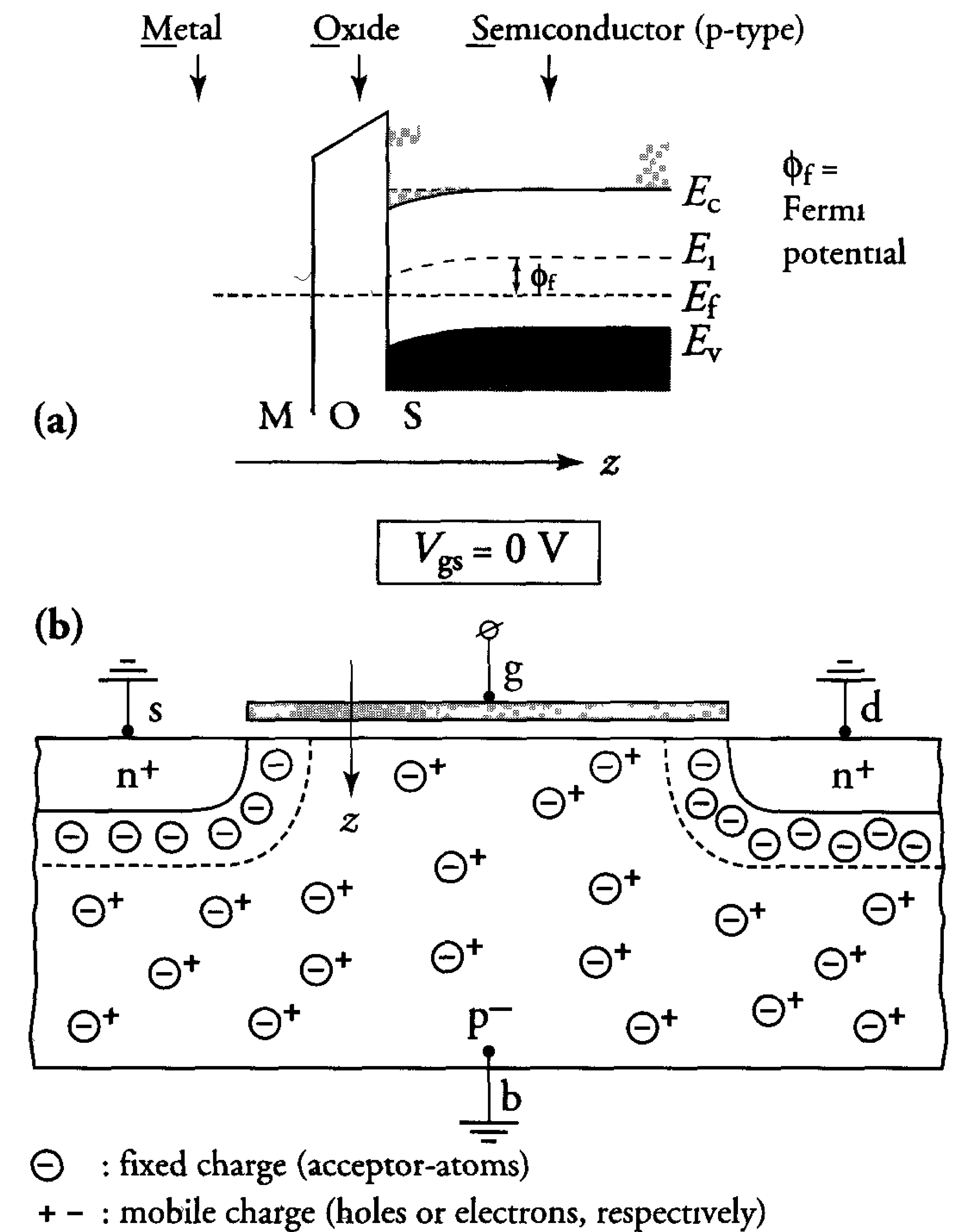


Figure 1.10: *Cross-section of a MOS transistor with $V_{gs}=V_{ds}=V_{sb}=0$ V and the associated energy band diagram*

It is assumed that the presence of the gate does *not* affect the distribution of holes and electrons in the semiconductor. With the exception of the depletion areas around the n^+ areas, the entire p-substrate is assumed to be homogeneous and devoid of an electric field ($E = 0$). There is no charge on the gate and no surface charge in the silicon. Generally,

the electron energies at the Fermi levels of the different materials in the structure will differ. Their work functions (i.e. the energy required to remove an electron from the Fermi level to vacuum) will also differ. When the voltage between the gate and source is zero ($V_{gs}= 0$) and the metal gate is short circuited to the semiconductor, electrons will flow from the metal to the semiconductor or vice versa until a voltage potential is built up between the two materials. This voltage potential counter-balances the difference in their work functions. The Fermi levels in the metal and the semiconductor are then aligned.

Therefore, there will be an electrostatic potential difference between the gate and substrate which will cause the energy bands to bend. The *'flat-band condition'* exists when there is no band-bending at the metal-semiconductor interface. The *'flat-band voltage'* V_{fb} is the gate voltage required to produce the flat-band condition. It is the difference between the work functions of the metal (ϕ_M) and the semiconductor (ϕ_S), i.e. $V_{fb} = \phi_{MS} = \phi_M - \phi_S$. Since equilibrium holds, the Fermi level in the semiconductor remains constant regardless of the value of the gate voltage.

A negative charge is induced in the semiconductor surface when a small positive voltage is applied to the gate, while the source, drain and substrate are at 0V, see also figure 1.11. The negative charge is caused by holes being pushed away from the insulator interface. The negatively charged acceptor atoms that are left behind form a negative space charge, i.e. a depletion layer. The thickness of this depletion layer is determined by the potential V_c at the silicon surface. The gate voltage V_{gs} now consists of two parts:

a. The voltage across the oxide $V_g - V_c$;

b. The voltage across the depletion layer V_c.

The capacitance between the gate and substrate now consists of the series connection of the oxide capacitance C_{ox} and the depletion-layer capacitance C_d.

The term V_T in figure 1.11 represents the *'threshold voltage'* of the transistor. This is the gate voltage at which the band-bending at the silicon surface is exactly $2\phi_f$. For the present, V_T is assumed to be positive for an inversion-layer nMOS transistor. This assumption is confirmed later in the text.

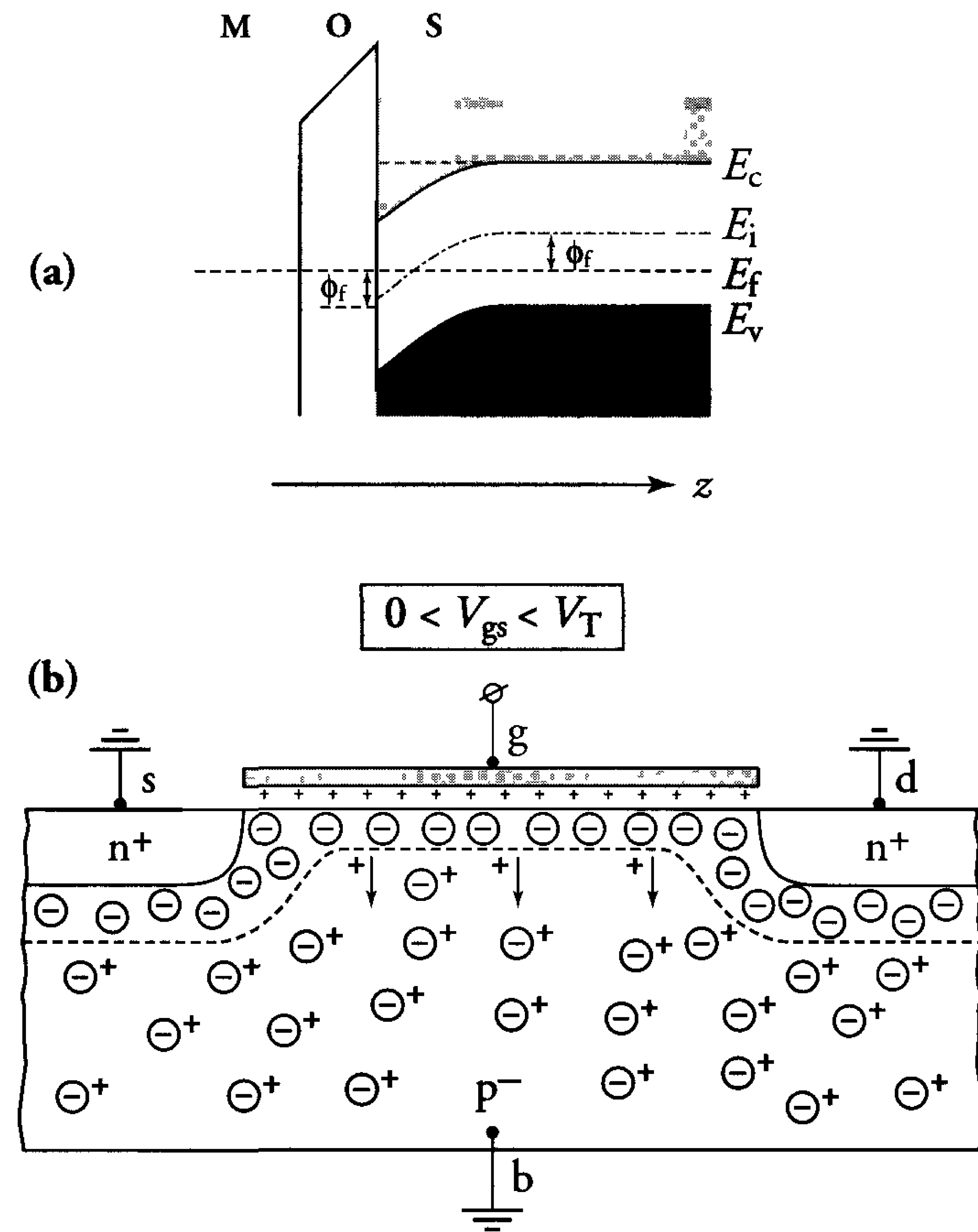


Figure 1.11: *Cross-section of a MOS transistor with $0<V_{gs}<V_T$ and $V_{ds}=V_{sb}=0$ V and its corresponding energy band diagram*

If the gate voltage is further increased ($V_{gs} \geq V_T$), then the band-bending at the silicon surface will be larger than $2\phi_f$. This situation is illustrated in figure 1.12. A comparison of figure 1.12 and figure 1.7c reveals that the energy band at the silicon surface corresponds to an n-type semiconductor.

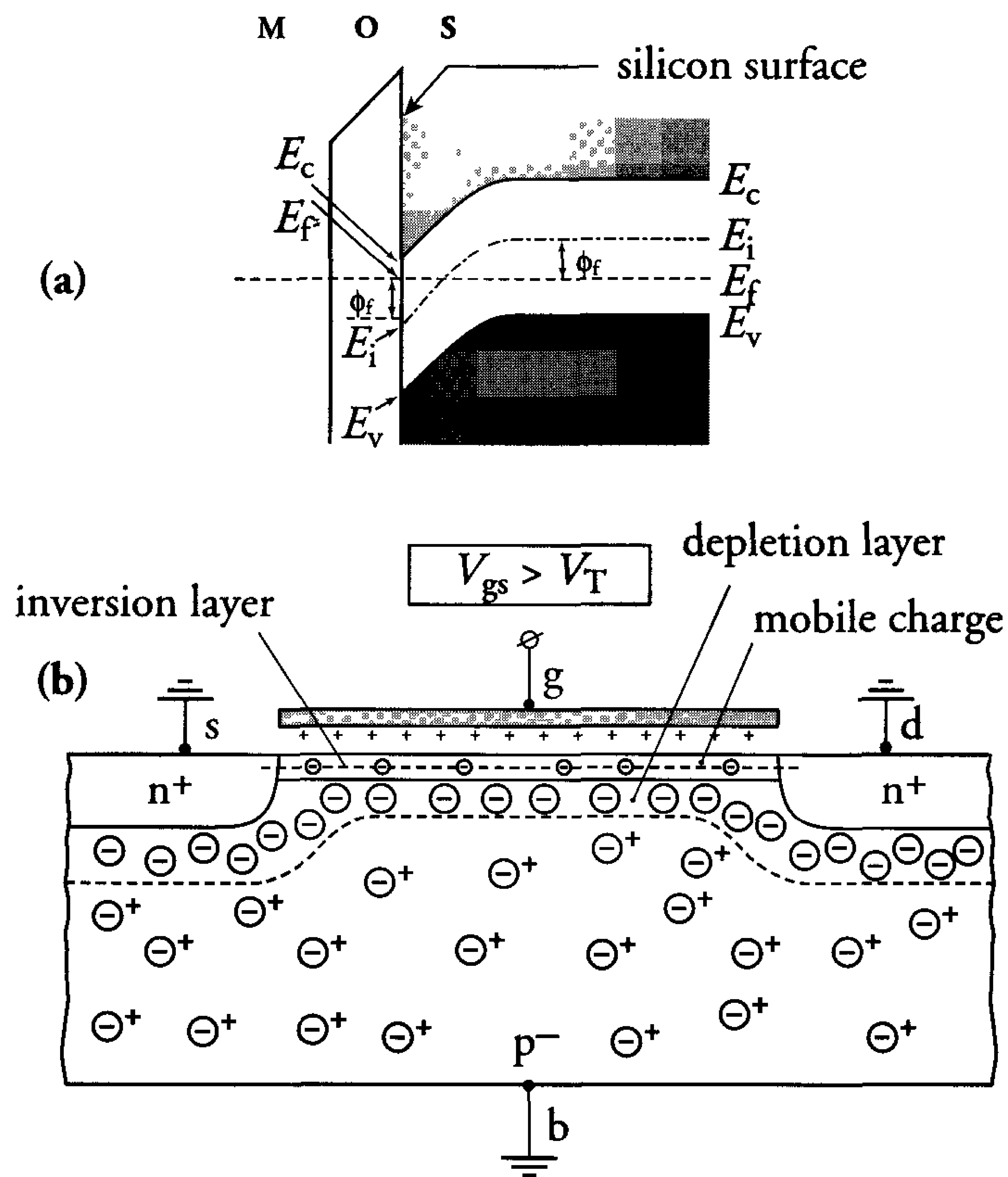


Figure 1.12: *Cross-section of a MOS transistor with* $V_{gs}>V_T$ *(*$V_T>0$*) and* $V_{ds}=V_{sb}=0$ *V and its corresponding energy band diagram*

Deep in the substrate, however, the energy band corresponds to a p-type semiconductor. A very narrow n-type layer has therefore been created at the surface of a p-type silicon substrate. In addition to the negative acceptor atoms already present, this *inversion layer* contains electrons which act as mobile negative charge carriers. Conduction in the n-type inversion layer is mainly performed by these electrons, which are minority carriers in the p-type substrate. The inversion layer forms a conducting *channel* between the transistor's source and drain. No current flows in this channel if there is no voltage difference between the drain and source terminals, i.e. $I_{ds} = 0$ A if $V_{ds} = 0$ V. The number of electrons in the channel can be controlled by the gate-source voltage V_{gs}.

Assuming that $V_{gs} > V_T$, the effects of increasing V_{ds} from 0 V are divided into the following regions:

1. $0 < V_{ds} < V_{gs} - V_T$.
   This is called the *linear* or *triode* region of the MOS transistor's operating characteristic.

2. $V_{ds} = V_{gs} - V_T$.
   At this point, a transition takes place from the linear to the so-called saturation region.

3. $V_{ds} > V_{gs} - V_T$.
   This is the *saturation* region of the MOS transistor's operating characteristic.

The three regions are discussed separately on the following pages.

The linear region $\begin{cases} V_{gs} > V_T > 0 \\ 0 < V_{ds} < V_{gs} - V_T \end{cases}$

Figure 1.13 shows the situation in the linear region, in which a current I_{ds} (which flows from drain to source) causes a voltage difference in the channel. The surface potential under the gate decreases from V_{ds} in the drain to 0 V in the source. The maximum potential difference between the gate and channel is at the source. Therefore, the strongest inversion and the highest concentration of electrons in the inversion layer occur adjacent to the source. The maximum potential difference between the channel and substrate is at the drain. The depletion layer is therefore thickest here. In the linear region, the drain current I_{ds} increases with increasing V_{ds} for a constant V_{gs}.

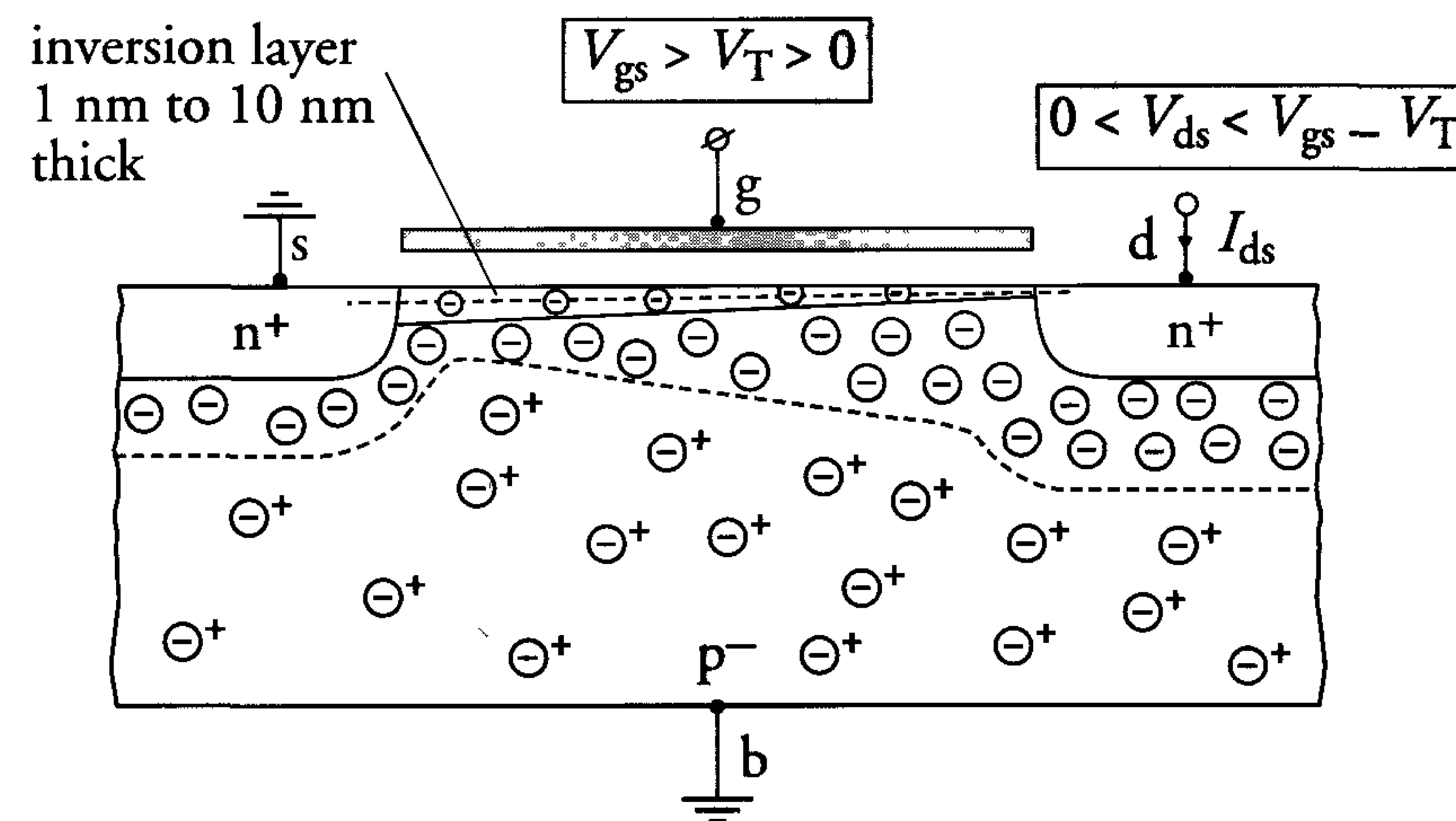


Figure 1.13: *Cross-section of a transistor operating in the linear (triode) region*

The transition region $\begin{cases} V_{gs} > V_T > 0 \\ V_{ds} = V_{gs} - V_T \end{cases}$

An increase in V_{ds}, with V_{gs} constant decreases the voltage difference between the gate and channel at the drain. The inversion layer disappears at the drain when the voltage difference between the gate and channel equals the threshold voltage V_T. The channel end then coincides with the drain-substrate junction. This situation occurs when $V_{ds} = V_{gs} - V_T$, and is shown in figure 1.14.

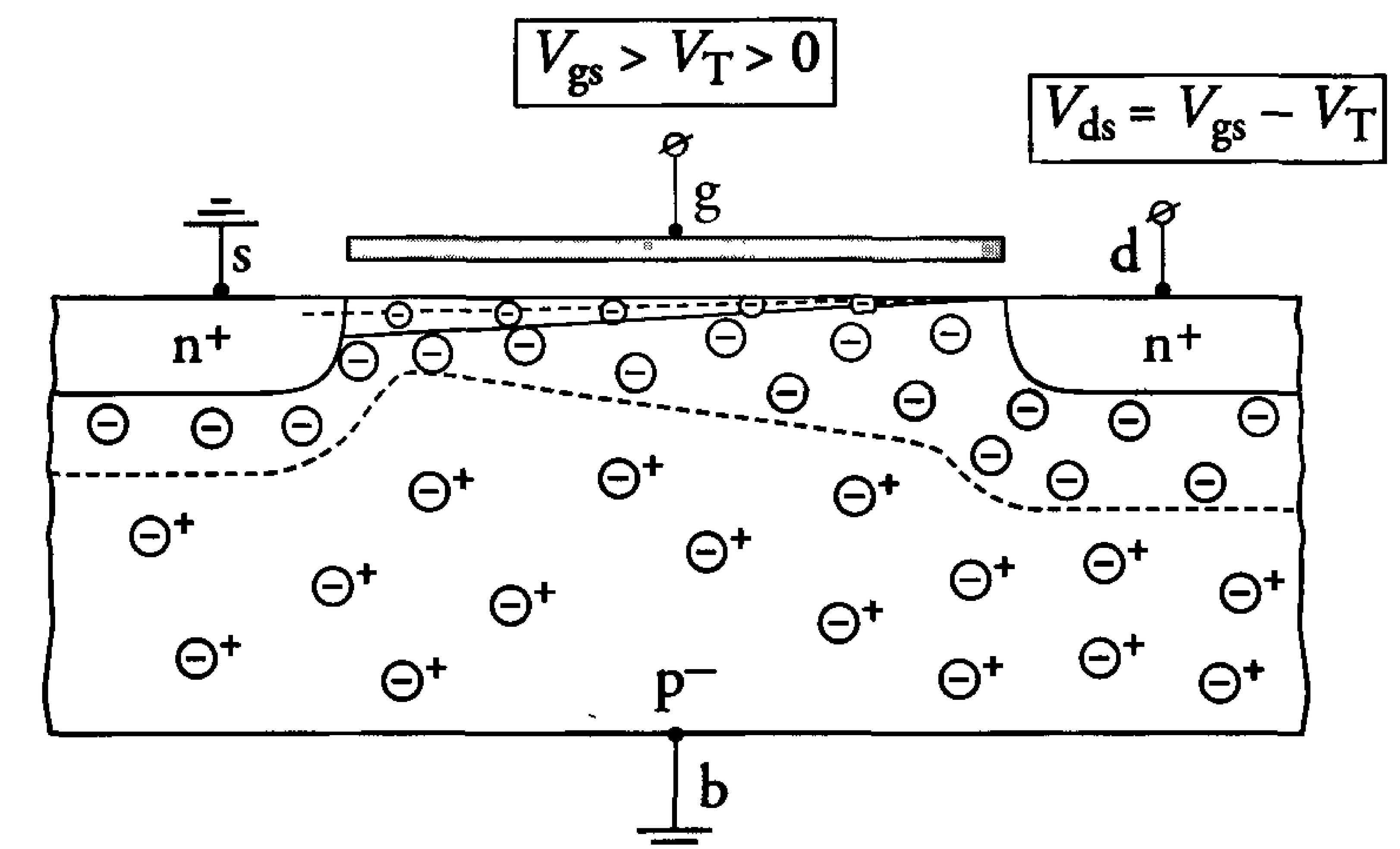


Figure 1.14: *Situation during transition from triode to saturation region, i.e.* $V_{ds}=V_{gs}-V_T$

The saturation region $\begin{cases} V_{gs} > V_T > 0 \\ V_{ds} > V_{gs} - V_T \end{cases}$

The channel end no longer coincides with the drain when V_{ds} is larger than $V_{gs} - V_T$. This situation is shown in figure 1.15.

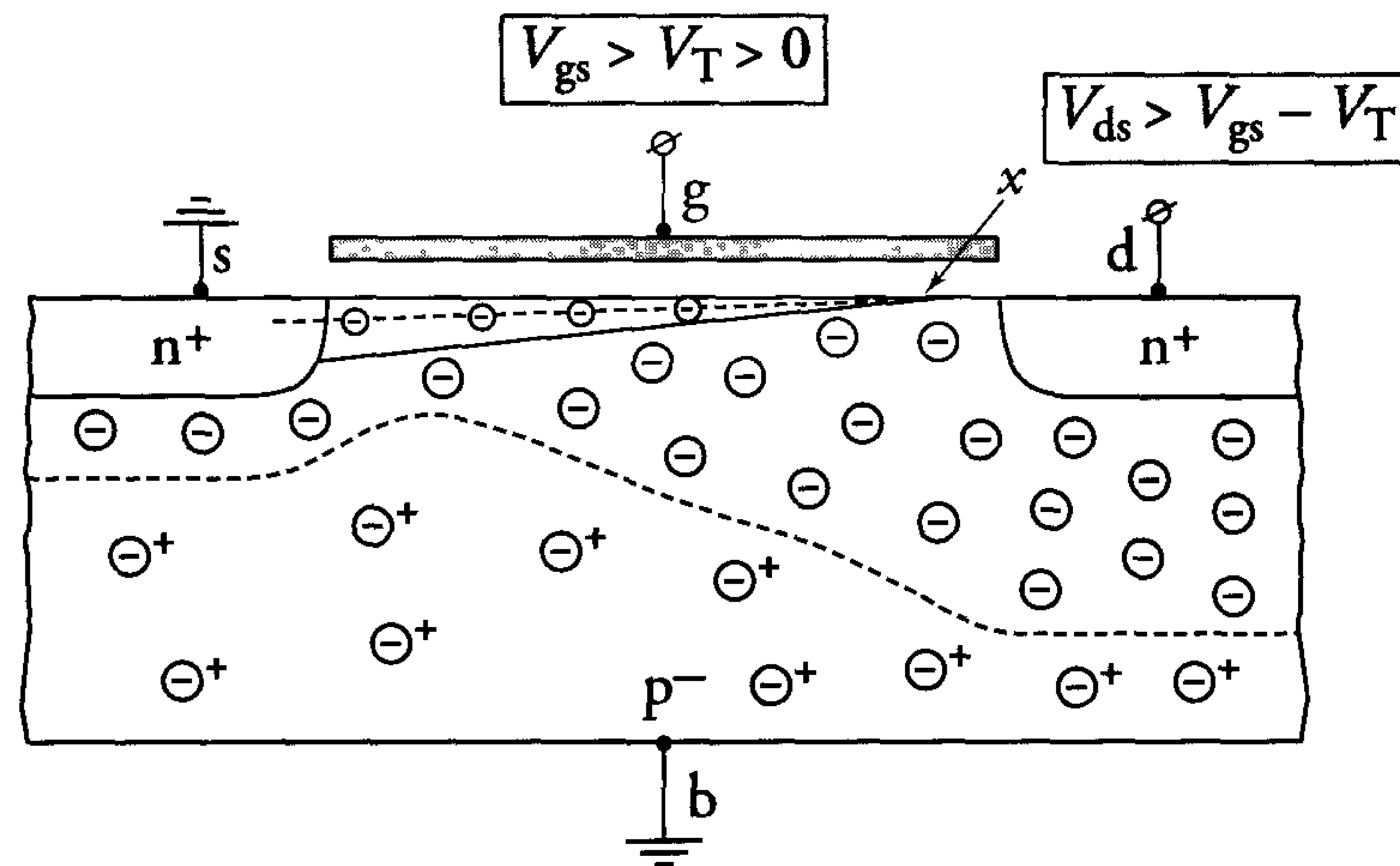


Figure 1.15: *Situation in the saturation region, i.e.* $V_{ds} > V_{gs} - V_T$

The voltage V_x at the end point x of the inversion layer equals $V_{gs} - V_T$. Therefore, V_T is the voltage difference between the gate and channel at position x. If this *pinch-off point* is considered to be the imaginary drain of the transistor, then I_{ds} is determined by the voltage $V_x = V_{gs} - V_T$. In other words, the drain current in the saturation region equals the drain current at the transition point between the linear and saturation regions. The value of the *saturation current* is clearly proportional to V_{gs}. Electrons are emitted from the inversion layer into the depletion layer at the pinch-off point. These electrons will be attracted and collected by the drain because $V_{ds} > V_x$.

Figure 1.16 shows the $I_{ds} = f(V_{ds})$ characteristic for various gate voltages. If $V_{ds} = 0\,\text{V}$, then $I_{ds} = 0\,\text{A}$. If V_{ds} is less than $V_{gs} - V_T$, then the transistor operates in the triode region and the current I_{ds} displays an almost linear relationship with V_{ds}. Current I_{ds} increases to its saturation value when $V_{ds} = V_{gs} - V_T$. Further increases of V_{ds} above $V_{gs} - V_T$ no longer cause increases in I_{ds}. The transition between the triode and saturation regions is characterised by the curve $V_{ds} = V_{gs} - V_T$.

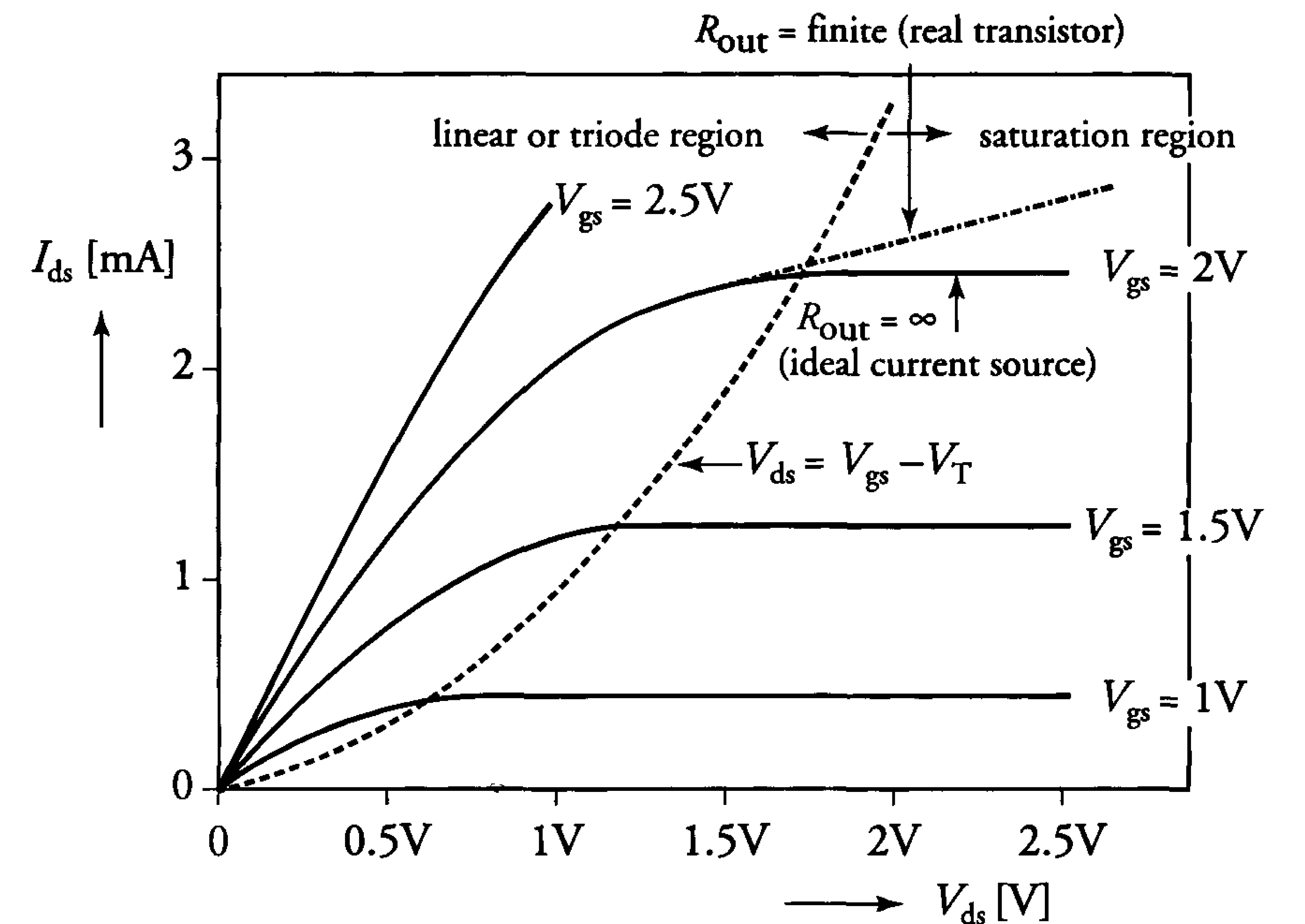


Figure 1.16: *The* $I_{ds} = f(V_{ds})$ *characteristic for various values of* V_{gs}

1.4 Derivation of simple MOS formulae

The inversion layer nMOS transistor shown in figure 1.17 has a width W perpendicular to the plane of the page and an oxide capacitance C_{ox} per unit area. A commonly-used unit for C_{ox} is $\text{fF}/\mu\text{m}^2$, where $1\text{fF} = 10^{-15}\text{F}$.

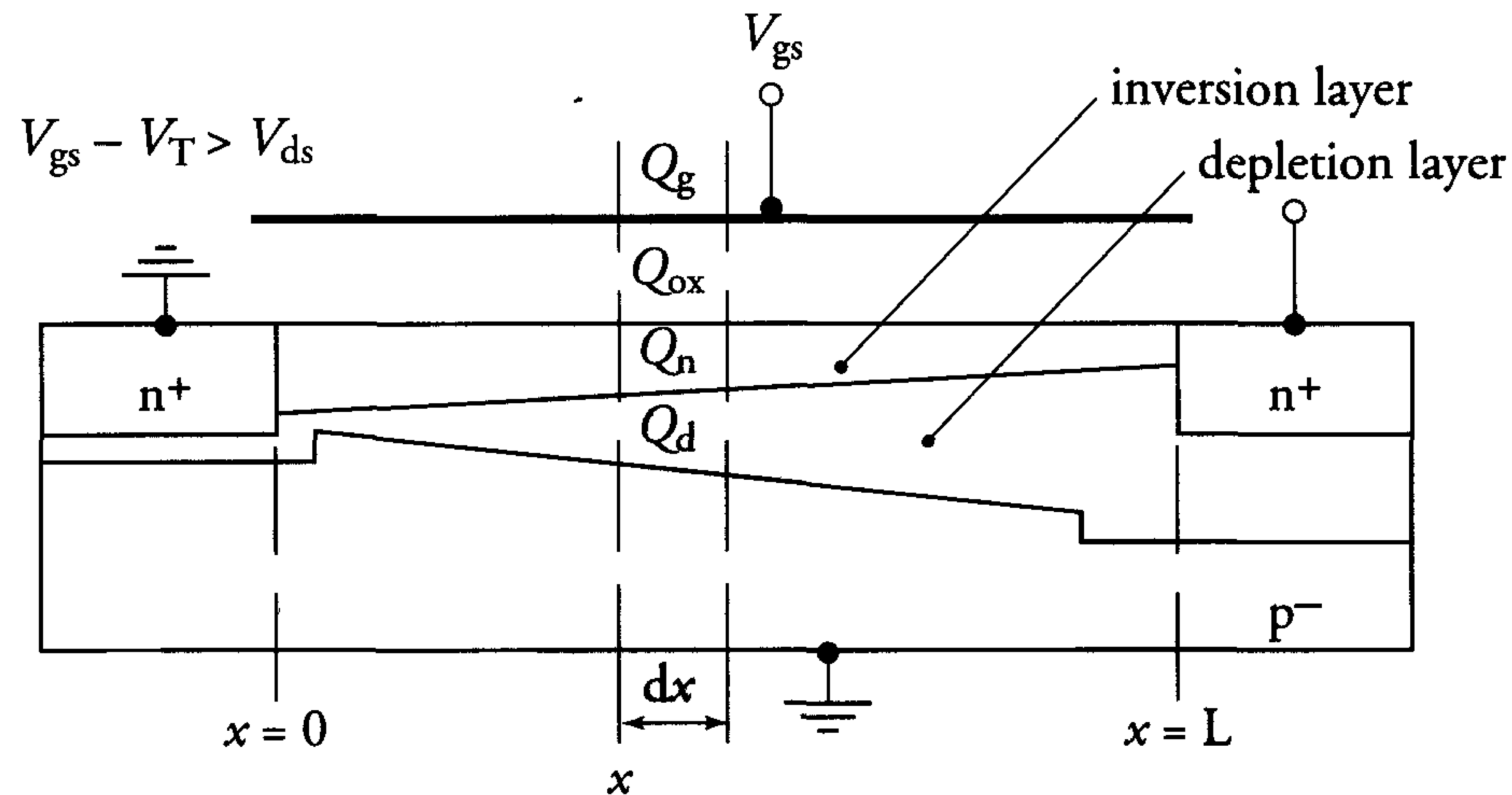


Figure 1.17: *Charges in a MOS transistor operating in the linear region*

Based on the *law for conservation of charge*, the following equality must hold at any position x between the source and drain:

$$Q_g + Q_{ox} + Q_n + Q_d = 0. \tag{1.3}$$

The components in this equation are charges per unit area, specified as follows:

Q_g = the gate charge $[C/m^2]$;
Q_{ox}= primarily a small fixed charge which in practice always appears to be present in the thin gate oxide $[C/m^2]$;
Q_n = the mobile charge in the inversion layer $[C/m^2]$;
Q_d = the fixed charge in the depletion layer $[C/m^2]$.

For gate voltages larger than V_T, the inversion layer shields the depletion layer from the gate. The charge in the depletion layer can then be considered constant:

$$Q_{ox} + Q_d = -C_{ox} \cdot V_T \tag{1.4}$$

The threshold voltage V_T is assumed to be constant. The potential in the channel at a position x is $V(x)$. With $Q_g = C_{ox}[V_{gs} - V(x)]$, substituting (1.3) into (1.4) yields:

$$Q_n = -C_{ox}[V_{gs} - V_T - V(x)]$$

The total mobile charge dQ_m in a section of the channel with length dx is defined as:

$$dQ_m = Q_n \cdot W \cdot dx = -W \cdot C_{ox}[V_{gs} - V_T - V(x)] \cdot dx \tag{1.5}$$

$$\Rightarrow \frac{dQ_m}{dx} = -W \cdot C_{ox}[V_{gs} - V_T - V(x)] \tag{1.6}$$

The drain current I_{ds} is expressed as:

$$I_{ds} = \frac{dQ_m}{dt} = \frac{dQ_m}{dx} \cdot \frac{dx}{dt} \tag{1.7}$$

where $\frac{dQ_m}{dx}$ is defined in equation (1.6) and
$\frac{dx}{dt}$ is the velocity v at which the charge Q_m moves from the source to the drain region.

This is the velocity of the electrons in the inversion layer and is expressed as:

$$v = \mu_n \cdot E = -\mu_n \cdot \frac{dV(x)}{dx} \tag{1.8}$$

where E is the electric field strength and
μ_n represents the electron mobility in the inversion layer.

In practice, this appears to be equal to half of the electron mobility in the substrate (see section 2.3). Combining equations (1.6), (1.7) and (1.8) yields:

$$I_{ds} = \mu_n \cdot C_{ox} \cdot W \cdot [V_{gs} - V_T - V(x)] \cdot \frac{dV(x)}{dx} \tag{1.9}$$

Substituting $\beta_\square = \mu_n \cdot C_{ox}$ yields:

$$I_{ds} \cdot dx = \beta_\square \cdot W \cdot [V_{gs} - V_T - V(x)] \cdot dV(x) \tag{1.10}$$

Integrating the left-hand side from 0 to L and the right-hand side from 0 to V_{ds} yields:

$$I_{ds} = \frac{W}{L} \cdot \beta_\square \cdot (V_{gs} - V_T - \frac{1}{2}V_{ds}) \cdot V_{ds} \tag{1.11}$$

Equation (1.11) has a maximum value when $V_{ds} = V_{gs} - V_T$. In this case, the current I_{ds} is expressed as:

$$I_{ds} = \frac{1}{2} \cdot \frac{W}{L} \cdot \beta_{\square} \cdot (V_{gs} - V_T)^2 \tag{1.12}$$

If $V_{gs} = V_T$ then $I_{ds} = 0\,\text{A}$. This clearly agrees with the earlier assumption that V_T is positive for an inversion-layer nMOS transistor. The term β is usually used to represent $\frac{W}{L} \cdot \beta_{\square}$. This factor is called the transistor *gain factor* and depends on geometry. The gain term $\beta_{\square}$ is a process parameter which depends on such things as the oxide thickness t_{ox}:

$$\beta_{\square} = \mu_n \cdot C_{ox} = \mu_n \cdot \frac{\epsilon_0 \epsilon_{ox}}{t_{ox}}$$

The unit of measurement for both β and $\beta_{\square}$ is A/V^2. However, $\mu\text{A/V}^2$ and mA/V^2 are the most commonly-used units. For an n-channel MOS transistor, $\beta_{\square}$ varies from $120\,\mu\text{A/V}^2$ to $240\,\mu\text{A/V}^2$ for oxide thicknesses of 10 nm and 5 nm, respectively.

According to equation (1.11), I_{ds} would reach a maximum value and then decrease for increasing V_{ds}. In the discussion concerning figures 1.15 and 1.16, however, it was stated that the current remains constant for an increasing V_{ds} once $V_{ds} > V_{gs} - V_T$. The transistor has two operating regions which are characterised by corresponding expressions for I_{ds}. These regions and their I_{ds} expressions are defined as follows:

1. The linear or triode region. $0 < V_{ds} < V_{gs} - V_T$.

$$I_{ds} = \beta \cdot (V_{gs} - V_T - \frac{1}{2} V_{ds}) \cdot V_{ds} \tag{1.13}$$

2. The saturation region. $V_{ds} \geq V_{gs} - V_T$.

$$I_{ds} = \frac{\beta}{2} \cdot (V_{gs} - V_T)^2 \tag{1.14}$$

According to equation (1.14), I_{ds} is independent of V_{ds} in the saturation region. The output impedance $\text{d}V_{ds}/\text{d}I_{ds}$ should then be infinite and the transistor should behave like an ideal current source. In practice, however, MOS transistors show a finite output impedance which is dependent on geometry . This is explained in chapter 2. Figure 1.16 shows both the ideal (theoretical) and the real current-voltage characteristics of a transistor for $\beta = 2.5\,\text{mA/V}^2$ and $V_T = 0.5\,\text{V}$.

The $I_{ds} = f(V_{ds})|_{V_{gs}=\text{constant}}$ curves in figure 1.16 are joined by the dotted curve $V_{ds} = V_{gs} - V_T$ at the points where equation (1.13) yields maximum values for I_{ds}. This curve divides the $I_{ds} - V_{ds}$ plane into two regions:

1. Left of the dotted curve: the triode or linear region, which is defined by equation (1.13);

2. Right of the dotted curve: the saturation region, which is defined by equation (1.14).

1.5 The back-bias effect (back-gate effect, body effect)

The simple MOS formulae derived in section 1.4 appear to be reasonably satisfactory in most cases. The very important *back-bias effect* is, however, not included in these formulae. This effect accounts for the modulation of the threshold voltage by the substrate bias and the subsequent effects on the drain current.

During normal operation (when $V_{gs} > V_T$ and $V_{ds} > V_{gs} - V_T$) a depletion layer is formed, as shown in figure 1.15. However, the thickness of the depletion region under the channel increases when a negative *back-bias voltage* (V_{sb}) is applied to the bulk (b) with respect to the source. This is caused by the increased reverse-bias voltage across the fictive channel-substrate junction. The increased depletion layer requires additional charge. The channel charge therefore decreases if V_{gs} is held constant. The channel conductivity can only be maintained if V_{gs} is increased. The threshold voltage is therefore related to the back-bias voltage V_{sb}. This dependence is expressed as follows:

$$V_T = V_x + K\sqrt{V_{sb} + 2\phi_f} \tag{1.15}$$

$$V_{T0} = V_x + K\sqrt{2\phi_f} \tag{1.16}$$

The terms in these formulae are as follows:

V_x = process-related constant threshold voltage term;
V_{T0} = $V_T|_{V_{sb}=0V}$;
K = process parameter equal to $\frac{1}{C_{ox}}\sqrt{2N_A q \epsilon_0 \epsilon_{si}}$, also known as the *'body factor'* or K-factor;
N_A = substrate (bulk) dope concentration;

V_{sb} = source-bulk (back-bias) voltage;
$2\,\phi_f$ = band bending where inversion first occurs.

The back-bias effect causes MOS transistors of the same type and dimensions to have different threshold voltages. The inverter shown in figure 1.18 serves as an example. Applying equation (1.15) yields the following equations for transistors T_1 and T_2, respectively:

$$V_{T1} = V_x + K\sqrt{V_{s_1b} + 2\phi_f}$$
$$V_{T2} = V_x + K\sqrt{V_{s_2b} + 2\phi_f}$$

If the output is 'high' ($\approx 1.8\,V$), the source-bulk voltages of T_1 and T_2 are $V_{s_1b} = V_{ss} - V_{bb} = 2\,V$ and $V_{s_2b} = V_{out} - V_{bb} = 3.8\,V$, respectively. The K-factor can therefore cause the threshold voltage V_{T2} of the upper transistor to be considerably larger than the threshold voltage V_{T1} of the lower transistor.

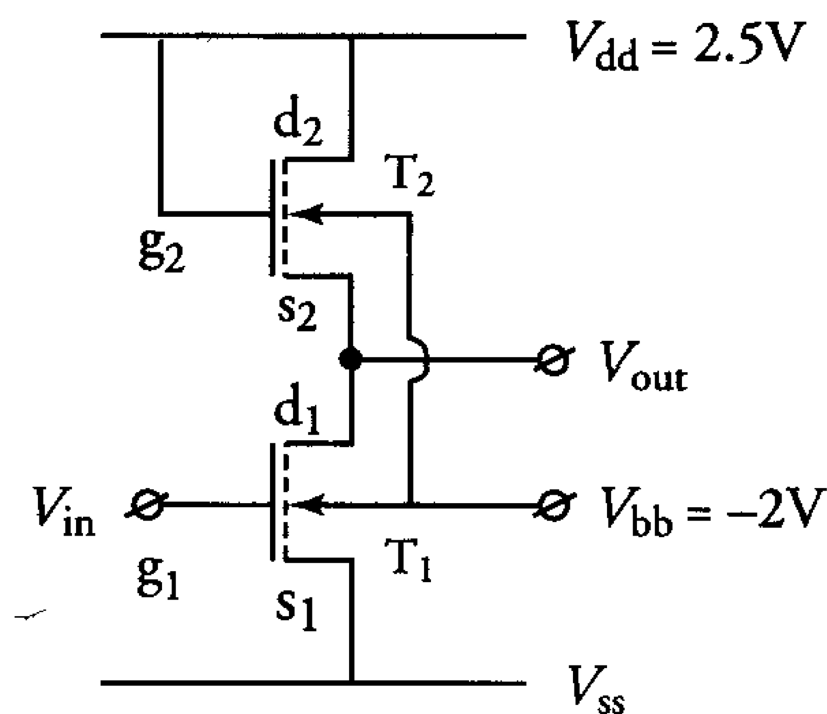


Figure 1.18: *nMOS-inverter with enhancement load*

Figure 1.19 shows the influence of the back-bias effect on different transistor characteristics. Formula (1.15) clearly shows that the threshold voltage V_T increases with an increasing back-gate voltage V_{sb}. For a constant V_{gs}, the drain-source current therefore decreases for an increasing V_{sb}. This is illustrated in figure 1.19b.

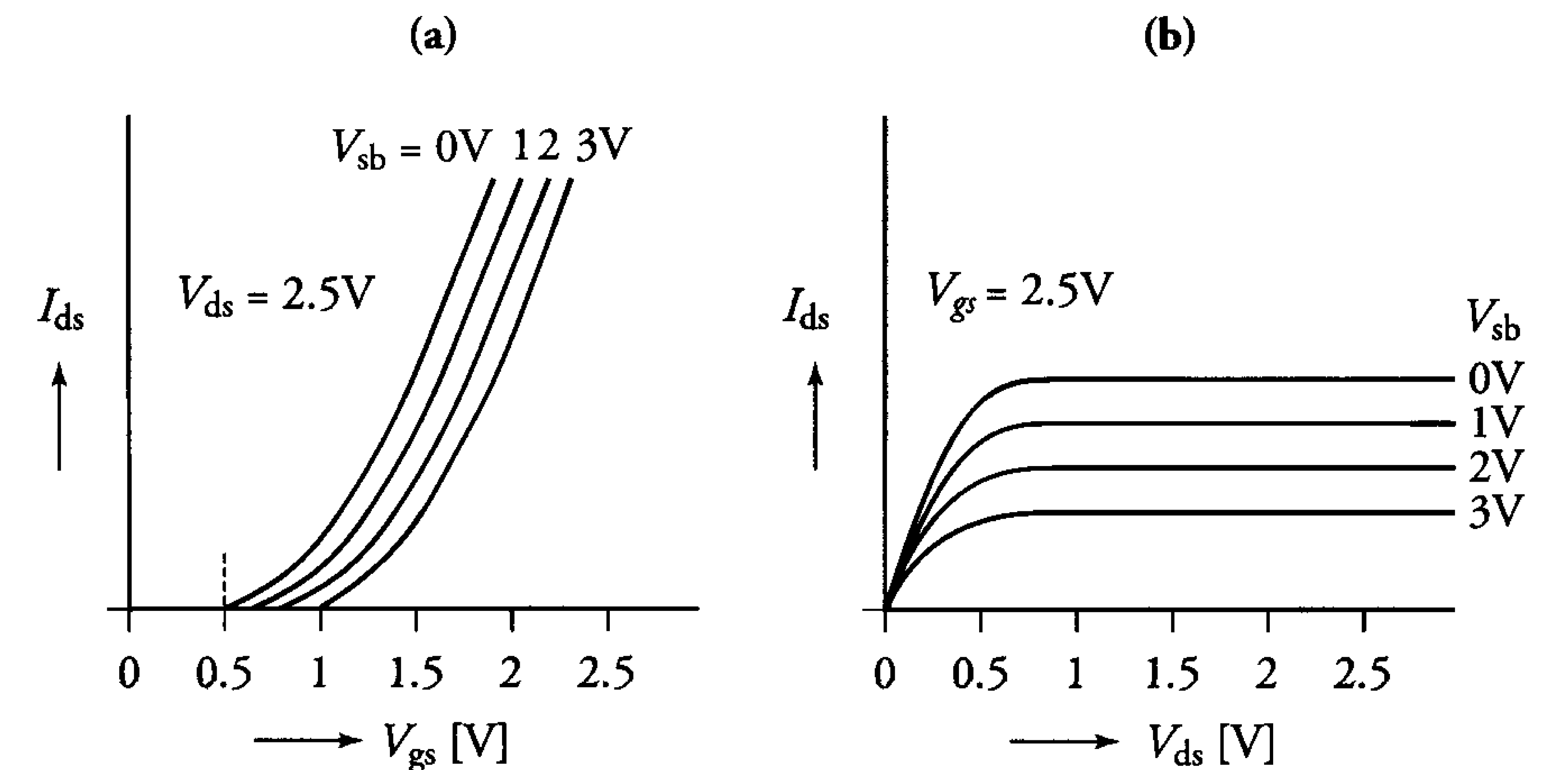


Figure 1.19: *Back-bias effect on MOS transistor characteristics:* (a) $I_{ds} = f(V_{gs})|_{V_{ds}=const}$ (b) $I_{ds} = f(V_{ds})|_{V_{gs}=const}$

Figure 1.20 shows the dependence of V_T on V_{sb}. The starting-point of this graph is determined by V_{T0} in equation (1.16) while its curve depends on the K-factor.

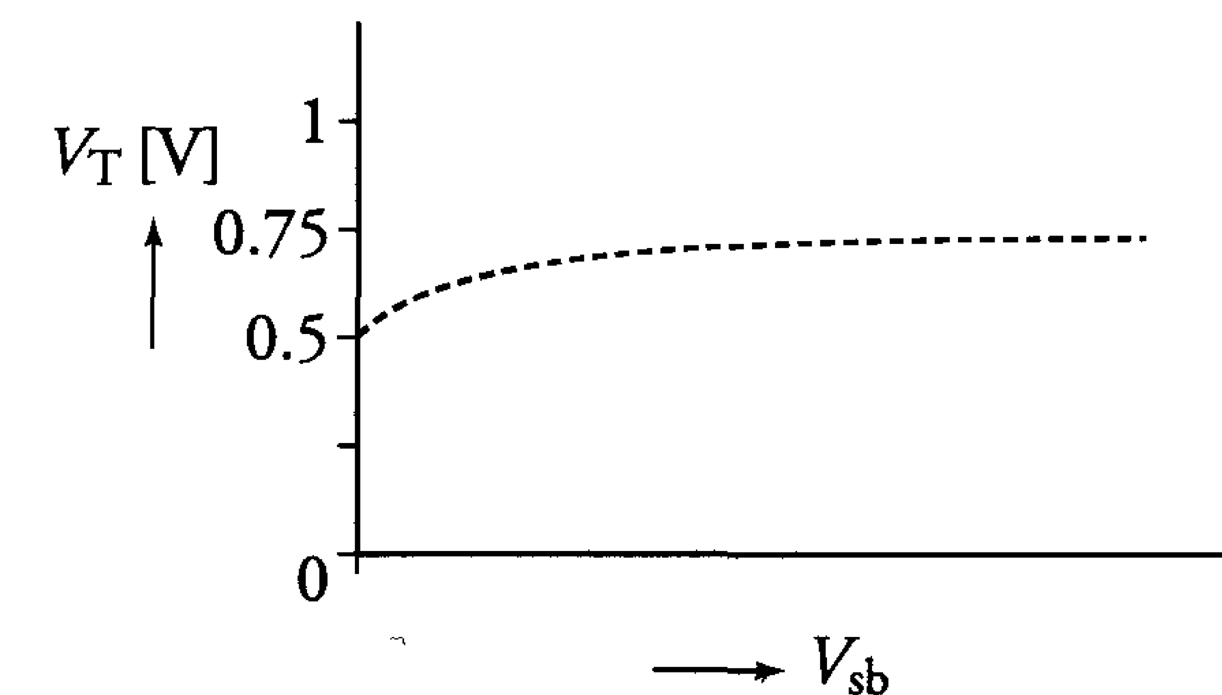


Figure 1.20: $V_T = f(V_{sb})$: *Threshold voltage as a function of source-bulk voltage*

The back-bias effect must be accurately treated when dimensioning MOS circuits. The most important reasons for using a back-bias voltage are as follows:

- Normally, the term V_x in equations (1.15) and (1.16) spreads more than the K-factor. The influence of the K-factor on the threshold voltage is larger when a back-bias voltage is applied. This results in a more stable threshold voltage.
- The depletion layer around the source and drain junctions of the MOS transistor becomes thicker as a result of the increased reverse voltage across these p-n junctions. This reduces the parasitic capacitances of the source and drain.
- Negative voltage pulses which may occur in dynamic MOS logic circuits may forward-bias the p-n diode between the substrate and a source or drain. Application of a negative voltage to the substrate virtually removes this possibility.

The MOS transistor formulae are summarised as follows:

$$\begin{array}{rcl} \text{– linear region} & : & I_{ds} = \beta(V_{gs} - V_T - \frac{V_{ds}}{2})V_{ds} \\ \text{– saturation region} & : & I_{ds} = I_{ds_{sat}} = \frac{\beta}{2}(V_{gs} - V_T)^2 \\ \text{where} & & V_T = V_x + K\sqrt{V_{sb} + 2\phi_f} \\ \text{and} & & V_{T0} = V_x + K\sqrt{2\phi_f} \end{array} \quad (1.17)$$

1.6 Factors which characterise the behaviour of the MOS transistor

The previously-discussed current-voltage characteristics represent the relationship between a transistor's current (I_{ds}) and its various applied voltages (V_{gs}, V_{ds} and V_{sb}). A number of important parameters which are frequently used to describe the behaviour of a transistor are explained below.

The *transconductance* g_m describes the relationship between the change δI_{ds} in the transistor current caused by a change δV_{gs} in the gate voltage:

$$g_m = \frac{\delta I_{ds}}{\delta V_{gs}}|V_{ds} = \text{const} \quad (1.18)$$

Referring to figure 1.16, it is clear that the value of g_m depends on the transistor's operating region:

$$\text{Linear region}: g_m = \beta \cdot V_{ds} \quad (1.19)$$

$$\text{Saturation region}: g_{m_{sat.}} = \beta \cdot (V_{gs} - V_T) \quad (1.20)$$

Another parameter that characterises conduction in a transistor is its output resistance. In the transistor's linear operating region, this resistance (which is also called the channel resistance) is defined as:

$$R_{out} = (\frac{dI_{ds}}{dV_{ds}})^{-1} = \{\beta(V_{gs} - V_T) - \beta V_{ds}\}^{-1} \quad (1.21)$$

If V_{ds} is small, then:

$$R_{out} = \frac{1}{\beta(V_{gs} - V_T)} \quad (1.22)$$

For an ideal MOS transistor operating in the saturation region, we have $\frac{dI_{ds}}{dV_{ds}} = 0$. The transistor current is then independent of V_{ds}. The output resistance is therefore infinite and the transistor acts as an ideal current source. In practice, however, the MOS transistor always has a finite output resistance and its current remains dependent on V_{ds}. This is illustrated in figure 1.16 and is treated in section 2.4.

1.7 Different types of MOS transistors

1. The previous discussions are all related to *n-channel* MOS transistors. The substrate material of these nMOS transistors is p-type and the drain and gate voltages are *positive* with respect to the source during normal operation. The substrate is the most *negative* electrode of an nMOS transistor.
2. *P-channel* MOS transistors are produced on an n-type substrate. The voltages at the gate and drain of these pMOS transistors are *negative* with respect to the source during normal operation. The substrate is the most *positive* electrode.

Generally, nMOS circuits are faster than those with pMOS transistors. The *power-delay* (τD) *product* of a logic gate is the product of its delay τ and dissipation D. The τD products of nMOS logic gates are lower than those of pMOS logic gates. This is because of the difference between the *mobility* of electrons and holes. Electron mobility is a factor 2.5 to 3.5 times higher than hole mobility in both the bulk silicon and inversion layers of the respective devices. Figure 2.1 illustrates this relationship, which is expressed as follows:

$$\mu_{\mathrm{n}} \approx 3 \cdot \mu_{\mathrm{p}}$$

The following relationship then follows from equation 1.13:

$$\beta_{\square_{\mathrm{n}}} \approx 3 \cdot \beta_{\square_{\mathrm{p}}}$$

An nMOS transistor therefore conducts approximately three times as much current as a pMOS transistor of equal dimensions and with equal absolute voltages.

Figure 1.21 shows a schematic overview of transistors which are distinguished on the basis of threshold voltage V_{T}. This distinction applies to both pMOS and nMOS transistors and results in the following types:

- *Enhancement* or *normally-off* transistors:
  No current flows through an enhancement transistor when $V_{\mathrm{gs}} = 0$. $V_{\mathrm{T}} > 0$ for an nMOS enhancement transistor and $V_{\mathrm{T}} < 0$ for a pMOS enhancement transistor.

- *Depletion* or *normally-on* transistors:
  Current flows through a depletion transistor when $V_{\mathrm{gs}} = 0$. $V_{\mathrm{T}} < 0$ for an nMOS depletion transistor and $V_{\mathrm{T}} > 0$ for a pMOS depletion transistor.

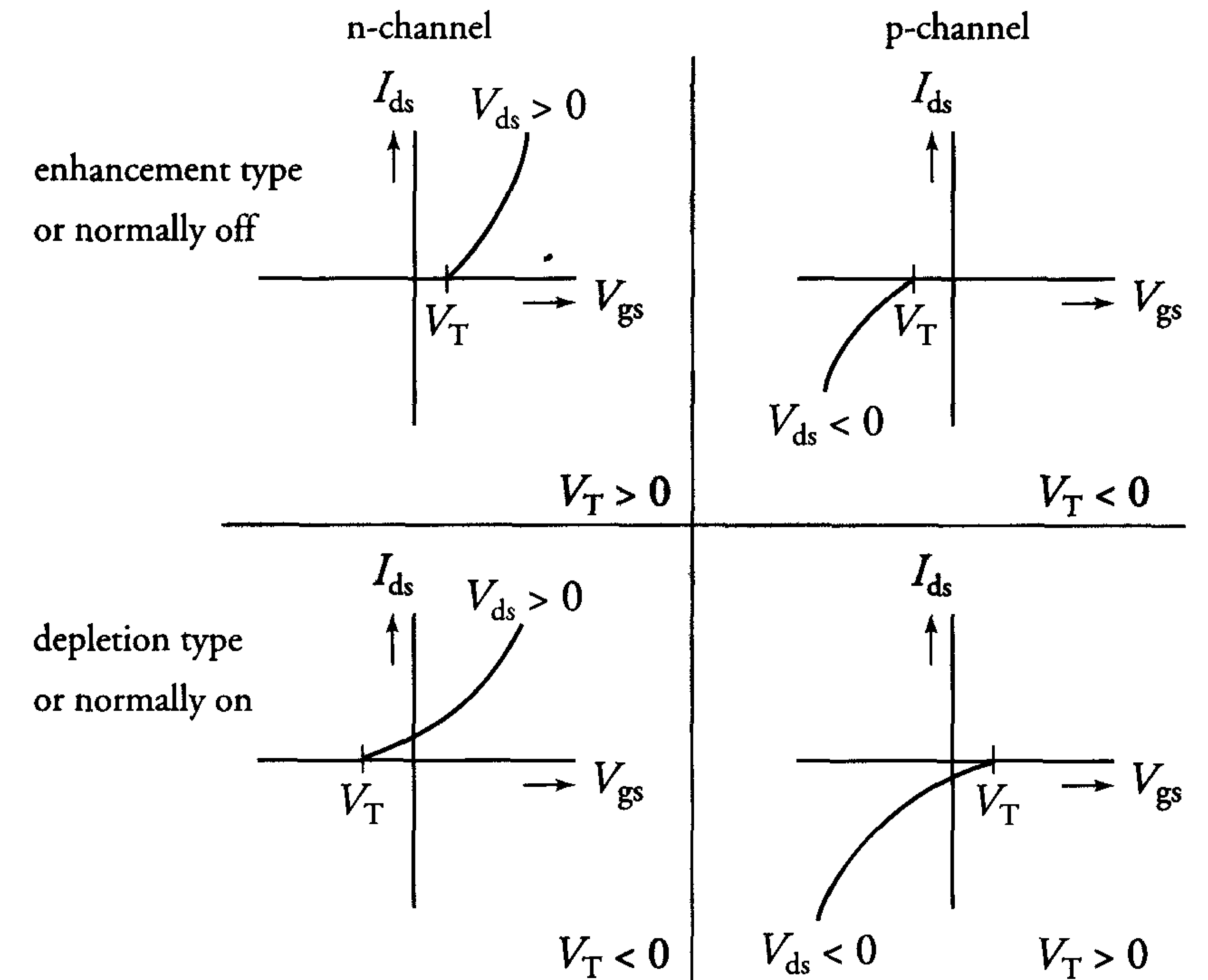


Figure 1.21: *Schematic overview of the different types of MOS transistors*

1.8 Parasitic MOS transistors

MOS (V)LSI circuits comprise many closely-packed transistors. This leads to the presence of parasitic MOS transistors, as illustrated in figure 1.22.

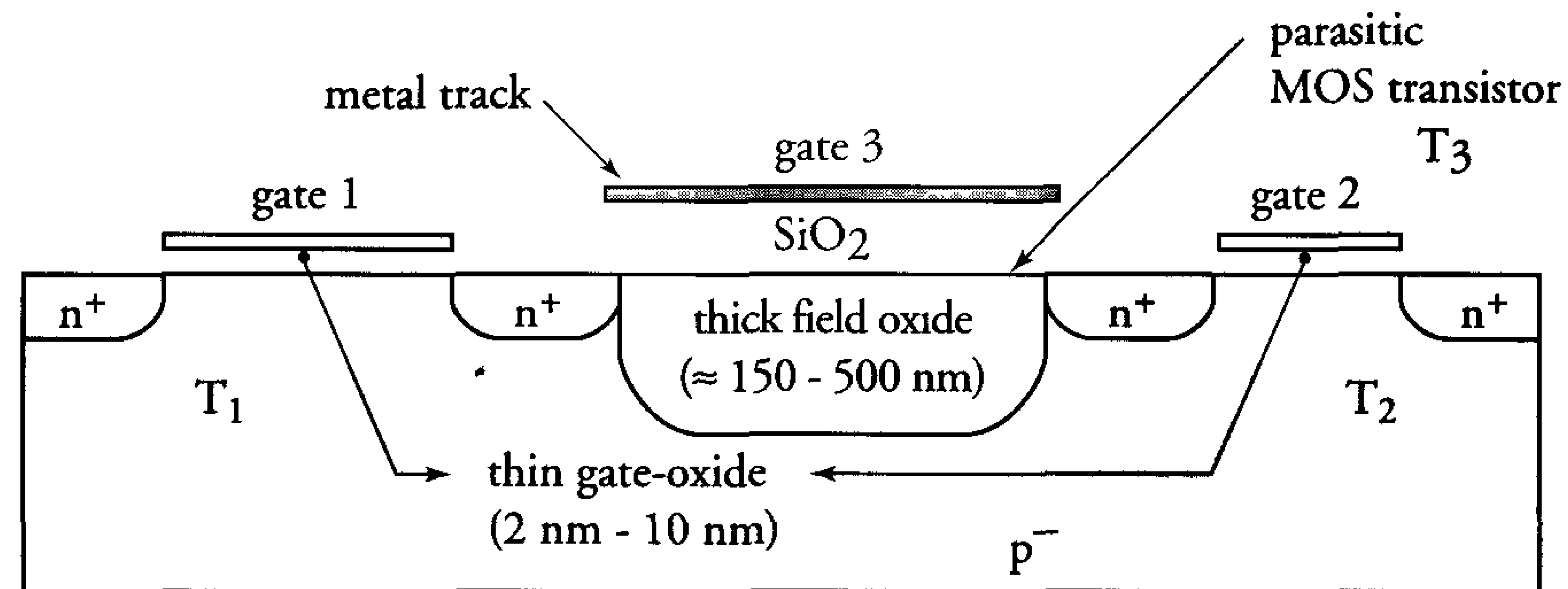


Figure 1.22: *Example of a parasitic MOS transistor*

Transistors T_1 and T_2 are separated by the *field oxide.* Parasitic MOS transistor T_3 is formed by an interconnection on the field oxide and the n^+ areas of transistors T_1 and T_2. This field oxide is thick in comparison with the gate oxide, which ensures that the threshold voltage $V_{T\text{par}}$ of transistor T_3 is larger than the threshold voltages of transistors T_1 and T_2. The field strength at the silicon surface in T_3 is therefore lower than in T_1 and T_2. Transistor T_3 will never conduct if its gate voltage never exceeds $V_{T\text{par}}$.

Many MOS production processes use an extra diffusion or ion implantation to artificially increase the threshold voltage $V_{T\text{par}}$ of parasitic transistors. For this purpose, boron is used to create a p-type layer beneath the thick oxide in processes that use p^--type substrates. This makes it much more difficult to create an n-type inversion layer in these areas.

Processes that use n^--type substrates use phosphorus to increase $|V_{T\text{par}}|$. The terms *channel stopper implants* and *guard ring* are used to refer to these boron and phosphorous implantations.

Note: Parasitic MOS transistors also appear in bipolar circuits. The absolute value of parasitic threshold voltages is always higher in n-type substrates than in p-type substrates. This is one of the reasons why planar IC technologies were mainly developed on n-epi layers.

1.9 MOS transistor symbols

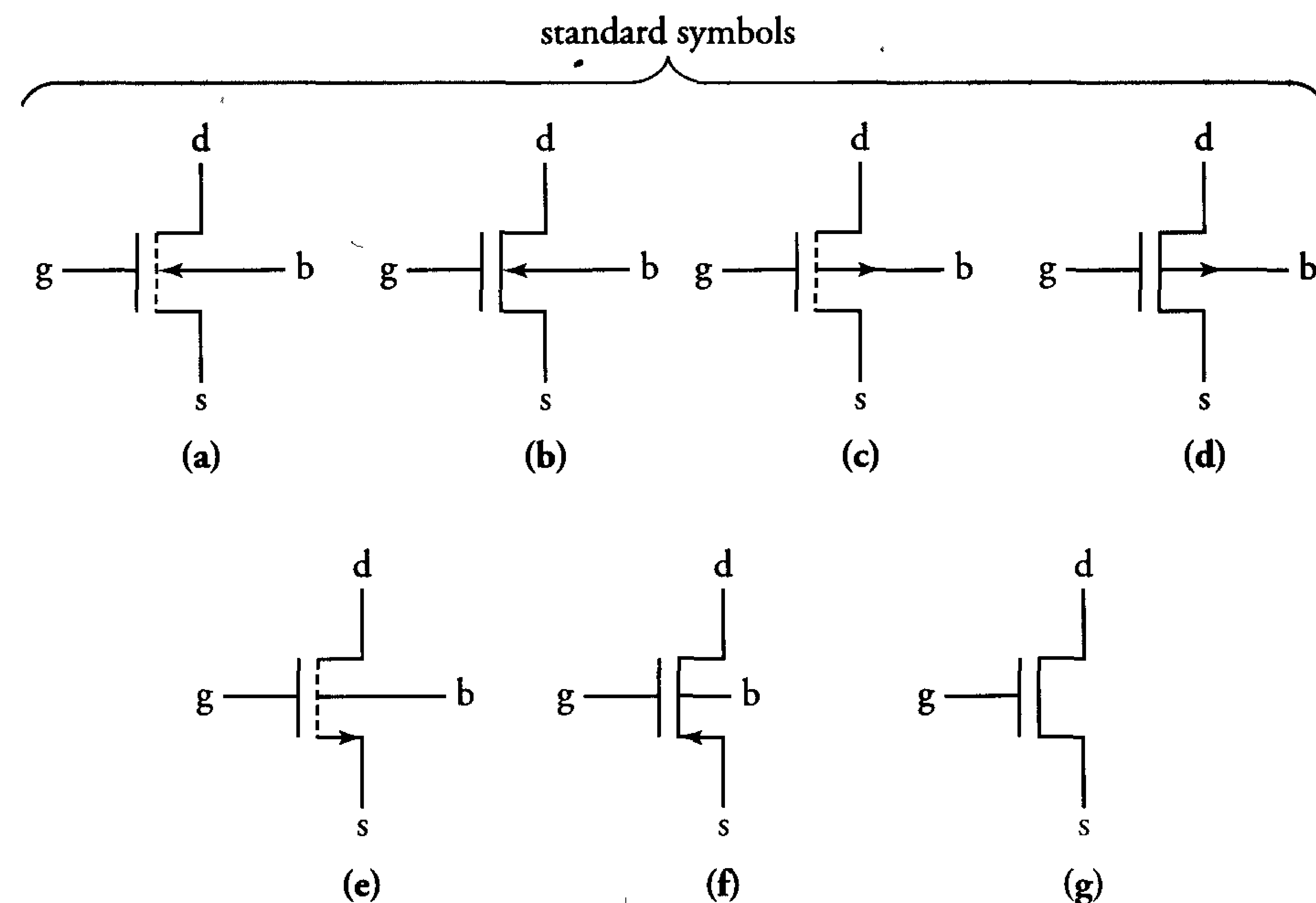


Figure 1.23: *Various transistor symbols*

Figure 1.23 shows various symbols used to represent MOS transistors. Their meanings are as follows:

a) The inward pointing arrow indicates that the transistor is n-channel and the broken line between s and d indicates that it is an enhancement transistor.

b) The solid line from s to d indicates that this n-channel transistor is a depletion device.

c) The outward pointing arrow indicates that the transistor is p-channel and the broken line between s and d indicates that it is an enhancement transistor.

d) The solid line from s to d indicates that this p-channel transistor is a depletion device.

e) This symbol for an n-channel enhancement transistor is analogous to the npn transistor symbol.

f) This p-channel transistor is by definition not necessarily an enhancement type.

g) This general symbol represents a MOS transistor of any type.

Adaptations of the above symbols are also used. MOS symbols must therefore be interpreted with caution. The following rules are generally applied:

1. A transistor symbol with a broken line between its source and drain is always an enhancement or normally-off type;
2. Arrows indicate the forward directions of the substrate-channel 'junctions'.

The symbols in figure 1.24 are used throughout this book.

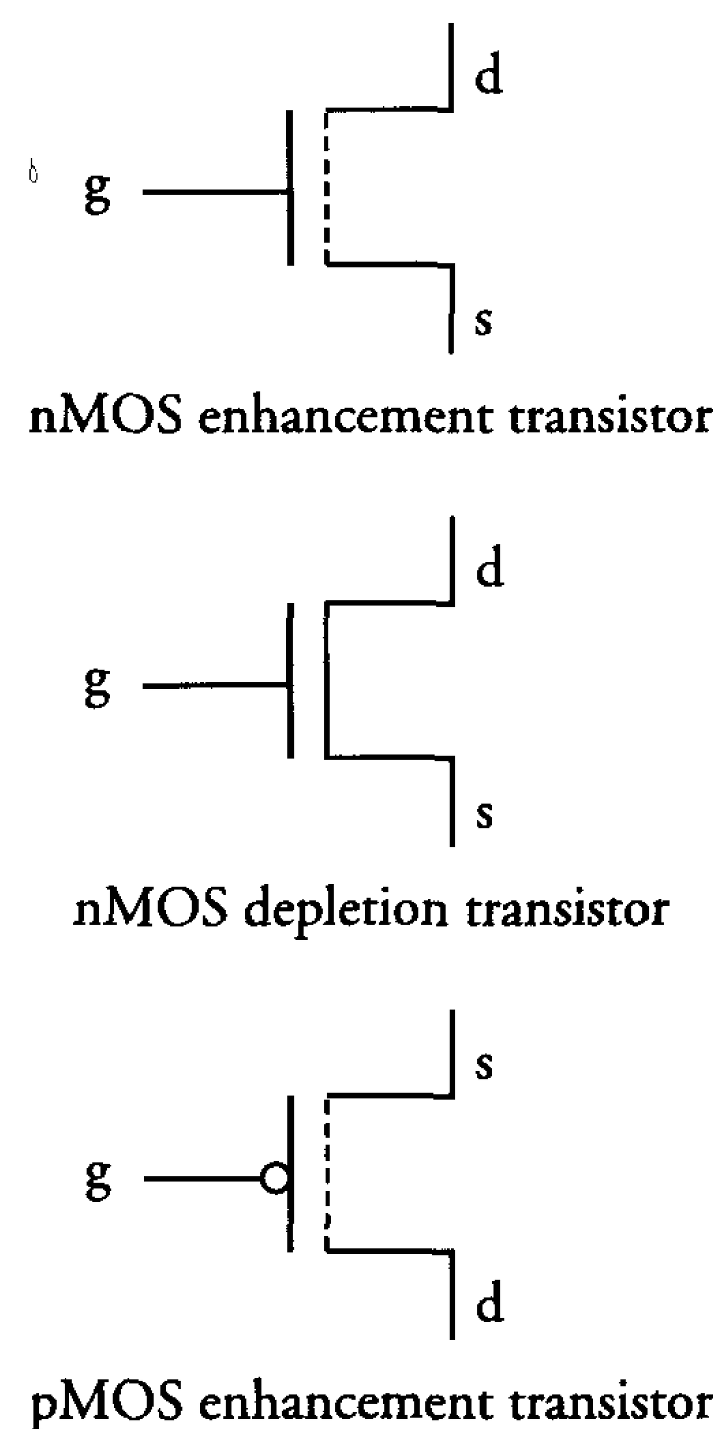


Figure 1.24: *Transistor symbols used throughout this book*

1.10 Capacitances in MOS structures

Figure 1.25 illustrates the MOS capacitance, whose value depends on such things as V_g and the frequency at which it varies. Section 1.3.1 describes the MOS capacitance and presents a qualitative discussion of its related charges, fields and voltages. Figure 1.26 shows a plot of the total capacitance C_t between the gate and ground terminals as a function of their voltage difference.

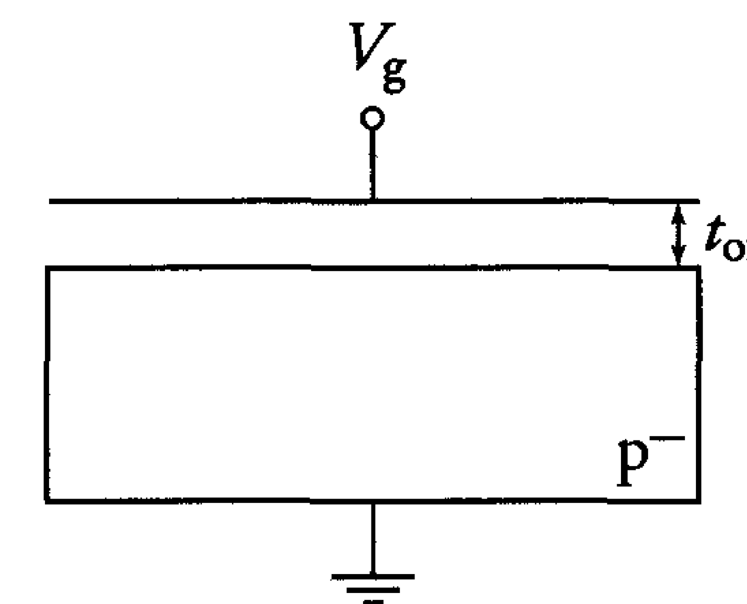


Figure 1.25: *The MOS capacitance*

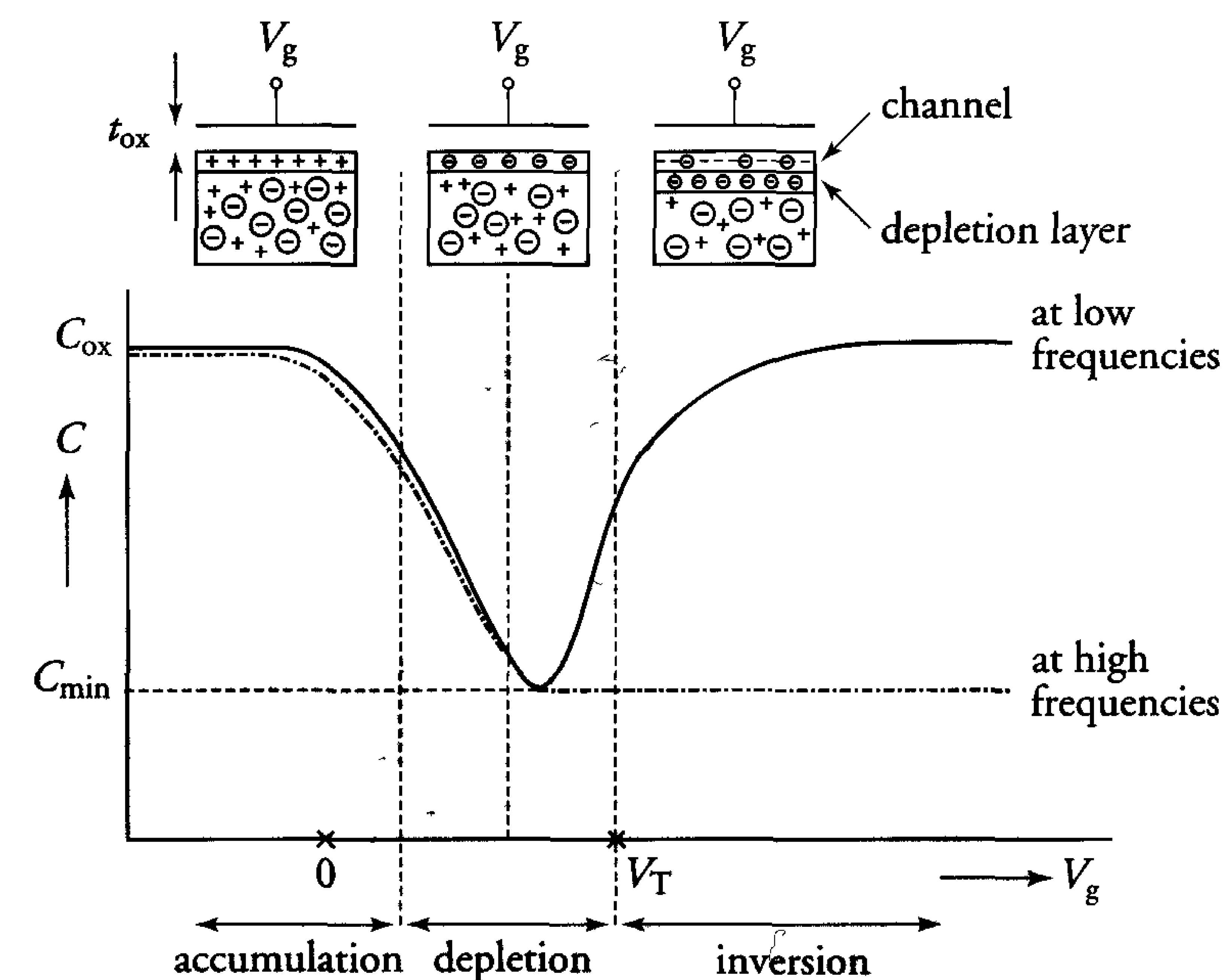


Figure 1.26: *Capacitance behaviour of a MOS structure*

The various regions of the $C - V$ *curve* in figure 1.26 are explained as follows:

1. $V_g \ll V_T$ for a p-type substrate; $V_g \gg V_T$ for an n-type substrate. Here, the surface potential ϕ_s is highly negative and majority carriers in the p-type substrate will form a surface layer of holes. This accumulation layer is thin in comparison with the oxide thickness and exists as long as V_g is much smaller than V_T. Now, the silicon behaves like a metal plate, and the MOS capacitance is equal to the oxide capacitance C_{ox}. Deviations only appear at very high frequencies ($> 1\,\text{GHz}$), where the *dielectric relaxation time* τ_R is important. For the 10Ωcm silicon, $\tau_R \approx 10\,\text{ps}$ ($=10^{-11}\,\text{s}$).

2. $V_g \approx V_T$, thus $\phi_s \approx 0...2\phi_f$.
   As V_g gradually becomes more positive, the accumulation layer decreases for a p-type substrate. A depletion layer is created under the gate when $\phi_s > 0$. A voltage change ΔV at the gate causes a change ΔQ in the charge at the edge of the depletion layer. In fact, the total capacitance is now determined by the series connection of the gate capacitance and the depletion layer capacitance. The capacitance therefore decreases.

3. $V_g \gg V_T$ for a p-type substrate; $V_g \ll V_T$ for an n-type substrate. Now, ϕ_s is highly positive and an inversion layer is created. This layer is thin compared to the oxide thickness. At low frequencies ($< 100\,\text{kHz}$), the capacitance will again be equal to the oxide capacitance C_{ox}. However, the inversion layer for a p-type substrate consists of electrons that are supplied and absorbed by the substrate. This relies on the process of thermal generation and recombination of minorities, i.e. the electrons. With a constant temperature, the speed of the generation/recombination process is limited. This accounts for the lower capacitance shown in figure 1.26 at higher frequencies ($> 1\,\text{MHz}$). At these high frequencies, the capacitance C_t will be about equal to the series connection of the gate capacitance and the depletion layer capacitance.

As discussed, the MOS capacitance can be considered as a series connection of two capacitances: the oxide capacitance C_{ox} between the gate and the silicon surface and a capacitance C_s between the silicon surface and the substrate interior. This is explained below.

The voltage V_g can be expressed as follows:

$$V_g = V_{ox} + \phi_{ms} + \phi_s \tag{1.23}$$

The law for conservation of charge yields the following equation:

$$Q_g + Q_{ox} + Q_n + Q_d = 0 \tag{1.24}$$

where:

V_{ox} = voltage across the oxide between gate and silicon surfaces;
ϕ_{ms} = contact potential between gate and substrate;
ϕ_s = surface potential of the silicon with respect to the substrate interior;
Q_g = charge on the gate;
Q_{ox} = charge in the oxide;
Q_n = charge in the inversion layer;
Q_d = charge in the depletion layer.

The following expression for a change ΔV_g in gate voltage can be derived from equation (1.23):

$$\Delta V_g = \Delta V_{ox} + \Delta\phi_s \quad (\phi_{ms} \text{ is constant, thus } \Delta\phi_{ms} = 0) \tag{1.25}$$

Substituting $Q_n + Q_d = Q_s$ in equation (1.24) yields:

$$\Delta Q_g = -\Delta Q_{ox} - \Delta Q_s \tag{1.26}$$

If Q_{ox} is considered constant, then:

$$\Delta Q_g = -\Delta Q_s \tag{1.27}$$

Equations (1.25) and (1.27) yield the following expressions:

$$\frac{\Delta V_g}{\Delta Q_g} = \frac{\Delta V_{ox}}{\Delta Q_g} + \frac{\Delta\phi_s}{\Delta Q_g} = \frac{\Delta V_{ox}}{\Delta Q_g} - \frac{\Delta\phi_s}{\Delta Q_s}$$

where:

$\frac{\Delta Q_g}{\Delta V_g} = C_t$ = the total capacitance of the MOS structure;

$\frac{\Delta Q_g}{\Delta V_{ox}} = C_{ox}$ = oxide capacitance;

$-\frac{\Delta Q_s}{\Delta\phi_s} = C_s$ = capacitance between the silicon surface and the semiconductor interior (depletion layer capacitance).

C_t can now be expressed as follows:

$$C_t = (\frac{1}{C_{ox}} + \frac{1}{C_s})^{-1} \tag{1.28}$$

Capacitance C_s is responsible for the drop in the $C - V$ curve. The value of C_s is determined by the substrate doping concentration and the potential difference across the depletion layer. The minimum value C_{min} in the $C - V$ curve is also determined by C_{ox}. A smaller C_{ox} leads to a larger $\frac{1}{C_{ox}}$ and a smaller C_{min}. C_{min} can be as low as $0.1C_{ox}$.

The $C - V$ curve is often used during MOS manufacturing processes to get a quick impression of the value of V_T.

Figure 1.27 shows a MOS capacitance with an additional n^+ area, which causes significant changes in the capacitance behaviour. The structure is in fact equivalent to a MOS transistor without a drain or to a MOS transistor with an external short circuit between its drain and source. This structure is generally called a *MOS capacitance* or a *MOS varactor*. Dynamic MOS circuits, in particular, use this device very often.

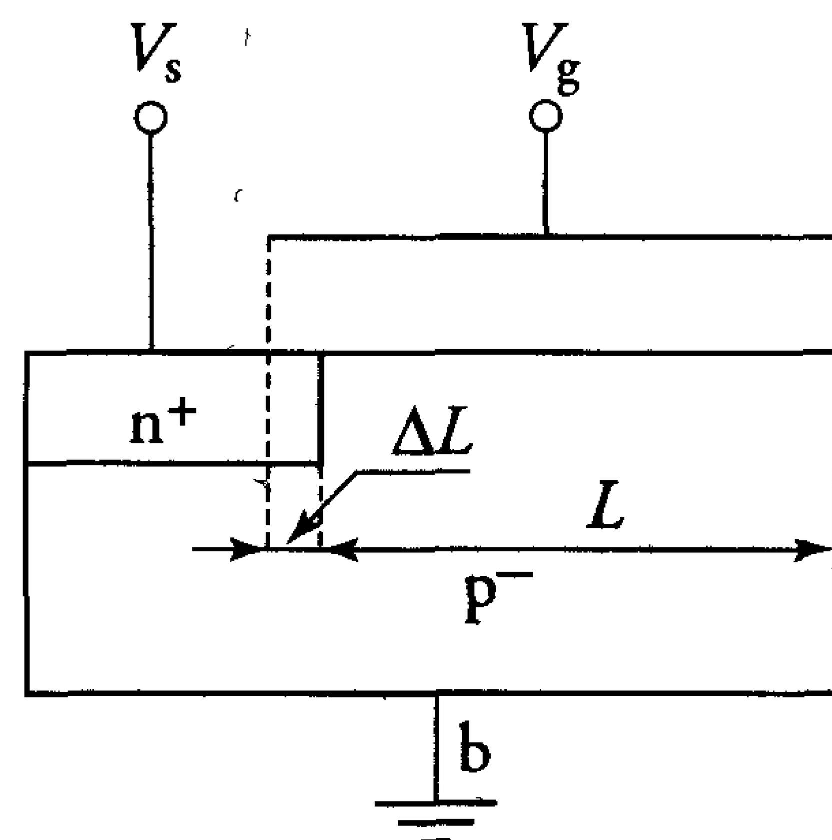


Figure 1.27: *MOS capacitance with source and/or drain area*

While $V_{gs} < V_T$, there is no inversion layer in a MOS capacitance, and the behaviour of the gate capacitance is unchanged. However, an inversion layer is created when $V_{gs} > V_T$. The electrons in this inversion layer are supplied by the n^+ area instead of by thermal generation/recombination processes of minorities in the substrate. This n^+ area can generate and absorb electrons at very high frequencies (> 1 GHz). Therefore,

C_t will now equal C_{ox} under all normal operating conditions. In this case, C_t represents the capacitance between the gate and source, i.e. $C_t = C_{gs} = C_{ox}(L + \Delta L) \cdot W$.

The dependence of the capacitance C_{gs} on the applied voltage V_{gs} is summarised as follows:

- When $V_{gs} < V_T$, there is no inversion layer. Here, the value of C_{gs} is determined by the channel width W and the gate overlap ΔL on the source/drain area: $C_{gs} = \Delta L \cdot W \cdot C_{ox}$.
- When $V_{gs} > V_T$, there is an inversion layer. Here, C_{gs} is determined by the channel length L: $C_{gs} = (L + \Delta L) \cdot W \cdot C_{ox}$.

The above non-linear behaviour of $C_{gs} = f(V_{gs})$ is shown in figure 1.28.

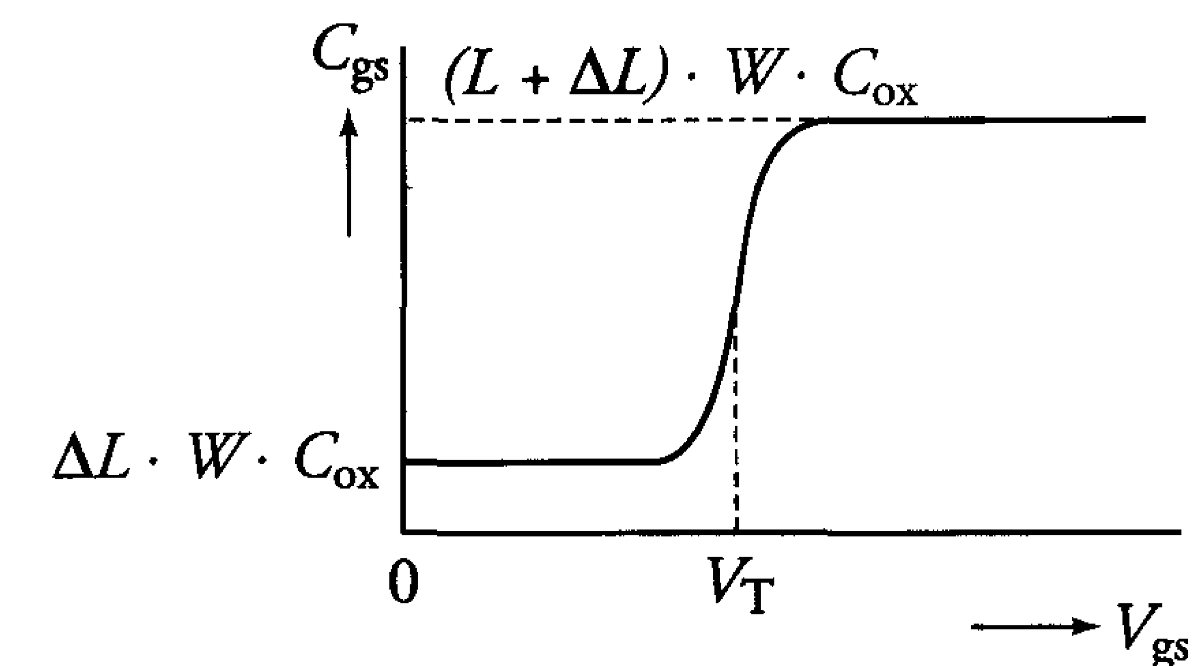


Figure 1.28: *Non-linear behaviour of a MOS capacitance*

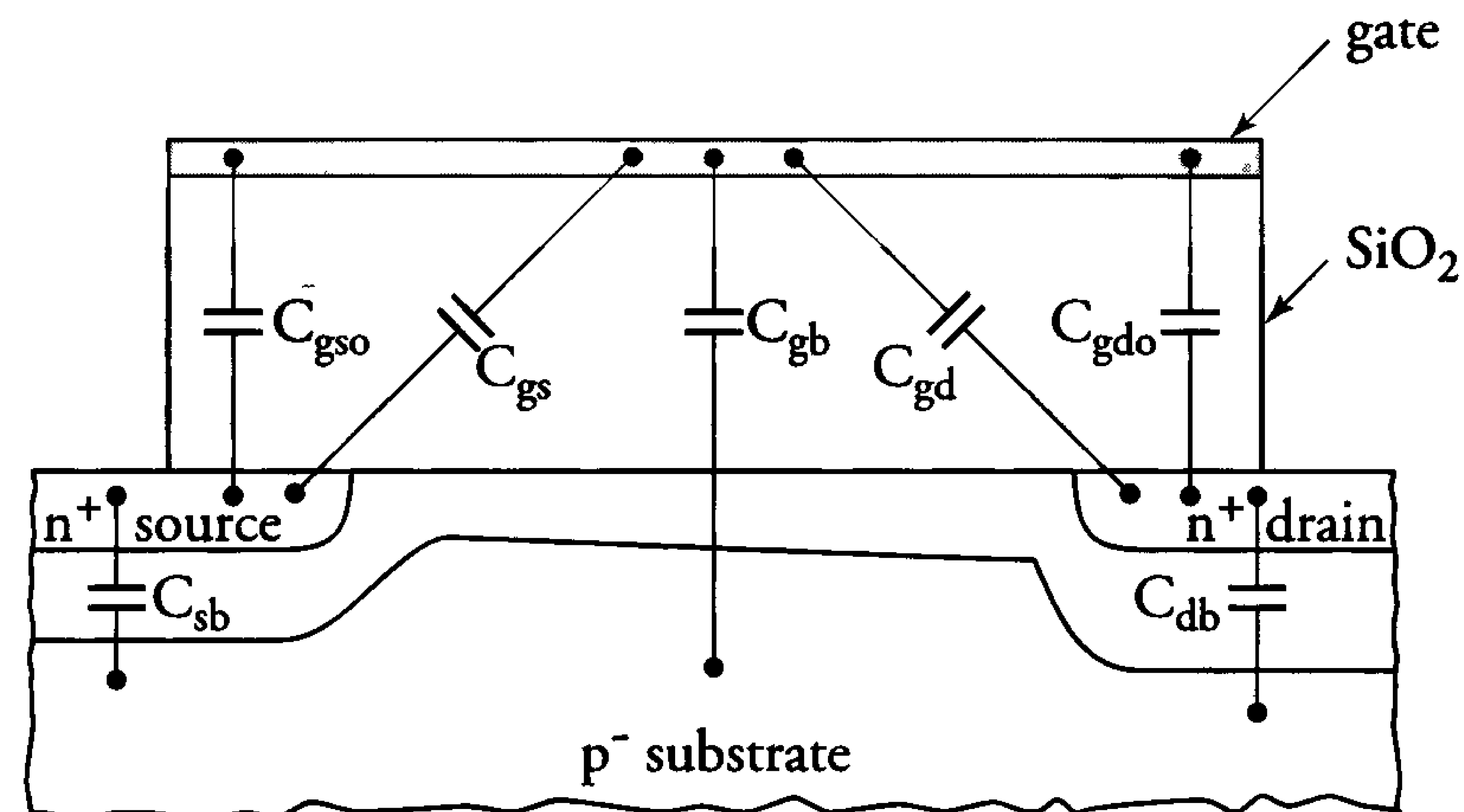
Note: There is no inversion layer when $V_{gs} < V_T$. Figure 1.26 shows how the gate-substrate capacitance then behaves.

Figure 1.29 shows the large number of capacitances in a real MOS transistor. These capacitances, which are largely non-linear, are defined as follows:

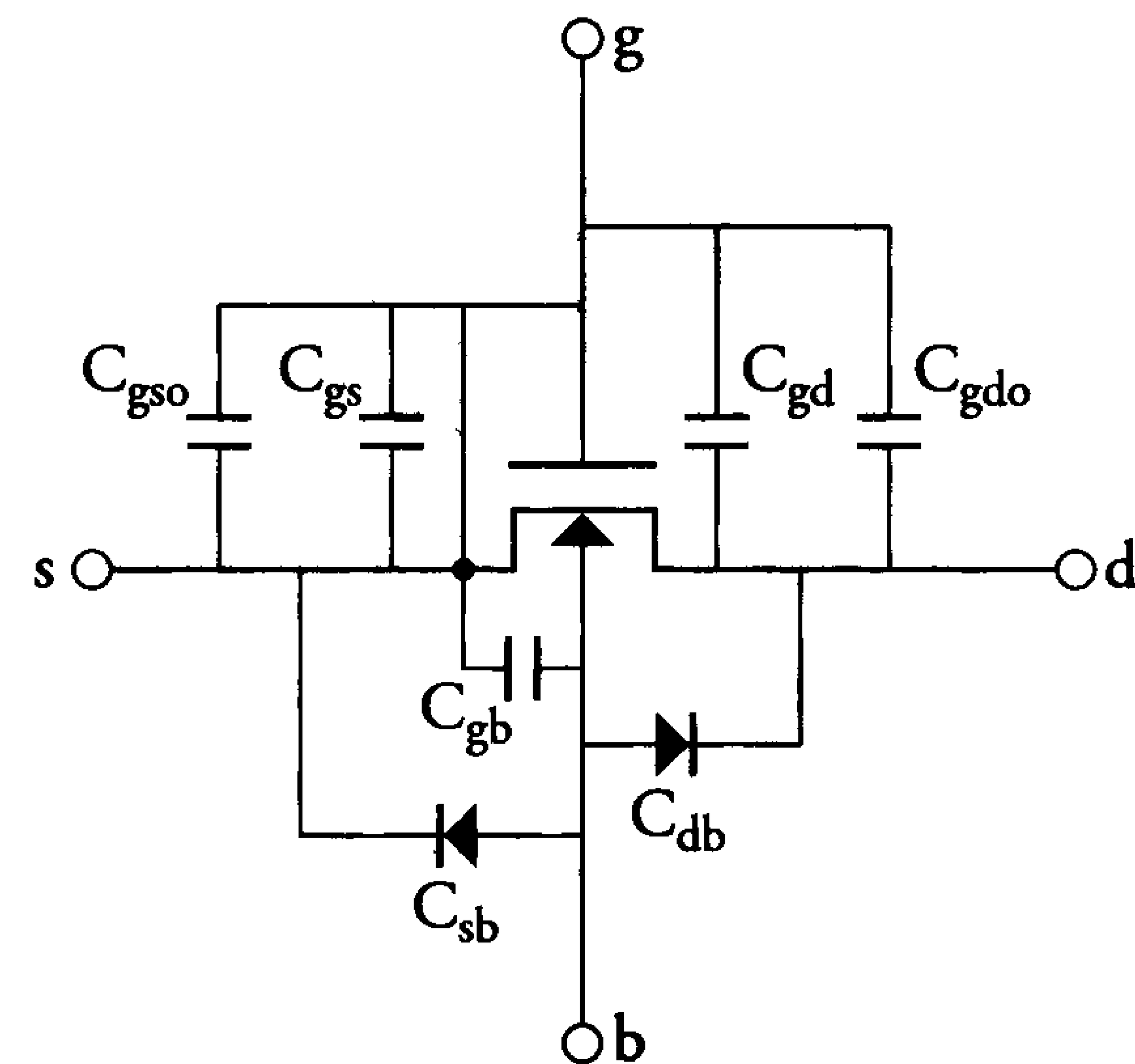
C_{db}, C_{sb}	: drain-substrate and source-substrate capacitances, which are non-linearly dependent on V_{db} and V_{sb}, respectively.
C_{gdo}, C_{gso}	: gate-drain and gate-source capacitances, which are voltage-independent.
C_{gd}, C_{gs}	: gate-drain and gate-source capacitances (via the

inversion layer), which are non-linearly dependent on V_{gs}, V_{gd} and V_{gb}.

C_{gb} : gate-substrate capacitance, which is non-linearly dependent on V_{gb}.



(a)



(b)

Figure 1.29: *Capacitances in a MOS transistor*

The values of the C_{db} and C_{sb} diode capacitances in figure 1.29 are expressed as follows:

$$C(V) = \frac{C_o}{(1 + \frac{V}{V_j})^{1/m}} \tag{1.29}$$

where:

C_o = capacitance when $V=0$;
V_j = junction voltage (0.6 V to 0.9 V);
m = grading factor, $2 \leq m \leq 3$: $m = 2$ for an abrupt junction and $m = 3$ for a linear junction.

Terms C_{gdo} and C_{gso} represent gate overlap capacitances that are determined by the transistor width, the length of the overlap on the drain and source areas, and the thickness of the gate oxide. These capacitances are clearly voltage-independent.

The gate-substrate capacitance C_{gb} is only important if $V_{gs} \ll V_T$. Now, C_{gb} is often expressed as $C_{gb} \approx (0.12 \text{ to } 0.2) \cdot W \cdot L \cdot C_{ox}$. The inversion layer shields the substrate from the gate and $C_{gb}=0$ when $V_{gs} \geq V_T$.

Terms C_{gd} and C_{gs} represent gate-drain and gate-source capacitances, respectively, which are present via the inversion layer (figure 1.28). The values of these capacitances depend strongly on the bias voltage on the terminals of the MOS transistor. The following cases are distinguished:

Case a $V_{gs} < V_T$; no inversion layer, thus $C_{gd}=C_{gs}=0$.

Case b $V_{gs} > V_T$ and $V_{ds}=0$.
For reasons of symmetry, $C_{gs}=C_{gd}=\frac{1}{2} \cdot W \cdot L \cdot C_{ox}$.

Case c $V_{gs} > V_T$ and $V_{ds} > V_{d_{sat}}$ ($V_{d_{sat}} = V_{gs} - V_T$).
The transistor is in saturation and there is no inversion layer at the drain: $C_{gd}=0$ and $C_{gs}= \frac{2}{3} \cdot W \cdot L \cdot C_{ox}$.
This expression for C_{gs} is derived below.

Case d $V_{gs} > V_T$ and $0 < V_{ds} < V_{d_{sat}}$.
In this case, a linear interpolation between the values in cases b and c closely corresponds to the actual values, which are shown in figure 1.30.

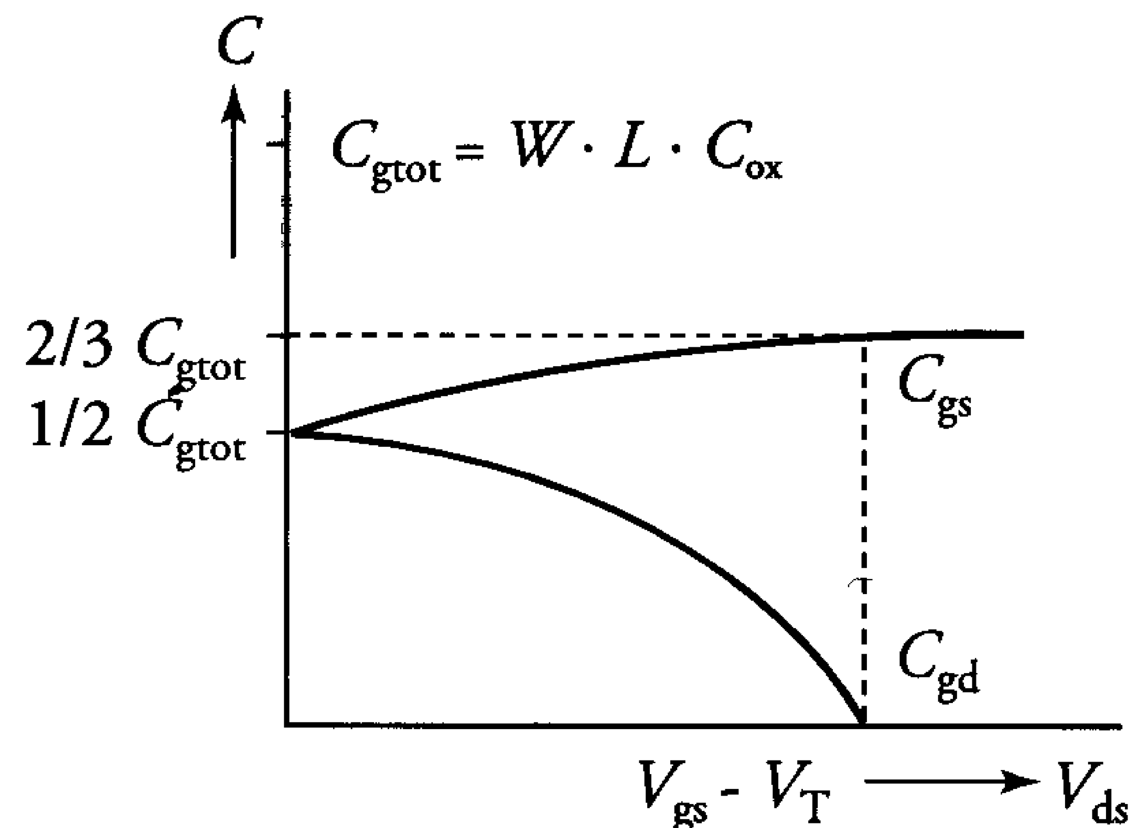


Figure 1.30: *C_{gs} and C_{gd} dependence on V_{ds} for $V_{gs} > V_T$*

The above expression in case c for the gate-source capacitance C_{gs} of a saturated MOS transistor is explained with the aid of figure 1.31. This figure shows a cross-section of a MOS transistor biased in the saturated region. The channel does not reach the drain area, but stops at a point where the channel potential is exactly $V_{gs}-V_T$.

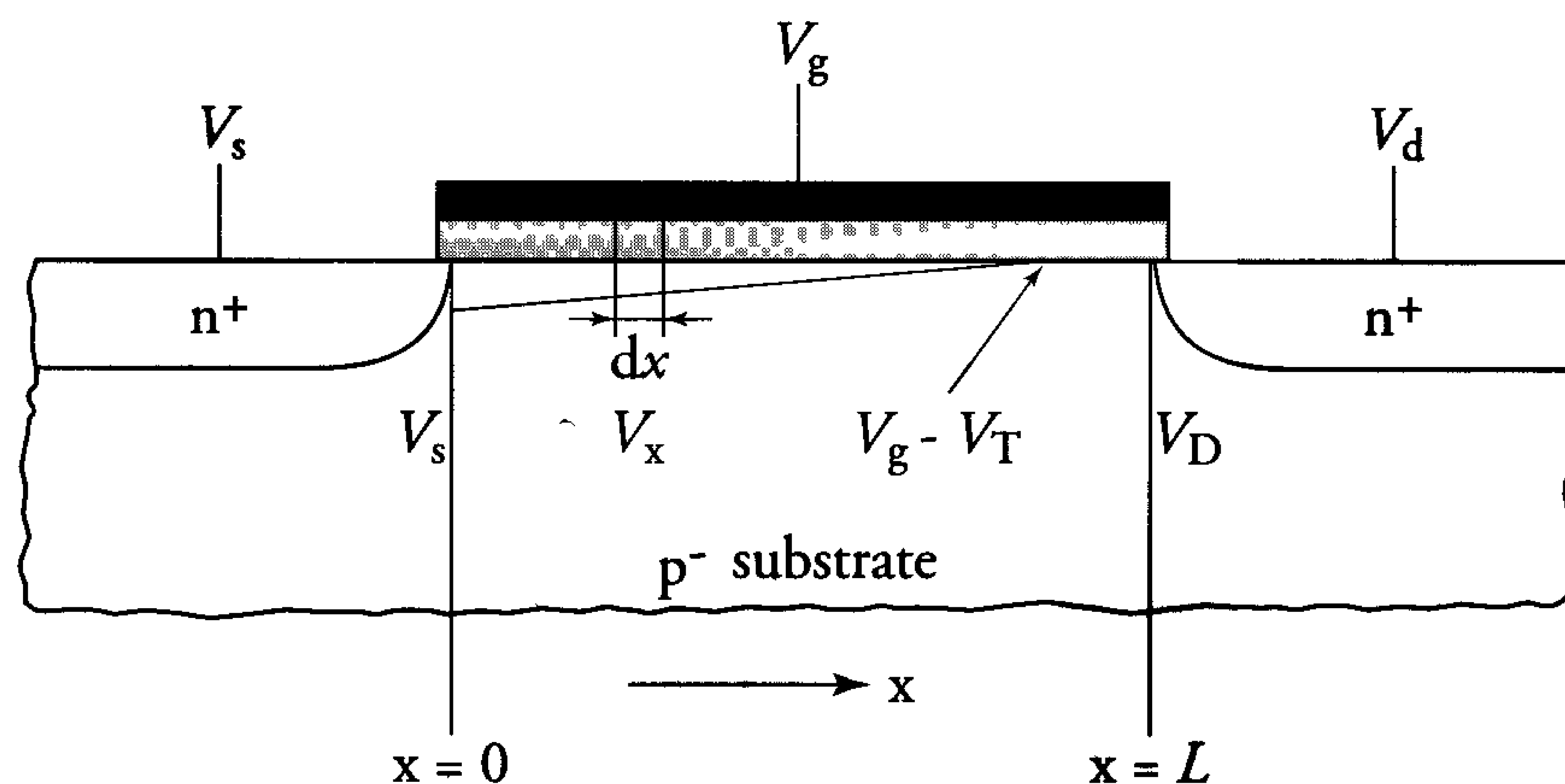


Figure 1.31: *Cross-section of a saturated MOS transistor.* $C_{gd} = 0$ *and* $C_{gs} = \frac{2}{3} \cdot W \cdot L \cdot C_{ox}$.

Equation (1.5) leads to the following expression for the charge dQ in a channel section of length dx at position x:

$$dQ(x) = Q_n \cdot W \cdot dx = -W \cdot C_{ox}[V_{gs} - V_T - V(x)] \cdot dx \qquad (1.30)$$

The following expression for dx is derived from equation (1.9):

$$dx = \mu_n \cdot C_{ox} \cdot W \cdot [V_{gs} - V_T - V(x)] \cdot \frac{dV(x)}{I_{ds}} \qquad (1.31)$$

Combining equations (1.30) and (1.31) yields the following expression for $dQ(x)$:

$$dQ(x) = \frac{\mu_n \cdot {C_{ox}}^2 \cdot W^2 \cdot [V_{gs} - V_T - V(x)]^2}{I_{ds}} \cdot dV(x) \qquad (1.32)$$

Equation (1.14) yields the following expression for the drain current I_{ds} in a saturated MOS transistor:

$$I_{ds} = \frac{\beta}{2} \cdot (V_{gs} - V_T)^2 = \frac{\mu_n \cdot C_{ox}}{2} \cdot \frac{W}{L} \cdot (V_{gs} - V_T)^2 \qquad (1.33)$$

Substituting equation (1.33) in equation (1.32) yields:

$$dQ(x) = \frac{C_{ox} \cdot W \cdot L \cdot 2 \cdot [V_{gs} - V_T - V(x)]^2}{(V_{gs} - V_T)^2} \cdot dV(x) \qquad (1.34)$$

Integrating equation (1.34) from the source to the imaginary drain gives:

$$Q = \int_{V_s}^{V_{gs}-V_T} \frac{C_{ox} \cdot W \cdot L \cdot 2 \cdot [V_{gs} - V_T - V(x)]^2}{(V_{gs} - V_T)^2} \cdot dV(x)$$

$$= \frac{C_{ox} \cdot W \cdot L \cdot 2}{(V_{gs} - V_T)^2} \cdot \left[-\frac{1}{3} \cdot [V_{gs} - V_T - V(x)]^3 \right|_{V_s}^{V_{gs}-V_T}$$

$$\Rightarrow Q = \frac{2}{3} \cdot W \cdot L \cdot C_{ox} \cdot (V_{gs} - V_T) \qquad (1.35)$$

The gate-source capacitance C_{gs} can be found by differentiating Q in equation (1.35) with respect to V_{gs}:

$$C_{gs} = \frac{dQ}{dV_{gs}} = \frac{2}{3} \cdot W \cdot L \cdot C_{ox} \qquad (1.36)$$

The C_{gs} of a saturated MOS transistor is therefore only two thirds of the total value, while the gate-drain capacitance is zero.

In summary:
Most capacitances in a MOS transistor are non-linearly dependent on the externally applied voltages. For each capacitance, these dependencies are as follows:

1. The diode capacitances C_{db} and C_{sb}:

   $C(V) = \frac{C_o}{(1+\frac{V}{V_j})^{1/m}}$, where $V_j \approx 0.6 \ldots 0.9\,\text{V}$ and $2 \leq m \leq 3$.

2. Figure 1.28 shows the voltage dependence of gate-channel capacitances C_{gd} and C_{gs} when the drain and source are short circuited, as is the case in a MOS capacitance. Figure 1.30 shows the voltage dependence of C_{gd} and C_{gs} when the drain and source are at different voltages, i.e. during normal transistor operation.

3. The gate-substrate capacitance C_{gb} is 0 when $V_{gs} > V_T$ and $C_{gb} = 0.2 \cdot W \cdot L \cdot C_{ox}$ if $V_{gs} < V_T$.

4. The overlap capacitances C_{gdo} and C_{gso} are the only capacitances which are not dependent on the externally applied voltages.

1.11 Conclusions

The basic principles of the operation of the MOS transistor can be explained in different ways. The fairly simple approach adopted in this chapter should provide a good fundamental understanding of this operation. The current-voltage characteristics presented are derived by means of the simplest mathematical expressions for MOS transistor behaviour.

Second-order and parasitic effects are not essential to an understanding of the basic principles of MOS transistor operation. They have therefore been neglected in this chapter. However, these effects should be included in accurate descriptions of MOS transistors and are therefore discussed in chapter 2. Most of these effects are included in the MOS transistor models used by commonly-used circuit simulation programs.

1.12 References

General basic physics

[1] R.S.C. Cobbold,
'Theory and applications of field effect transistors',
Wiley, New York

[2] S.M. Sze,
'Physics of Semiconductor Devices',
2nd edition, Wiley, 1981

[3] A.S. Grove,
'Physics and Technology of Semiconductor Devices',
Wiley, New York

[4] Y.P. Tsividis,
'Operation and modelling of the MOS transistor',
Mc Graw-Hill, 1987

[5] C. Kittel,
'Introduction to Solid State Physics',
Wiley, 1976, New York

MOS capacitances

[6] E.W. Greenwich,
'An Analytical Model for the gate Capacity of Small-Geometry MOS structures',
IEEE Transactions on Electron Devices,
ED-30, pp 1838-1839, 1983

[7] J.J.Paulos, D.A. Antoniadis, and Y.P. Tsividis,
'Measurement of Intrinsic Capacitances of MOS Transistors',
ISSCC Digest of technical papers, pp 238-239, 1982

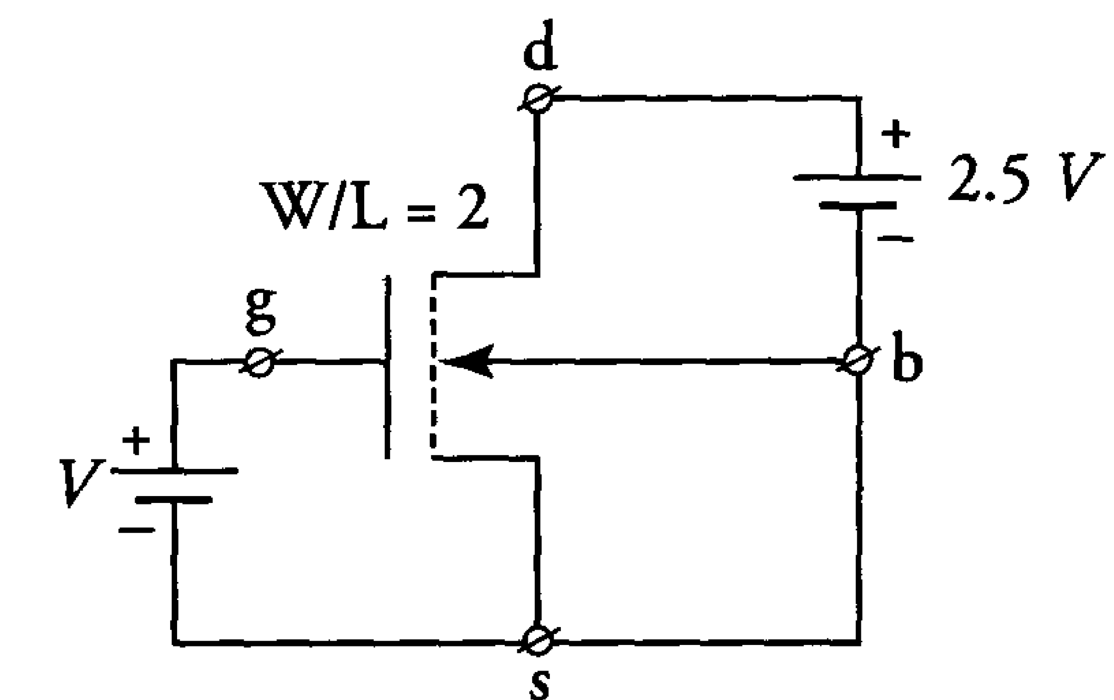
[8] D.E. Ward and R.W. Dutton,
'A Charge-Oriented Model for MOS Transistor Capacitances',
IEEE Journal of Solid-State Circuits, pp 703-707, 1978

1.13 Exercises

Note: $2\phi_f$=0.64 V throughout these exercises.

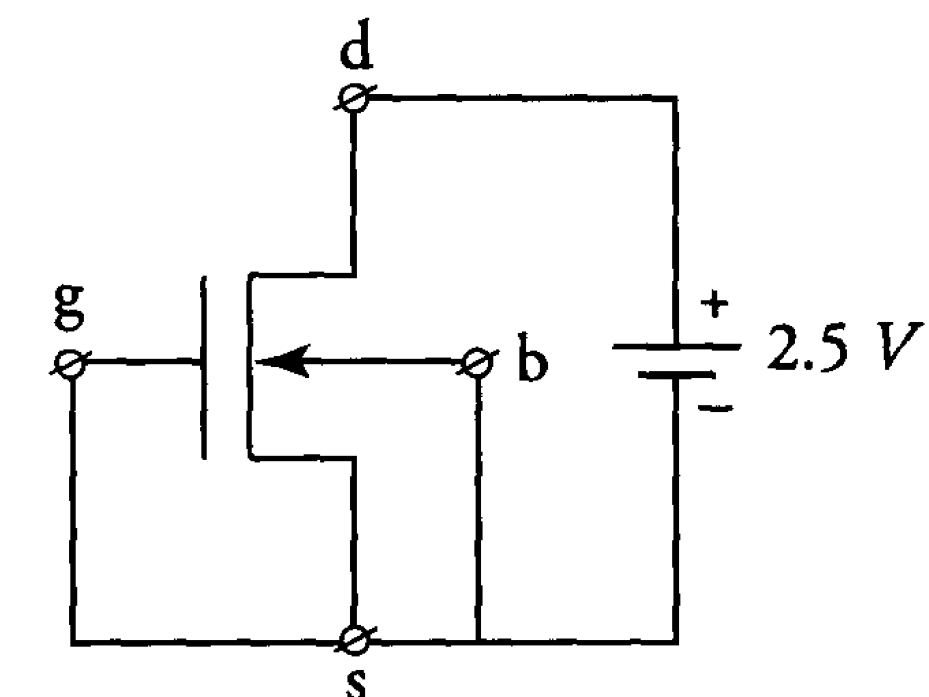
1. What happens to the depletion layer in figure 1.12 when the substrate (b) is connected to a negative voltage (≈ -2 V) instead of ground?
   What effect does this have on the threshold voltage V_T ?

2. Current I_{ds} in a transistor ($\frac{W}{L}$=2) is 62.5 μA when its gate-source voltage V is 1 V. The current is 1 mA when V = 2.5 V.



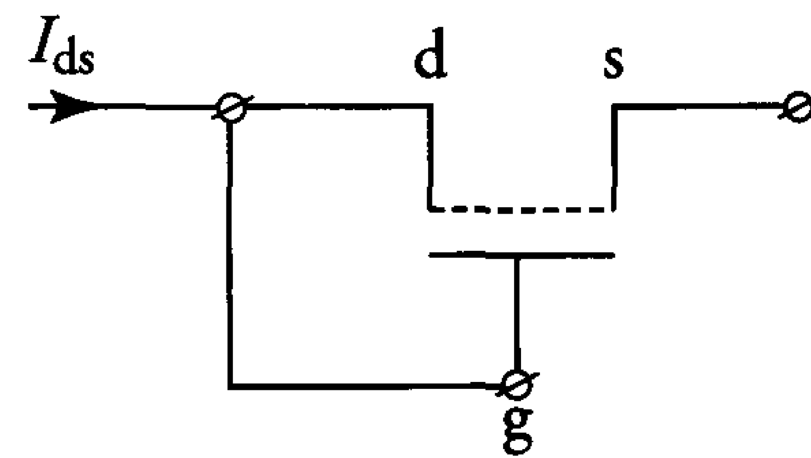
   a) Which transistor operating regions (linear or saturated) do these values of V correspond to?
   b) Calculate $\beta_{\square}$ and V_T for the given transistor.

3. Given:



   a) What type is the transistor shown?
   b) Calculate I_{ds} when this transistor has the same β as the transistor in exercise 2 and V_T= -2 V.

4. Given:

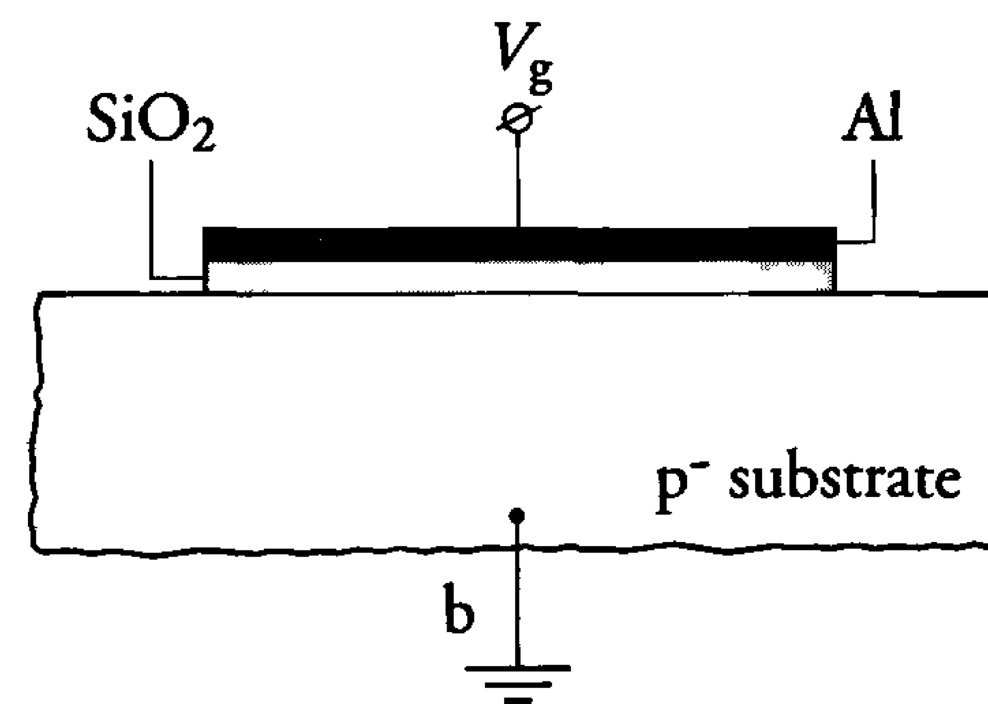


If this is an n-type enhancement MOS transistor and the current $I_{ds} > 0$, explain the following:

a) This transistor is always in its saturation region.

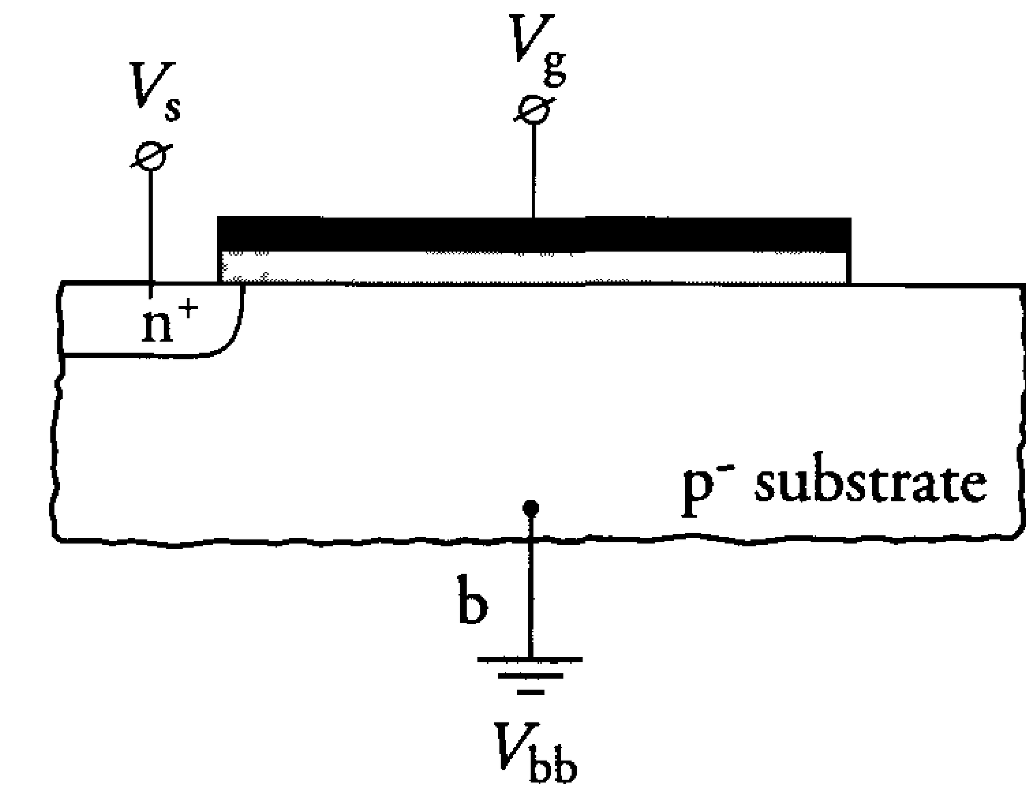
b) This connection is often called a MOS diode.

5. For this exercise, the threshold voltage V_T is 0.5 V. There is *no* thermal generation of electron/hole pairs.



a) The above structure exists when the source and drain areas of an nMOS transistor are excluded.
Copy this structure and include the possible depletion and inversion layers for the following values of V_g: −1 V, 0.25 V, 1 V and 2.5 V.

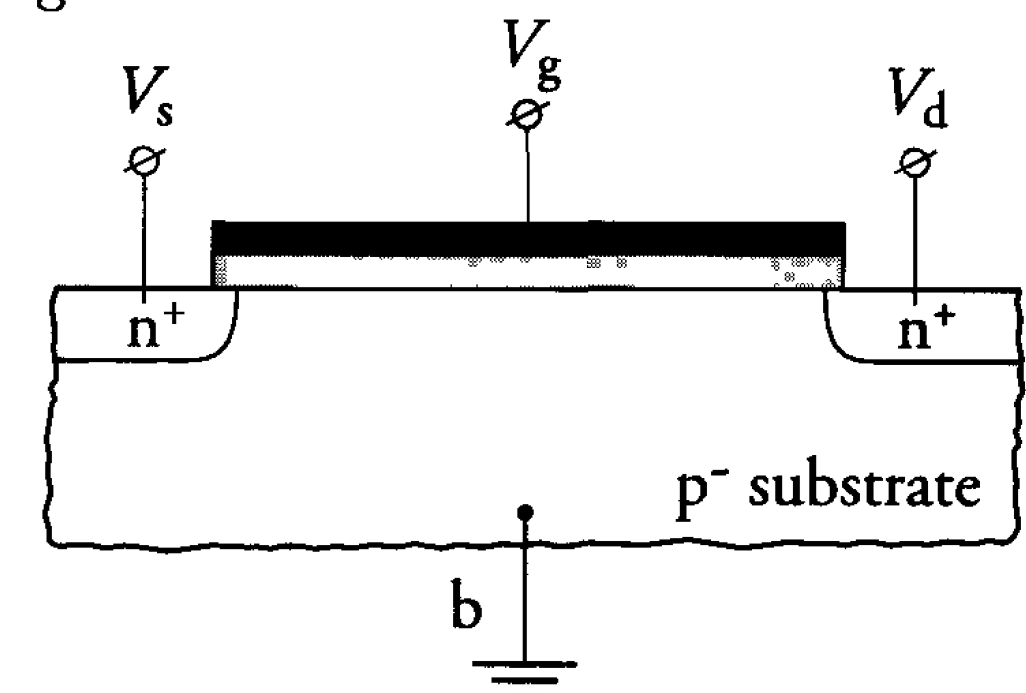
b) An n^+ area is now added to the structure in exercise 5a.



Repeat exercise 5a for $V_s = 0$ V and for $V_s = 1$ V.

c) The substrate of the structure in exercise 5b is connected to a negative voltage: V_{bb}=−2 V.
What happens to the depletion and inversion layers if $V_s = 0$ V and $V_g = 1$ V?

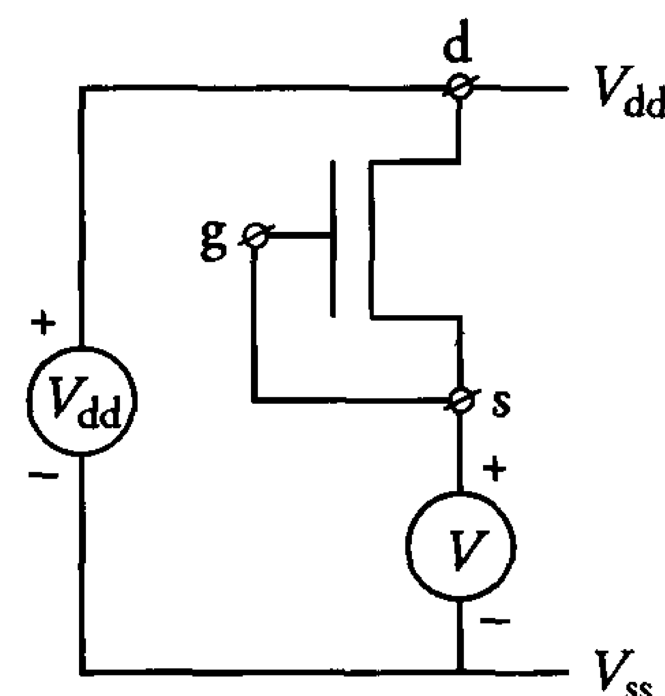
d) A second n^+ area is added to the structure of exercise 5b to yield the following structure.



Repeat exercise 5a for $V_s = 0$ V and $V_d = 1.5$ V.

e) In practice, there are thermally-generated electron hole pairs in the silicon substrate. The resulting free electrons in the depletion layer move in the opposite direction to the applied external electric field.
Draw the direction of movement of the thermally-generated electrons and holes for V_g=2.5 V in the structure of exercise 5a.
If this situation continues for a longer period, a new equilibrium is reached and the electrons and holes accumulate in the structure. Draw this situation.

6. The following values apply in the figure shown: V_{dd}=2.5 V, β=1 mA/V^2, V_x=−1.5 V, V_{bb}=−2 V.



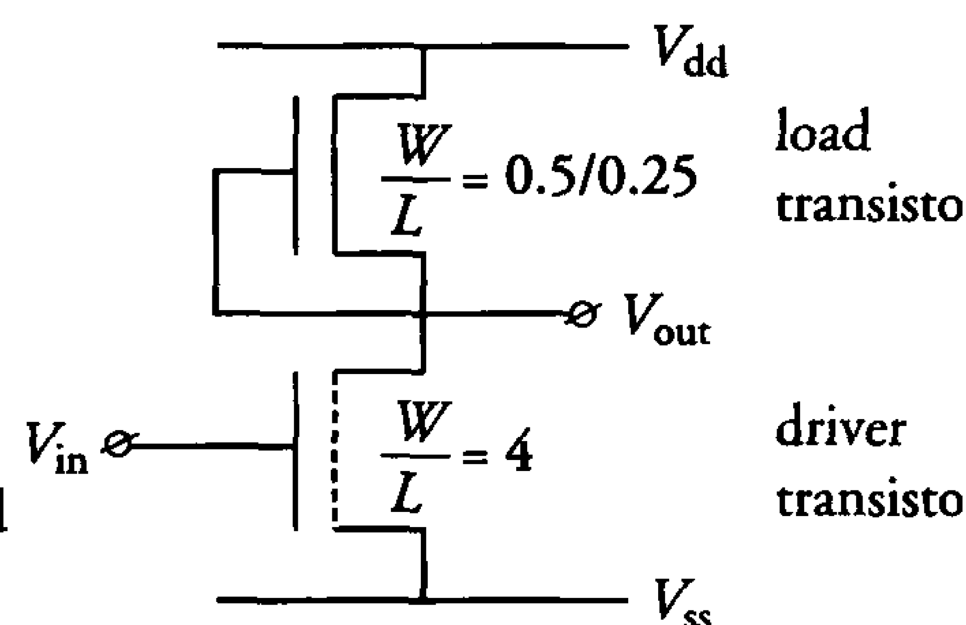
a) What type is the transistor and why?

b) Calculate and draw the graph $I_{ds}=f(V_{ds})$ for K=0 $V^{1/2}$ and V_{ds}=0, 0.5, 1, 1.5, 2 and 2.5 V.

c) Repeat b) for K=0.3 $V^{1/2}$.

d) Assuming K=0.3 $V^{1/2}$, calculate the output impedance of the transistor for V_{ds}=50 mV and for V_{ds}=1 V. (Note: the drain remains at 2.5 V).

7. The following values apply for the circuit shown: V_{dd} = 2.5 V, V_{bb} = −1.5 V, V_{ss} = 0 V, K=0.3 $V^{1/2}$, $\beta_\square$ = 240 μA/V^2, V_{X_L} = −1.5 V and V_{X_D} = 0.3 V.



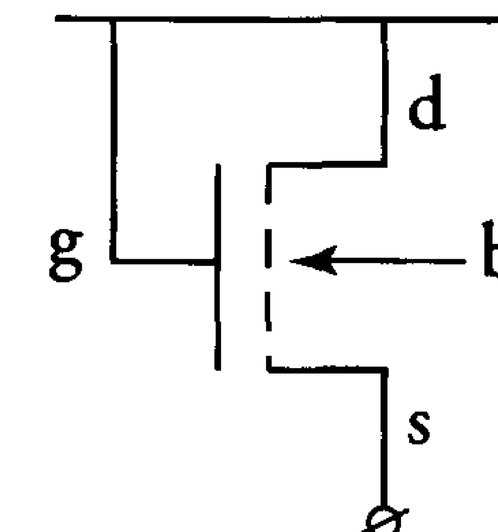
a) Calculate V_{out} for V_{in}=2.5 V.

b) Determine the transconductance of both MOS transistors for this situation.

c) What value does V_{out} reach when V_{in}=0.25 V?

d) The same low output level must be maintained when the load transistor is replaced by an enhancement-type transistor with its gate at V_{dd}. Does this require a driver transistor with the same $\frac{W}{L}$ and with a smaller or a larger channel width W? Explain your answer.

8. The aspect ratio of this transistor is $W/L = 0.4/0.25$. Results of measurements on it are summarised in the following table:



V_{sb}[V]	I_{ds}[μA]		
	V_{gs}= 1V	V_{gs}= 2V	V_{gs}= 2.5V
0	50	512	–
1.32	19	–	

a) Determine V_x, K and $\beta_\square$ for this transistor.

b) Calculate and draw the graph $V_T=f(V_{sb})$ for at least five V_{sb} values (0 V<V_{sb}<4 V).

9. Define an expression for the transconductance with respect to the substrate voltage V_{sb} when the transconductance with respect to the normal gate voltage is defined as $g_m=\frac{\delta I_{ds}}{\delta V_{gs}}$.

Chapter 2

Physical and geometrical effects on the behaviour of the MOS transistor

2.1 Introduction

The simple formulae derived in sections 1.4 and 1.5 account for the first-order effects which influence the behaviour of MOS transistors. Until the mid-seventies, formulae (1.17) appeared quite adequate for predicting the performance of MOS circuits. However, these transistor formulae ignore several physical and geometrical effects which significantly degrade the behaviour of MOS transistors. The results are therefore considerably more optimistic than the actual performance observed in MOS circuits. The deviation becomes more significant as MOS transistor sizes decrease in VLSI circuits.

This chapter contains a brief overview of the most important effects which degrade the performance of MOS devices. The temperature effects described are first order. The remaining effects are second-order.

2.2 The zero field mobility

As discussed in chapter 1, the MOS transistor current is heavily determined by the *gain factor* β of the transistor:

$$\beta = \frac{W}{L} \cdot \beta_{\square} = \frac{W}{L} \cdot \mu \cdot C_{ox} \tag{2.1}$$

where W and L represent the transistor channel width and length respectively. C_{ox} represents the gate oxide capacitance per unit of area and μ represents the actual *mobility factor* of the carriers in the channel. This mobility can be quite different from the zero-field or substrate mobility factor μ_0, which depends on the doping concentration in the substrate. Figure 2.1 shows electron and hole mobilities in silicon at room temperature as a function of the doping concentration.

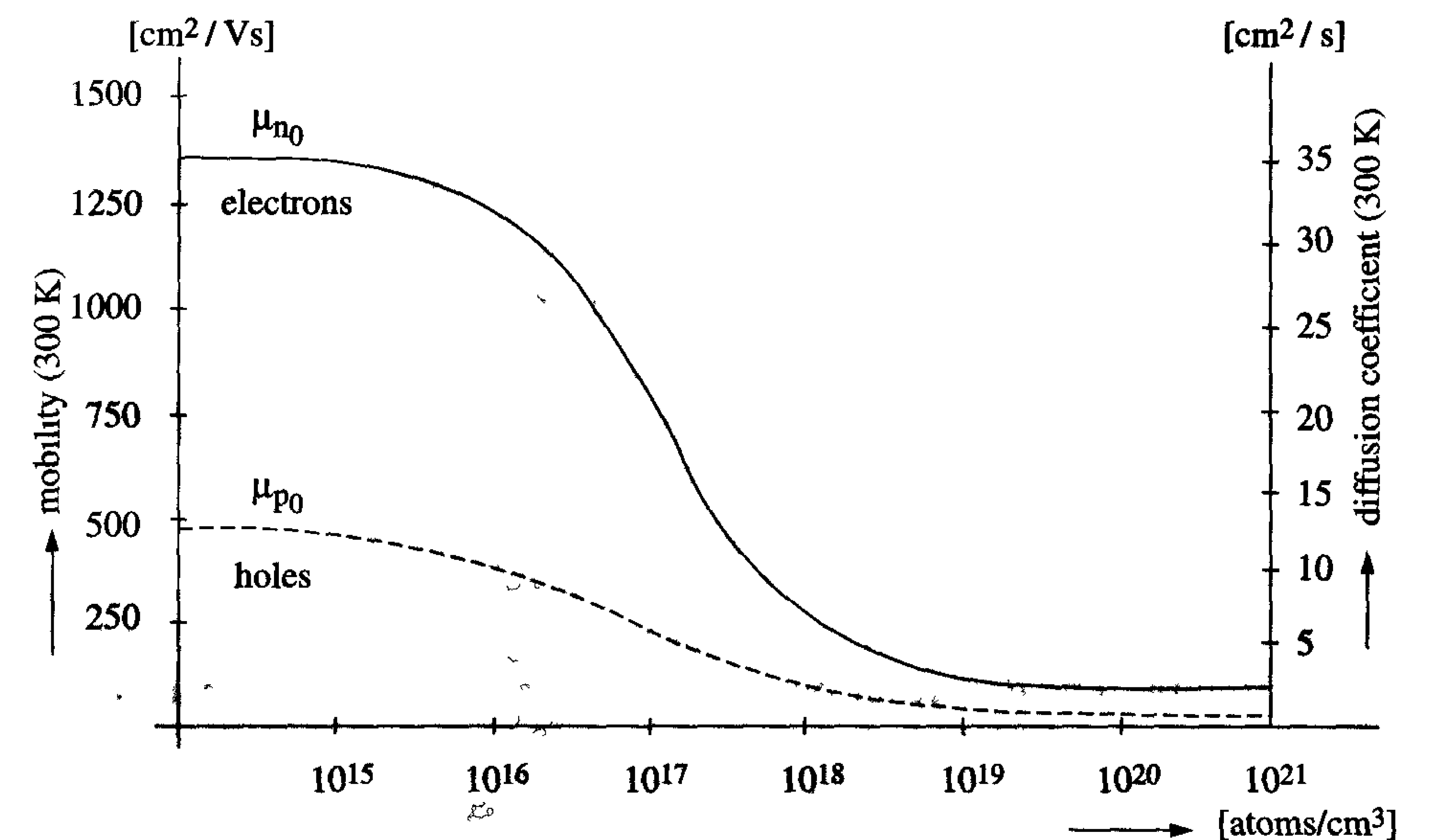


Figure 2.1: *Carrier mobility and diffusion coefficient as a function of doping concentration in silicon at room temperature*

For a substrate doping concentration of 10^{17} atoms/cm^3, the mobility of electrons (μ_{n_0}) is about three times that of holes (μ_{p_0}), in the absence of an electric field. This is the major reason for the β_n of an nMOS transistor being about three times as high as the β_p of an equally sized pMOS transistor. However, several effects dramatically reduce the mobility of the carriers in the channel. These are discussed in the next subsection.

2.3 Carrier mobility degradation

During normal transistor operation, electrical fields are applied in both the lateral (horizontal) and vertical directions, which influence the mobility of the charge carriers in the channel. Moreover, when the chip temperature is increased, either by an increase of the ambient temperature or by its own dissipation, this will have a negative effect on the carrier mobility and thus on the β of each transistor.

2.3.1 Temperature-dependent carrier mobility reduction

An increase in the operating temperature of a MOS transistor affects its behaviour in two different ways:

1. The mobility of the majority carriers, e.g. electrons in an nMOS transistor, in the channel decreases. Consequently, the transistor gain factor $\beta_\square$ also decreases. Its temperature dependence is expressed as follows [1]:

$$\beta_\square(T) = \beta_\square(298\,\mathrm{K}) \cdot \left(\frac{298}{Temp}\right)^{3/2} \tag{2.2}$$

2. The threshold voltage V_T of both nMOS and pMOS transistors decreases slightly. The magnitude of the influence of temperature change on threshold voltage variation ΔV_T depends on the substrate doping level. A variation of -1 mV/°C is quite typical.

As a result of these effects, a temperature increase of 80°C may cause a 30% decrease in transistor gain and a drop of about 240 mV in threshold voltage. In conclusion, an increase in temperature reduces the current through MOS transistors and leads to a reduction in the maximum operating speed.

2.3.2 Vertical and lateral field carrier mobility degradation

During normal operation, the effective mobility μ of the carriers in the transistor channel is degraded by components indicated in figure 2.2. These include the vertical electric field E_z, the horizontal electric field E_x and the carrier velocity v.

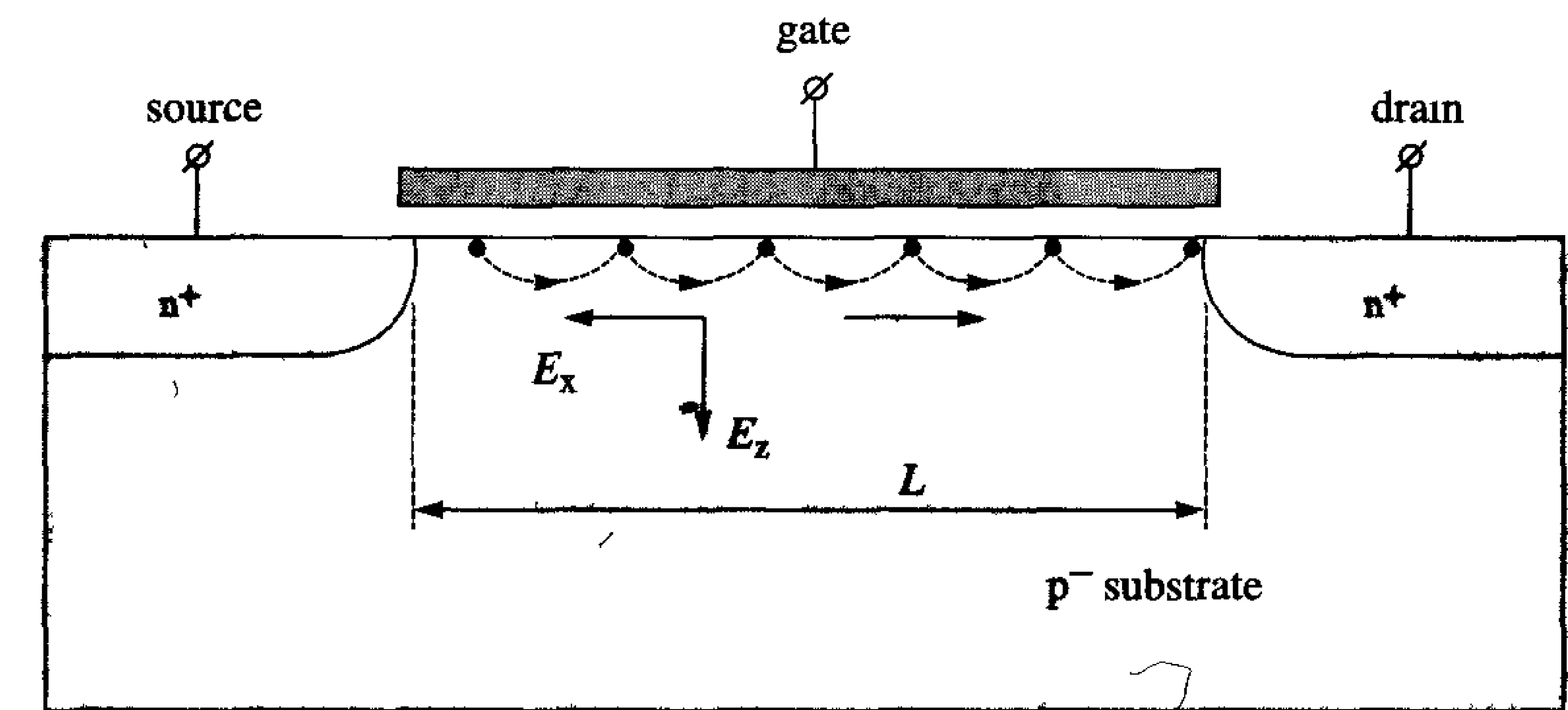


Figure 2.2: *Components which affect carrier mobility in MOS transistors*

When the *vertical electric field* E_z is high, the carriers in an n-channel device are strongly attracted to the silicon surface, where they rebound. The resulting *'surface scattering'* is indicated by the dashed lines in figure 2.2. This causes a reduction in the recombination time and the mobility μ of the carriers.

In [7], some experimental results are presented with respect to the vertical field carrier mobility degradation. The vertical electric field depends on the gate voltage and on the substrate voltage. The relationship between these voltages and the mobility can be expressed as follows:

$$\mu = \frac{\mu_0}{1 + \theta_1(V_\mathrm{gs} - V_\mathrm{T}) + \theta_2(\sqrt{V_\mathrm{sb} + 2\phi_\mathrm{F}} - \sqrt{2\phi_\mathrm{F}})} \tag{2.3}$$

where, μ_0 represents the substrate mobility, ϕ_F represents the Fermi level in the substrate and θ_1 and θ_2 are constants.

The actual mobility is equal to the substrate mobility when $E_\mathrm{z} = 0$. Some transistor models include the series resistance of the source and drain in the surface scattering factor θ_1.

The carriers in the transistor channel are accelerated to a maximum velocity when the *horizontal electric field* E_x is high. This means that, above a critical field $E_{\mathrm{x_c}}$, the carrier velocity is no longer related to E_x and reaches a constant level. A good first-order approximation for this 'velocity saturation' phenomenon is:

$$\mu = \frac{\mu_0}{1 + E_\mathrm{x}/E_{\mathrm{x_c}}} \quad \text{for } E_\mathrm{x} \leq E_{\mathrm{x_c}} \tag{2.4}$$

$$\text{where} \quad E_\mathrm{x} = \frac{V_\mathrm{ds}}{L} \tag{2.5}$$

$$\text{and} \qquad \frac{1}{L \cdot E_{x_c}} = \theta_3 \tag{2.6}$$

The term L represents the channel length. Substituting equations (2.5) and (2.6) in equation (2.4) yields:

$$\mu = \frac{\mu_0}{1 + \theta_3 \cdot V_{ds}} \tag{2.7}$$

The above effects are included in the following expression for carrier mobility:

$$\mu = \frac{\mu_0}{(1 + \theta_1(V_{gs} - V_T) + \theta_2(\sqrt{V_{sb} + 2\phi_F} - \sqrt{2\phi_F}))(1 + \theta_3 V_{ds})} \tag{2.8}$$

At high gate voltages, the vertical field influence (represented by the voltage terms containing V_{gs} and V_{sb}) may reduce the current from a transistor's drain to its source by about 50 %. The horizontal field influence may be of the same magnitude. Note that this horizontal field close to the source dominates the drain-source current. At a level of about 1 V/μm, this horizontal field also reduces the electron mobility in the channel of an nMOS transistor by almost 50 %. Thus, the total field-dependent mobility reduction can amount to a factor four.

As a result of the ultra short channel lengths, most carriers in the channel travel at a maximum *saturation velocity* v_{sat}. This almost eliminates the channel length's influence on the current, which can then be expressed as a linear relation with the V_{dd}.

$$I_{ds_{sat}} = v_{sat} \cdot C_{ox} \cdot W(V_{dd} - V_T) \tag{2.9}$$

For a normalised width W, $I_{ds_{sat}}$ is proportional to $(V_{dd} - V_T)/t_{ox}$. Therefore, the transistor's *drive current* is expected to show negligible increases over many technology generations to come [16].

2.4 Channel length modulation and static drain feedback

The ideal I_{ds} vs V_{ds} characteristics illustrated in figure 1.16 do not show the influence of V_{ds} on I_{ds} in the saturation region. In practice, an increase in V_{ds} in the saturation region causes an increase in I_{ds}. This phenomenon is particularly obvious in short-channel devices. It is caused by two effects, namely: *channel length modulation* and *static drain feedback*. These effects are discussed in this section.

2.4.1 Channel length modulation

The distribution of carriers in an nMOS transistor operating in the saturation region ($V_{ds} > V_{dssat} = V_{gs} - V_T$) is illustrated in figure 2.3. The operation of the basic MOS transistor in this region is discussed in section 1.3. Clearly, the end of the inversion layer (which is called the imaginary drain) does not reach the actual drain. The effective channel length therefore equals $L - \Delta L$.

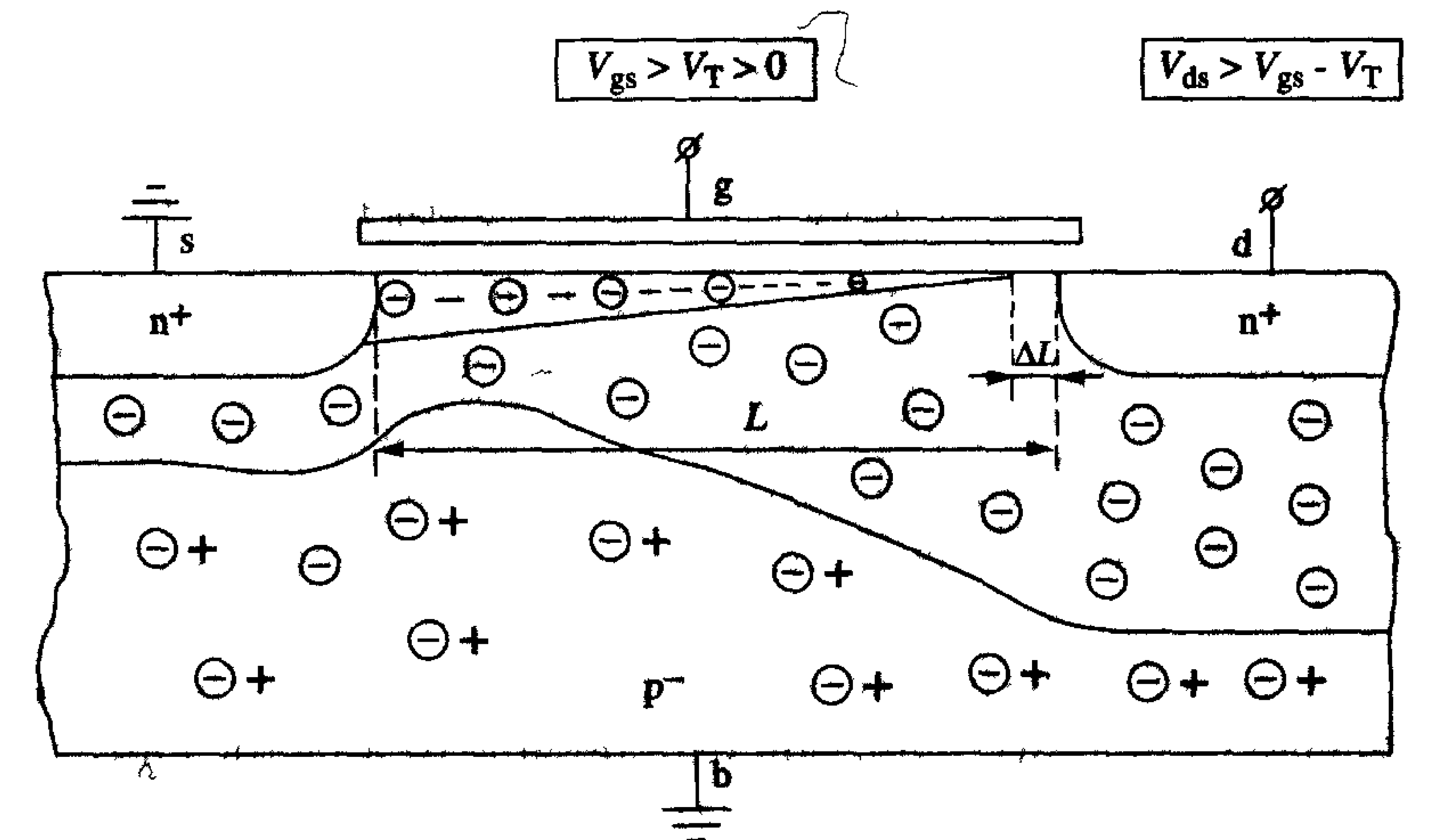


Figure 2.3: *A MOS transistor in the saturation region* ($V_{ds} > V_{gs} - V_T$)

The saturation current specified in equation (1.14) must be changed to account for the effective channel length. The modified expression is as shown in equation (2.10). In this expression, the total field-dependent mobility degradation, as discussed before, is not included.

$$I_{ds_{sat}} = \frac{W}{L - \Delta L} \cdot \frac{\beta_\square}{2} \cdot (V_{gs} - V_T)^2 \tag{2.10}$$

where ΔL is the length of the depletion region at the silicon surface between the inversion layer and the drain.

The voltage $V_{ds} - V_{dssat}$ across this *'pinch-off' region* modulates ΔL. This effect can be modelled by:

$$\frac{\Delta L}{L} = \alpha \ln(1 + \frac{V_{ds} - V_{ds_{sat}}}{\alpha V_P}) \tag{2.11}$$

where α and V_P are constants, which may vary with the transistor geometry.

The above discussions show that the additional contribution to the drain current of a MOS transistor operating in the saturation region is proportional to $V_{ds} - V_{ds_{sat}}$

2.4.2 Static drain feedback

The depletion area below the gate is influenced by the channel potential, which varies from source to drain, as illustrated in figure 2.3. The depletion charge is therefore largely determined by V_{ds}. The depth of the drain depletion region is directly proportional to the drain potential. A deeper depletion region contains more minority carriers. These carriers reduce the barrier that a gate potential must surmount when creating an inversion layer. An increase in drain potential therefore results in a reduction of the threshold voltage. The following empirical expression, in which ΔV_T is proportional to V_{ds}, appears reasonably satisfactory:

$$\Delta V_T = -\gamma V_{ds}^{0.6} \tag{2.12}$$

The factor γ is constant and expresses the relationship between drain-source voltage and threshold voltage variation. The effect of this static drain feedback is considerable in short-channel transistors that operate close to the threshold voltage.

Channel length modulation and static drain feedback cause the drain current to vary with the transistor's output impedance. Consider, for example, a circuit comprising two identical transistors operating in the saturation region with equal gate voltages but different drain voltages. The transistor with the higher drain voltage will have a shorter channel and a reduced threshold voltage. Consequently, its drain to source current will be higher.

2.4.3 The Early voltage

Another means of describing the previous effects is to use the *Early voltage*. This voltage is commonly-used in bipolar transistor theory and, by analogy, the MOS Early voltage (V_E), which can be defined as:

$$V_E = \frac{I_{ds}}{\delta I_{ds}/\delta V_{ds}} = \frac{I_{ds}}{g_{ds}} \tag{2.13}$$

where g_{ds} represents the output transconductance.

Figure 2.4 shows the effect of the Early voltage on the MOS transistor characteristics.

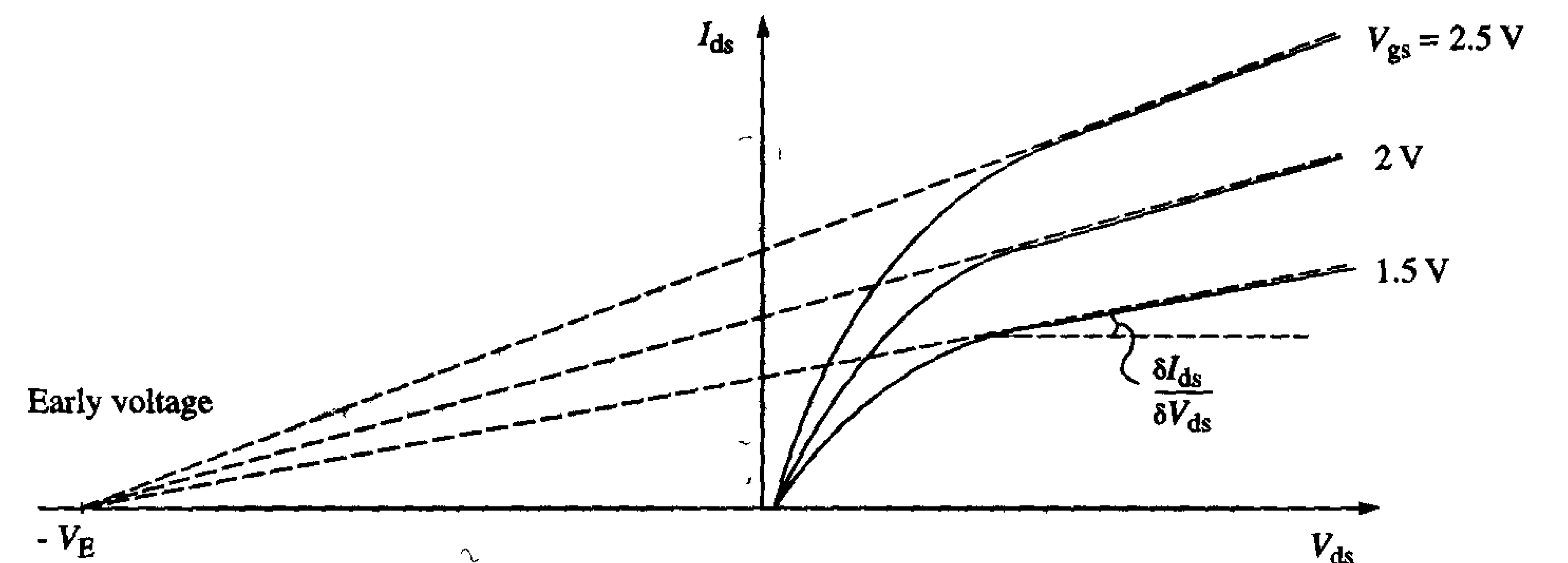


Figure 2.4: *The Early voltage effect on the MOS transistor characteristic*

Figure 1.16 suggests that the output resistance of a MOS transistor operating in the saturation region is infinite. Figure 2.4, however, reveals that this output resistance is in fact finite and is largely determined by the value of the Early voltage. Empirically, the Early voltage has been found to have the following properties:

- It increases with higher substrate doping levels;
- It decreases with decreasing channel length;
- It is characteristic for each process;
- It typically lies between 5V and 50V.

2.5 Small-channel effects

The electrical behaviour of a MOS transistor is primarily determined by its gain factor β, its threshold voltage V_T and its body factor K. Generally, the values of these parameters are largely dependent on the width W and length L of a transistor. The influence of these dependencies increases as transistor dimensions decrease. These small-channel effects, which are discussed below, are particularly significant in submicron and deep-submicron MOS processes.

2.5.1 Short-channel effect

The cross-section of a short-channel transistor presented in figure 2.5 is used to explain short-channel effects.

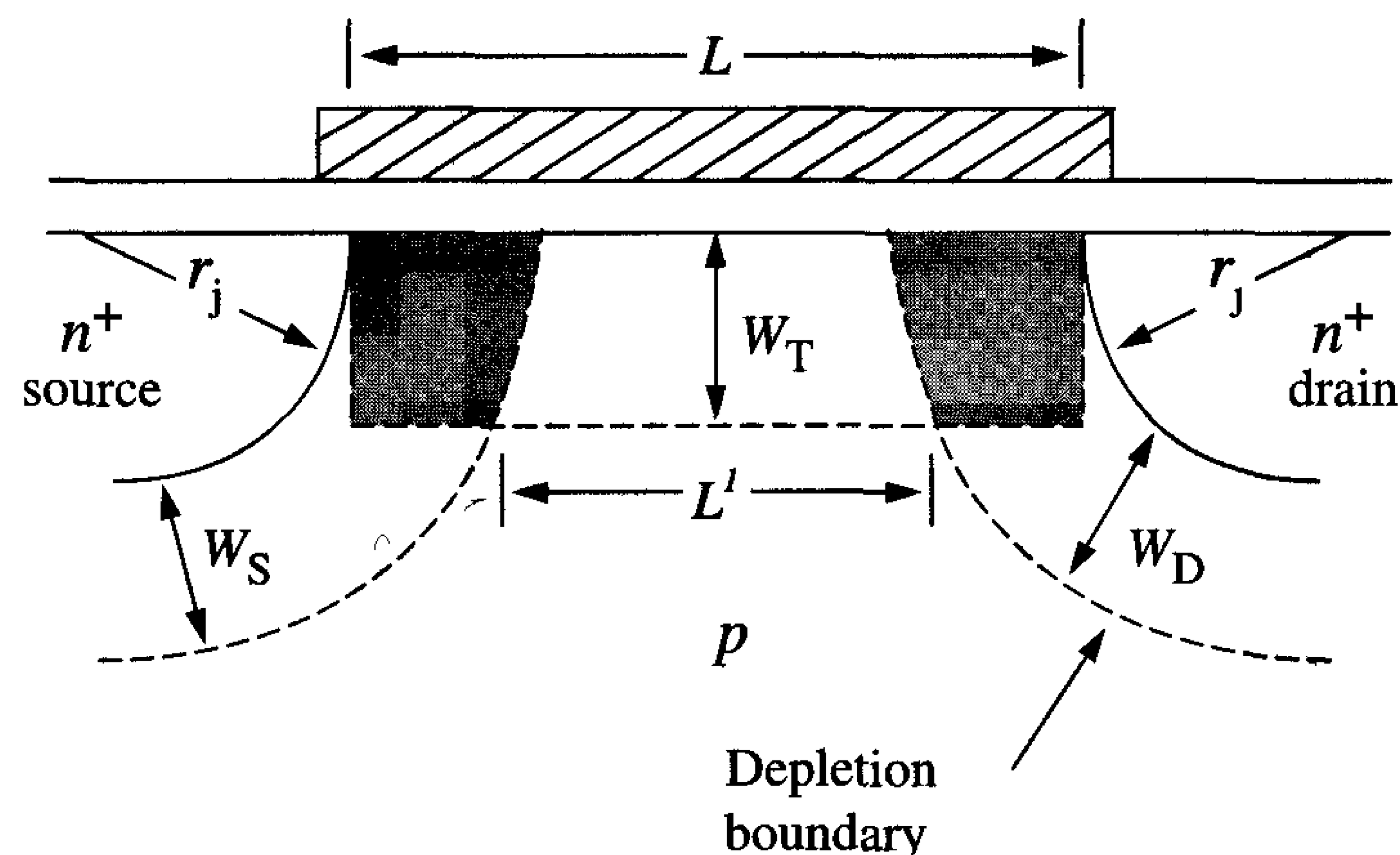


Figure 2.5: *Cross-section of a short-channel transistor, showing several depletion areas that affect each other*

Even in the absence of a gate voltage, the regions under the gate close to the source and drain are inherently depleted of majority carriers, i.e. holes and electrons in nMOS and pMOS transistors, respectively. In a short-channel transistor, the distance between these depletion regions is small. The creation of a complete depletion area under the gate therefore requires a relatively small gate voltage. In other words, the threshold voltage is reduced. This is a typical two-dimensional effect, which can be reduced by shallow source and drain diffusions. However, the associated smaller diffusion edge radii cause a higher electric field near the

drain edge in the channel when $V_{ds}>V_{gs}>V_T$. One way to overcome this problem is to reduce the supply voltage. This short-channel effect on the threshold voltage occurs at shorter gate lengths and causes *threshold voltage roll-off*, see figure 2.6. The figure also shows the effect of the ΔV_T implantation, which is discussed below.

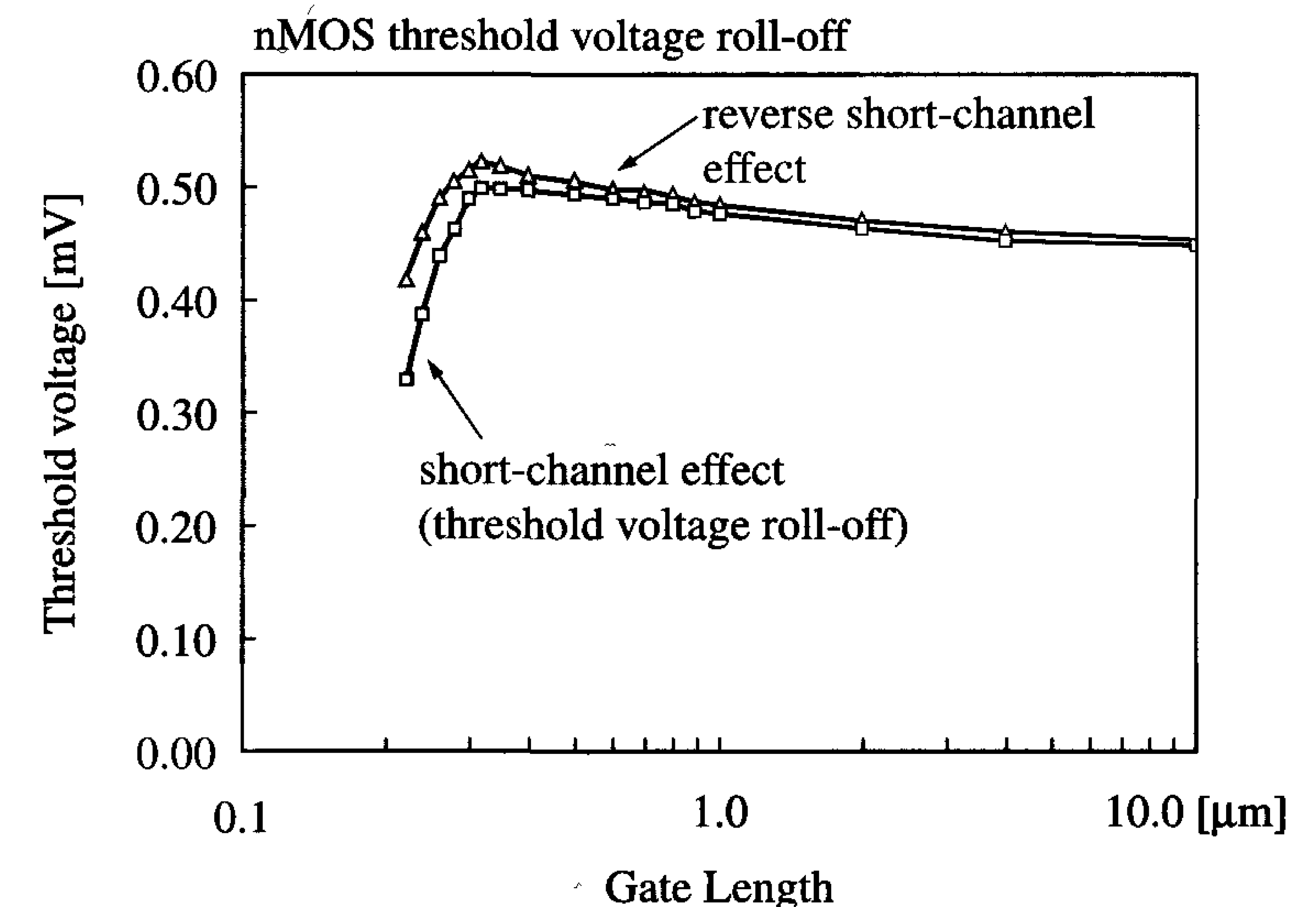


Figure 2.6: *Short-channel and reverse short-channel effect on the threshold voltage V_T of an nMOS transistor with and without a ΔV_T implantation*

In addition to the previously-mentioned undesirable increased electric field near the drain, the use of shallow source and drain diffusions (to reduce the short-channel effect on the threshold voltage) also increases the resistance of these areas. This in turn increases the need to use silicides (see section 3.7.2).

An alternative approach to using shallow source and drain diffusions uses a ΔV_T *implant* to compensate the V_T drop in short-channel devices. The implant is optimised for transistors with the smallest channel lengths in a given process. These transistors will have the nominal threshold voltage while transistors with longer channels will have higher threshold voltages.

A second effect that depends on the channel length is the *reverse short-channel effect*. This effect, increasing threshold voltage V_T with decreasing gate length, is attributed to a lateral non-uniform channel

doping induced by locally enhanced diffusion. During processing, an excess of silicon interstitial silicon atoms is created in the neighbourhood of the polysilicon gate edges as a result of implantation and oxidation steps. These interstitials enhance diffusion in a certain region under the gate.

The lateral extent of this enhanced diffusion is strongly influenced by the SiO_2/Si interface that forms an effective sink for interstitials and creates large concentration gradients. The flux of interstitials results in an enhanced diffusion of dopant atoms such as boron at the gate edges towards the interface. The redistribution of the boron results in a modification of the channel profile of the nMOS devices and a substantial increase of the threshold voltage is observed. This effect is more pronounced at shorter gate lengths because the channel profile is only modified at the edges. At even shorter gate lengths, the normal short-channel effect starts to dominate and the threshold voltage roll-off is observed, see figure 2.6. The effect is more pronounced for nMOS devices because the channel dopant atom used is boron, which is more susceptible to these enhanced diffusion effects than arsenic, which is used for the pMOS channel profile.

2.5.2 Narrow-channel effect

Also, the width of an active device influences the threshold voltage. The depletion layer extends under the edges of the gate, where the gate electrode crosses the field oxide. With a LOCOS type of field isolation, see figure 2.7, this effect is primarily caused by the encroachment of the channel stop dopant at the edge of the field isolation.

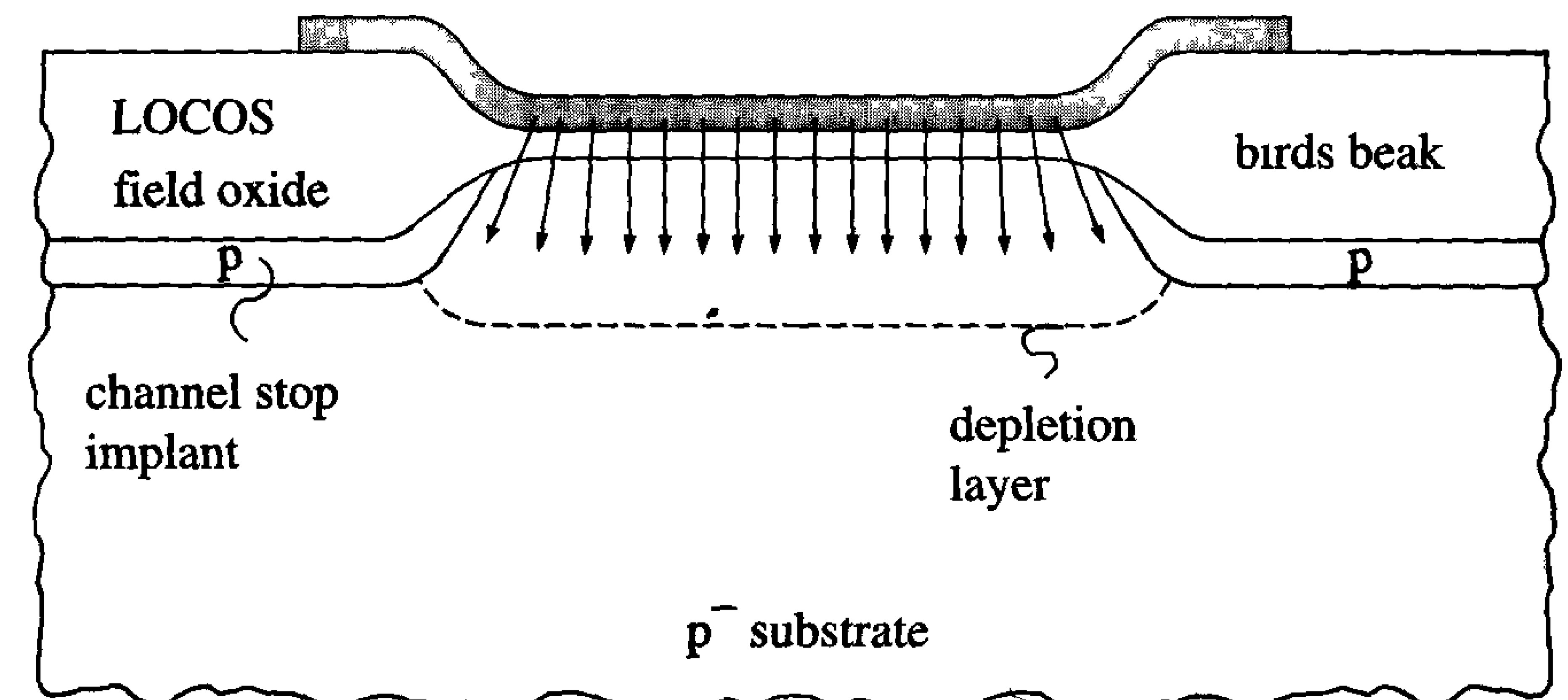


Figure 2.7: *Cross-section of a narrow-channel transistor showing the distribution of electric field lines under the gate*

The additional depletion region charge has to be compensated by an additional gate voltage. This results in an increase of the threshold voltage at reduced width of the device. The encroachment of channel stop dopant is especially pronounced for a conventional diffused well technology. The channel stop dopants are implanted prior to the high-temperature LOCOS oxidation and cause a large shift in V_T. In a retrograde implanted well process, the field oxidation is performed prior to the well implants and less encroachment of dopant atoms occurs under the gate edge. However, the threshold voltage is increased as a result of the bird's beak and two-dimensional spreading of the field lines at the edge. Figure 2.8 shows this *narrow-channel effect*, together with the influence of the channel width on the threshold voltage in a Shallow-Trench Isolation (section 3.8) scheme. In contrast to the conventional narrow-width effect, the threshold voltage shift is limited and even decreased at very narrow channel widths of around $0.2\,\mu m$.

This *Inverse Narrow-Width Effect* (INWE) is attributed to a sharp corner at the top of the shallow-trench isolation. The fringing field at this corner results in an increased electrical field strength and reduces the threshold voltage. Also, the quality of the oxide used to fill the trench is not as good as the thermal oxide of the field oxide. A positive fixed oxide charge is present in the oxide and, in nMOS devices, it contributes to the decreased threshold voltage. This contribution of the fixed oxide charge is less severe than the fringing field compound and depends also on the deposition method used to fill the trench.

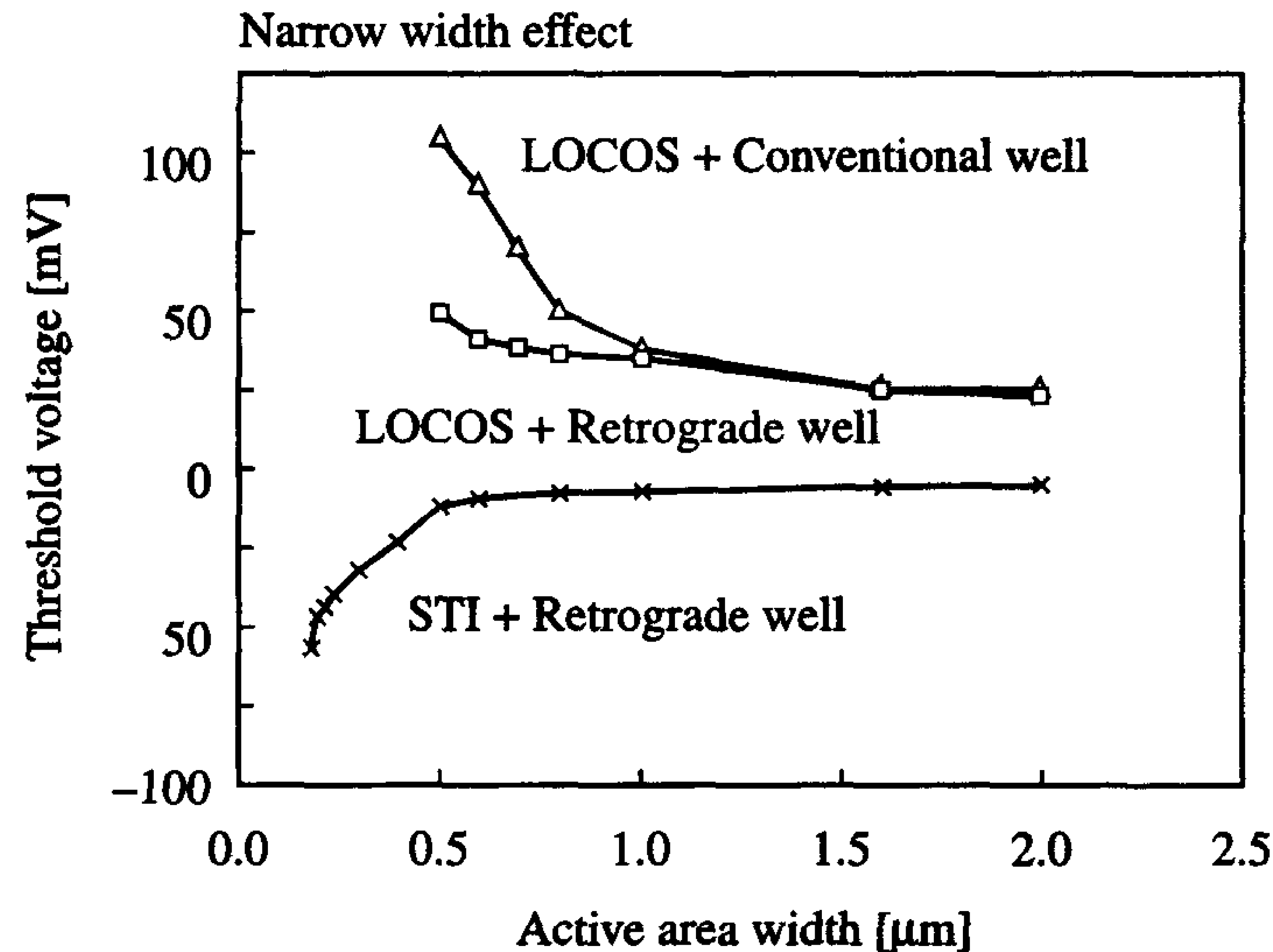


Figure 2.8: *Shift of threshold voltage of nMOS devices as function of the active area width for different well technology and field isolation schemes*

2.5.3 Modelling small-channel effects

The previous sections show that short-channel effects tend to lower the threshold voltage while narrow-channel effects increase it. Therefore, these effects compensate each other partly, if not completely, in a transistor with a both short and narrow channel. For modelling purposes, both effects can simply be added. The influence of a transistor's dimensions on its body-effect factor K are modelled in a similar manner. These relationships are summarised as follows:

- A shorter channel results in a lower V_T, a lower μ and a lower K-factor.
- A narrower channel results in a higher V_T, a lower μ and a higher K-factor.

The dependence of a transistor's V_T, for example, on its geometry can be modelled as follows:

$$V_T = V_{T0} + D_{l_1}\left(\frac{1}{L_{ref}} - \frac{1}{L}\right) + D_{l_2}\left(\frac{1}{L_{ref}^2} - \frac{1}{L^2}\right) + D_w\left(\frac{1}{W_{ref}} - \frac{1}{W}\right)$$

where D_{l_1} and D_{l_2} represent the short-channel and the reverse short-channel effect, with $D_{l_1} > 0$ and $D_{l_2} < 0$.
D_w represents the channel width dependence (narrow width effect), with $D_w > 0$.
W_{ref} and L_{ref} are width and length of the reference transistor respectively.

Often, the transistor with minimum channel length and width is chosen as the reference transistor. Consequently, the capabilities of a process technology can be easily evaluated from the parameters of the reference transistor.

2.6 Punch-through

The drain and source depletion regions of a MOS transistor may merge when a sufficiently large reverse-bias voltage is applied to the drain-substrate junction. This is particularly likely to occur in MOS transistors with short channel lengths. The energy barrier, which keeps electrons in the source of an n-channel device, is lowered when the drain and source depletion regions merge. Consequently, many electrons start to flow from the source to the drain even when the gate voltage is below the threshold value and the transistor is supposedly not conducting. This effect is known as *punch-through*. The drain-source voltage V_{PT} at which punch-through occurs is approximated as follows:

$$V_{PT} = \frac{q}{2\epsilon_0\epsilon_r} \cdot N_A \cdot L^2 \qquad (2.14)$$

where N represents the substrate dope,
L represents the transistor channel length and
q represents the charge of an electron.
The punch-through effect can be reduced during processing by increasing the doping level of the substrate with a *anti-punch-through* (APT) implantation. The associated increase in the threshold voltage of the transistor can be compensated by reducing the oxide thickness.

2.7 Hot-carrier effect

2.7.1 Introduction

Impact ionisation and hot-carrier degradation occur in a MOS transistor when a significant number of the mobile charge carriers in the channel become sufficiently energetic. This happens when the horizontal electric field in the transistor becomes larger than 2×10^5 V/cm. Impact ionisation increases drain current and produces a substrate current. Hot-carrier degradation causes a permanent change in a transistor's $I - V$ characteristic. For the same operating voltage, the above effects are much more significant in n-channel transistors than in p-channel transistors.

Therefore, this section is limited to a discussion of these effects in n-channel transistors. Hot-carrier effects are very strongly related to the maximum value of the horizontal electric field in a MOS transistor. The distribution of this electric field and its dependence on external and internal parameters is therefore described. The effects of impact ionisation and several aspects of hot-carrier degradation are then discussed. Finally, two methods are presented to reduce the electric field in small MOS transistors.

2.7.2 The electric field in MOS transistors

Figure 2.9a shows a typical example of the potential distribution in an n-channel MOS transistor operating in the saturation region. Figure 2.9b shows the cross-section of the relevant transistor. Figure 2.9c shows the horizontal electric field $E_x = dV/dx$. The bulk of the voltage drop occurs in the transistor pinch-off region near the drain junction. This causes the sharp peak in the horizontal electric field. Two-dimensional effects produce a maximum electric-field value E_{mx} which is slightly within the drain region instead of at the channel-drain metallurgical junction.

The value of E_{mx} depends on two groups of parameters. The first group contains the bias voltages on the transistor terminals. Here, the strongest dependence is on V_{ds}. Value E_{mx} is in fact directly proportional to V_{ds} if the ratio V_{gs}/V_{ds} is kept constant. An increase in V_{gs} causes a more even distribution of the electric field over the length of the transistor and E_{mx} therefore decreases. The dependence on the substrate voltage is very small. The drain-substrate voltage only determines the maximum electric field at the drain-substrate junction. Extremely

large drain-substrate voltages cause a breakdown in this junction and have a negative effect on the operation of the transistor.

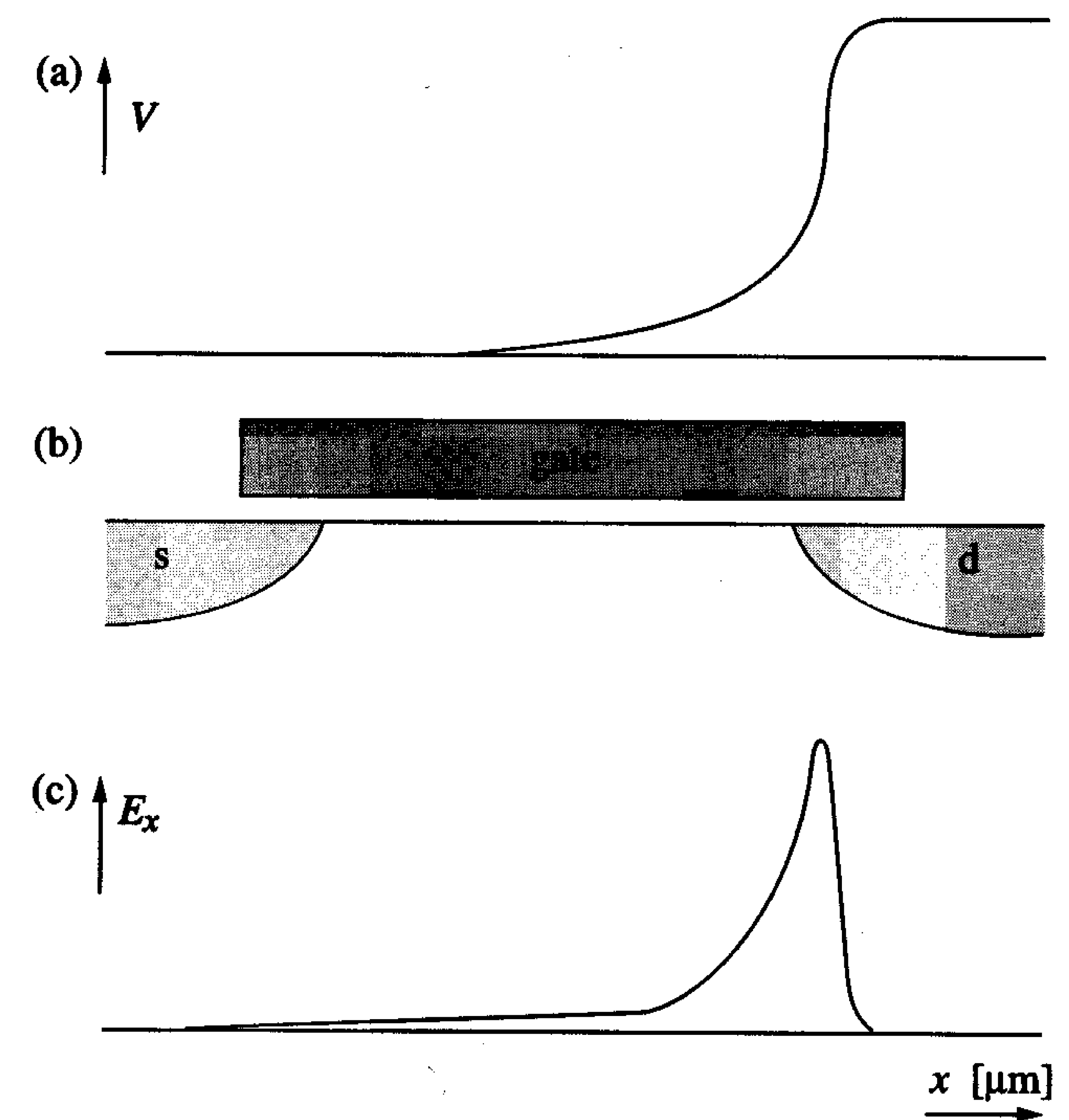


Figure 2.9: *(a) A typical example of the potential distribution in a saturated n-channel MOS transistor. (b) A cross-section of the MOS transistor. (c) The horizontal electric field $E_x = dV/dx$ as a function of x.*

The second group of parameters that determine the value of E_{mx} contains the geometrical parameters of the transistor. E_{mx} increases when the transistor length decreases and when the gate oxide becomes thinner. The doping profile of the drain area, however, is of greater influence. The depletion layer does not extend very far into the drain area when the junction between the highly-doped drain and the substrate is very abrupt. The distribution of the electric field E_x is then very asymmetrical and the value of E_{mx} is high. Field E_x is much more symmetrical and the value of E_{mx} is much lower when the drain-substrate junction is less abrupt. This is also the case when operating conditions produce a constant donor concentration in the channel which is almost as high as the drain concentration.

Methods to reduce E_{mx} are discussed in section 2.7.5.

2.7.3 Impact ionisation

A conduction electron may collide with a silicon atom at a lattice point in an nMOS transistor channel. If such a collision involves a high-energy electron, it can cause the transition of an electron from the valence band to the conduction band. This produces an extra conduction electron and a hole. Both electrons flow to the drain and the hole drifts to the substrate. This effect is called *impact ionisation.* This gives rise to an increase in the drain-source current I_{ds} and to a substrate current I_b. Figure 2.10 shows the dependence of I_b on the gate-source voltage V_{gs} for different values of V_{ds}, where I_b is plotted on a linear scale. The degree in which impact ionisation occurs and the magnitude of I_b are directly proportional to the square of E_{mx}. A reduction in E_{mx} therefore causes a reduction in I_b.

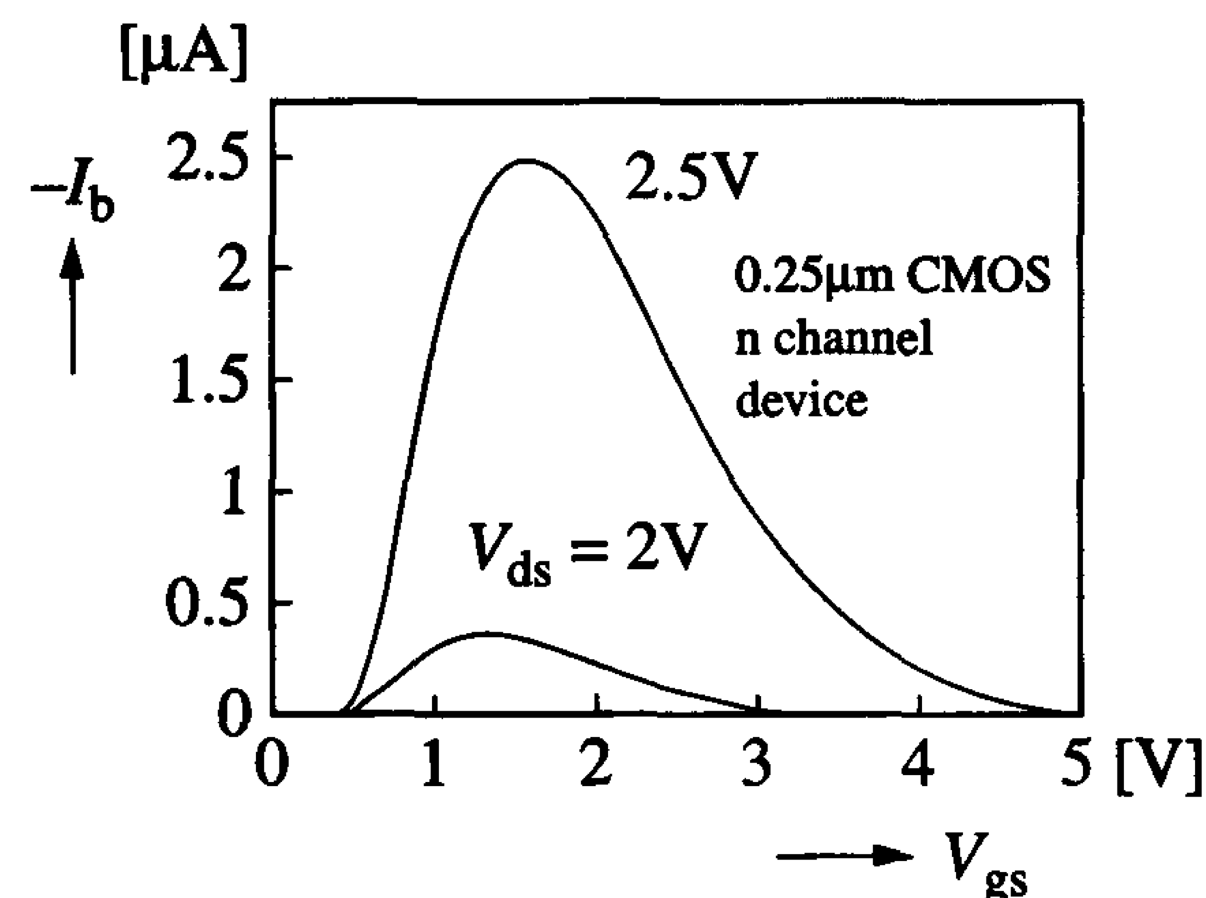


Figure 2.10: *Substrate current I_b as a function of V_{gs} for two values of V_{ds}*

2.7.4 Hot-carrier degradation

Impact ionisation generates high-energy electrons and holes. These carriers can scatter towards the Si-SiO_2 interface and enter the SiO_2 if they have sufficient energy. The required energy is at least 3.2 eV for electrons and 3.8 eV for holes. The number of carriers that satisfy these

conditions is very small. The carriers may become trapped in the $Si0_2$, giving rise to oxide charge. In addition to this charge-trapping effect, interface states also occur. The sign and magnitude of the resulting extra space charge are dependent on the bias conditions of the transistor. The degradation of transistor current-voltage characteristics which is caused by this injection of both types of carriers into $Si0_2$ is called *hot-carrier degradation.* This effect was commonly called hot electron degradation before it was realised that it is also influenced by holes.

2.7.5 Reducing the maximum electric field in a MOS transistor

Graded drain and lightly doped drain structures are used to reduce the maximum value of the electric field in small transistors and thus prevent hot-carrier degradation.

The *graded drain* transistor is a very simple adaptation of the conventional transistor. The junction between the drain and the substrate is made much more gradual by simply implanting phosphorous with a relatively low concentration in the highly concentrated n^+ area. The phosphorous has a much higher diffusion coefficient than the arsenic in this area and therefore diffuses much further. This results in a donor profile with a low gradient; an example is shown in figure 2.11. The graded drain reduces the maximum electric field by about 30 %. This implies that the operating voltage can be increased by 50 % for given transistor dimensions.

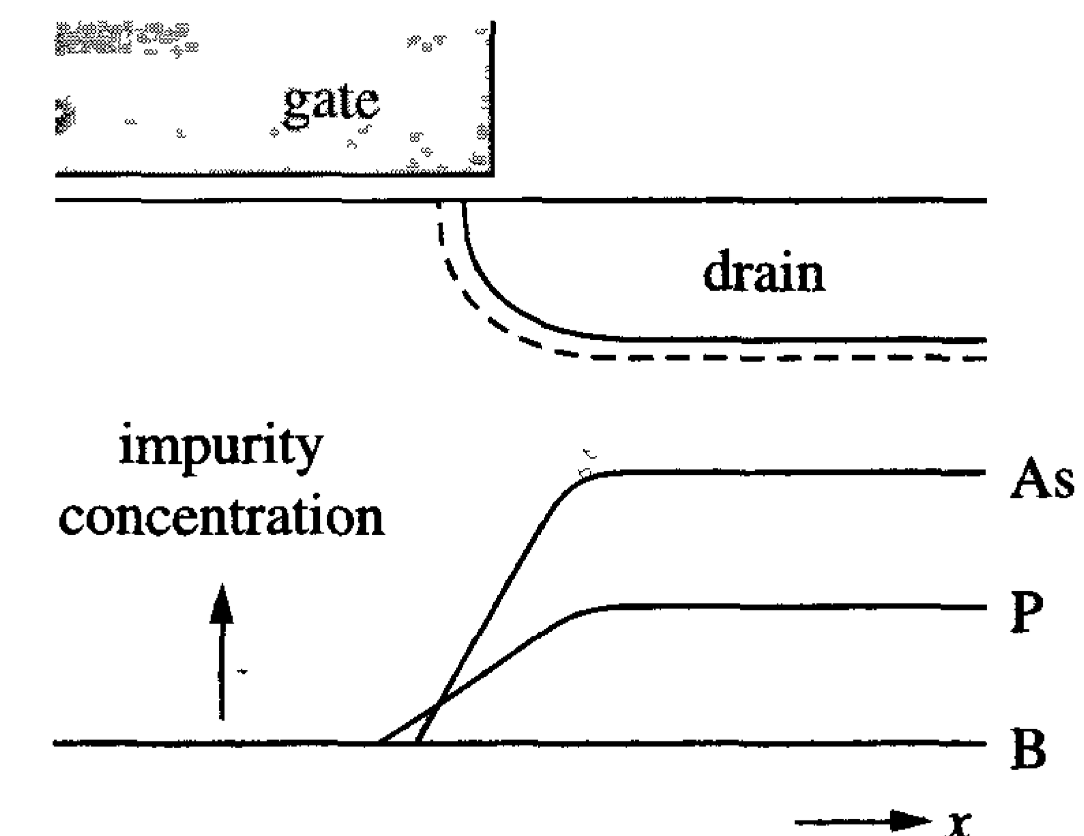


Figure 2.11: *Phosphorous (P) halo around arsenic (As) in the cross-section of a graded drain transistor and the concentrations as a function of the position x*

The lightly doped drain (LDD) method is a more difficult means of reducing the drain-substrate concentration gradient. The maximum electric field obtained is lower than that achieved with the graded drain. The various LDD process steps are explained with the aid of figure 2.12. A 0.35 μm nMOS transistor with a gate oxide thickness of about 7 nm thickness is shown in figure 2.12a.

Normal processing, which is extensively described in chapter 3, is used to create the gate oxide. Phosphorous with a concentration that varies from 1×10^{18} to 4×10^{18} atoms per cm^3 is subsequently implanted. The oxide layer of about 0.35 μm thickness shown in figure 2.12b is then grown. This is followed by an anisotropic etch, which leaves the oxide spacers shown on both sides of the gate in figure 2.12c. A subsequent highly concentrated implantation of arsenic and a drive-in diffusion produce the resulting n^- and n^+ areas shown in figure 2.12d. The magnitude of the transistor's horizontal electric field as a function of the channel position x is shown in figure 2.12e. Its maximum value is 50 % of that obtained in a comparable transistor with conventional arsenic drain and source areas. Two factors account for this significant reduction. The first is the relatively long region with a low donor concentration. A depletion area will form much sooner in this area than in the n^+ area. A large proportion of the drain-source voltage drop is distributed over this area. The second factor is the extra separation between the gate and the n^+ drain area. This also reduces the influence of the second-order effects previously discussed in this chapter.

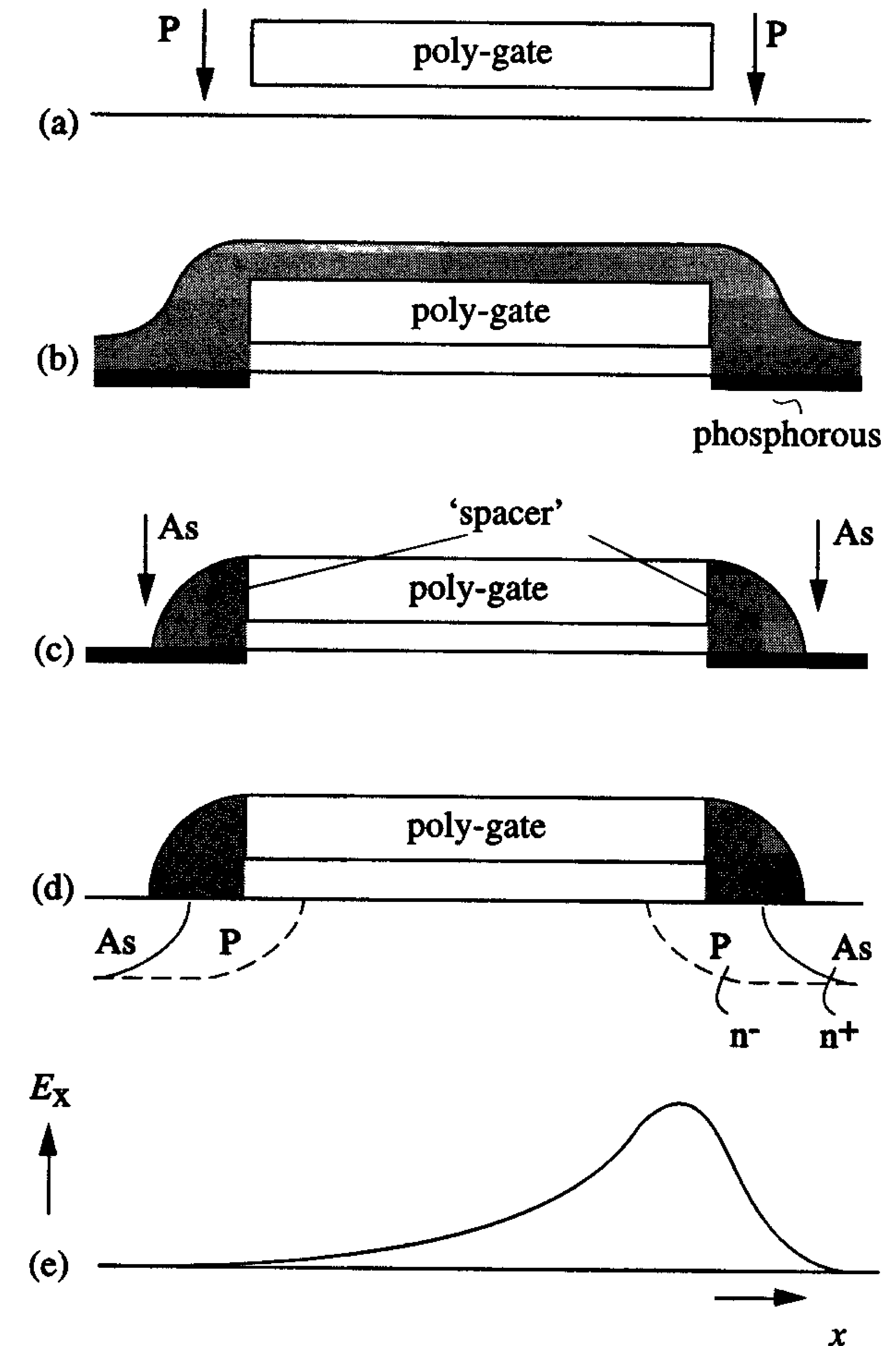


Figure 2.12: *Process steps for the creation of an LDD transistor and the resulting reduced horizontal electric field distribution*

The LDD transistor is difficult to create and has the added disadvantage of possible increased series resistance in the source and drain as a result of the n^- areas. Careful optimisation, however, yields small transistors with high operational voltages that can deliver high currents. LDD implants are included in CMOS technologies down to 0.35 μm channel lengths. As discussed in section 2.7.4, the required energy for carriers to cross the Si-Si0_2 interface barrier is at least 3.2 eV for electrons and 3.8 eV for holes. As supply voltages reduce with the advent of new process generations, these carriers can hardly ever reach such energies when the supply voltage is 2.5 V or less. In a 0.25 μm CMOS process,

with $V_{dd} = 2.5\,\mathrm{V}$, the electron energy can reach $qv = 2.5\,\mathrm{eV}$, except for those electrons that achieve increased energy through one or more collisions. However, the number of such high-energy electrons is rapidly decreasing with reducing voltages below 3 V. As a result, LDD implants are no longer required in $0.25\,\mu\mathrm{m}$ CMOS processes and below. These are then replaced by a more highly doped drain extension, as discussed in section 3.8.3.

2.8 Weak-inversion behaviour of the MOS transistor

An nMOS transistor operates in the *'weak-inversion'* region when its gate-source voltage (V_{gs}) is just below its threshold voltage (V_T), see figure 2.13.

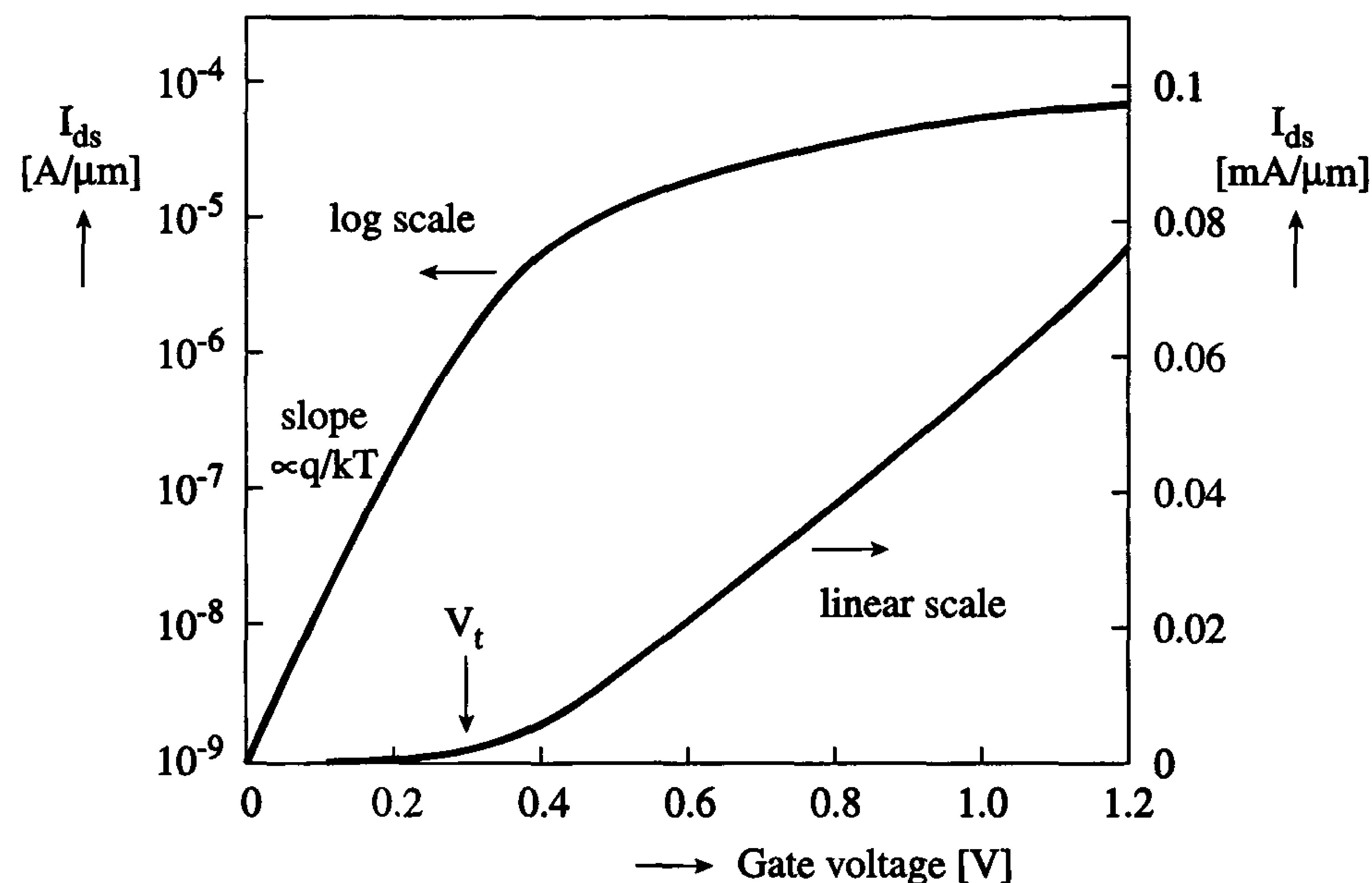


Figure 2.13: *(a) Sub-threshold current representation and (b) on enlarged scales*

Below the threshold voltage, the current decreases exponentially. On a logarithmic scale, the slope is inversely proportional to the thermal energy kT. Some electrons still have enough thermal energy to cross the gate-controlled potential barrier (figure 2.14) and travel to the drain [13]. At (very) low threshold voltages, the resulting *subthreshold current* may lead to an unacceptibly high power consumption. This leakage current should not exceed a few tens of nano-amperes for a one hundred million transistor chip in standby mode (no circuit activity and at zero gate voltage). This operating region is also called the *'sub-threshold region'*. The normal characteristics that apply to the previously-considered full-inversion triode or linear operating region do not apply to the weak-inversion region. The drain-source current in a transistor with a long channel and a constant drain-source voltage operating in the weak-inversion region is expressed as follows:

$$I_{\mathrm{dssub}} = \frac{W}{L} \cdot C\, I_{\mathrm{ds0}}\, \mathrm{e}^{V_{\mathrm{gb}}/mU_{\mathrm{T}}} \tag{2.15}$$

The terms in equation (2.15) are defined as follows:

$$\begin{aligned} C &= \mathrm{e}^{-V_{\mathrm{sb}}/U_{\mathrm{T}}} - \mathrm{e}^{-V_{\mathrm{db}}/U_{\mathrm{T}}} \quad (C \text{ is constant here}); \\ U_{\mathrm{T}} &= \frac{kT}{q} \approx 25\,\mathrm{mV} \text{ at room temperature;} \\ I_{\mathrm{ds0}} &= \text{characteristic current at } V_{\mathrm{gb}} = 0\,\mathrm{V}; \\ m &= \text{slope } \approx 1.5. \end{aligned}$$

Equation (2.15) applies when V_{gs} is no more than a few U_T below V_T. The sub-threshold transistor current I_{dssub} can lead to a considerable *stand-by current* in transistors that are supposedly inactive.

An accurate description of the behaviour of a transistor operating in the weak-inversion region is contained in references [11] and [12]. The following statements briefly summarise this operating region:

1. If V_T is low ($\leq 0.6\,\mathrm{V}$), there is always a sub-threshold current when $V_{gs} = 0\,\mathrm{V}$. This has the following consequences:
   (a) There is a considerable stand-by current in (C)MOS VLSI and memory circuits;
   (b) The minimum clock frequency of dynamic circuits is increased as a result of leakage currents. DRAMs are among the circuits affected.
2. In normal transistor operation, carriers face a barrier close to the source, which they have to cross. Figure 2.14 shows the influence

of the drain voltage on the barrier height. An increase of the drain-source voltage will reduce the barrier height. This *Drain-Induced Barrier Lowering effect* (DIBL) leads to a reduction of the threshold voltage V_T in both the weak-inversion and strong-inversion operation regions of the transistor.

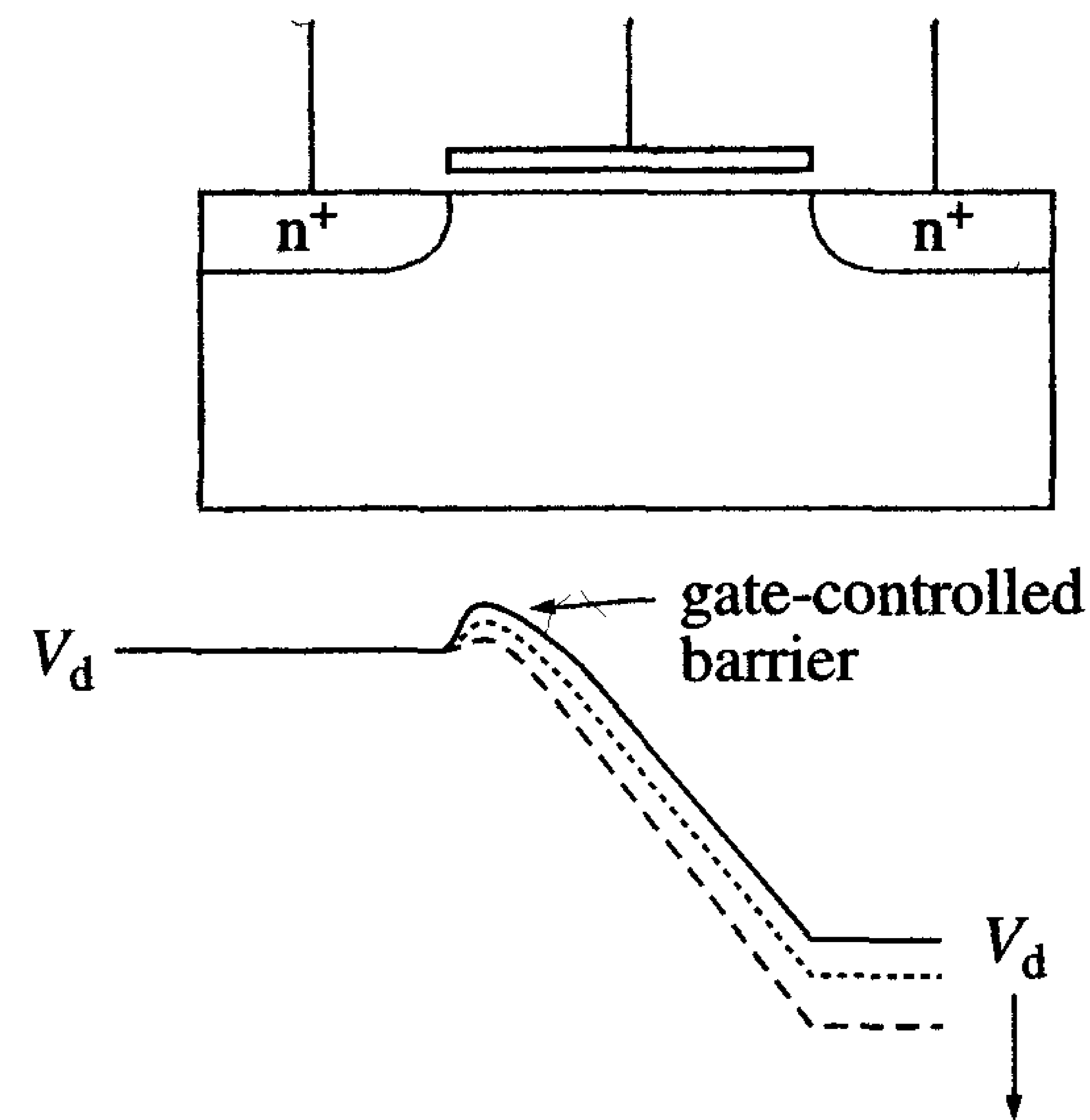


Figure 2.14: *The effect of the drain voltage on lowering the barrier*

In a 0.18 μm CMOS process, an increase of V_{ds} from 0.6 V to 1.8 V may result in a threshold reduction of ΔV_T of more than 100 mV. The effect manifests itself more at lower effective channel lengths (DIBL $\div \frac{V_{ds}}{L_{eff}^2}$). It is clear that this effect is very important in the development and the modelling of deep-submicron technologies.

3. Analogue circuit techniques use weak-inversion behaviour in low-current applications. The gain of a MOS transistor operating in the weak-inversion region is relatively high and comparable to the gain of bipolar transistors.

2.9 Conclusions

The formulae derived in chapter 1 provide a good insight into the fundamental behaviour of MOS devices. These formulae were used to predict circuit behaviour with reasonable accuracy until the mid-seventies. The continuous drive for higher circuit densities with smaller transistors, however, gave rise to an increased contribution from physical and geometrical effects. These effects cause deviations from the ideal transistor behaviour assumed in chapter 1. In addition, the magnitude of these deviations increases as transistor dimensions shrink. These effects combine to reduce the ideal transistor current by more than a factor four for channel lengths of 0.25 μm and below. There are also effects that permanently degrade the performance of a MOS transistor and therefore reduce its lifetime. Reducing the influence of these effects requires both technological and design measures.

2.10 References

General basic physics

[1] R.S.C. Cobbold,
'Theory and applications of field effect transistors',
John Wiley & Sons, Inc. New York

[2] S.M. Sze,
'Modern Semiconductor Device Physics',
John Wiley & Sons, 1997

[3] A.S. Grove,
'Physics and Technology of Semiconductor Devices',
John Wiley & Sons, Inc. New York

[4] Y.P. Tsividis,
'Operation and modeling of the Mos transistor',
Mc Graw-Hill, 1987

Mobility reduction of charge carriers

[5] S.C. Sun, J.D. Plummer,
'Electron mobility in inversion and accumulation layers on thermally oxidized silicon surfaces',
IEEE Transactions on Electron Devices, Vol 8, 1980

[6] S.M. Sze (ref [2]), chapter 1

[7] A.J. Walker and P.H. Woerlee,
'A mobility model for MOSFET device simulations',
Journal de Physique, colloque C4, Vol. 49, number 9, Sept. 1988, p 256

Short-channel effects

[8] G.Merkel,
'A simple model of the threshold voltage for short and narrow channel effects',
Solid-State Electronics, Vol. 23, number 12-B, 1980

[9] L.M. Dang,
'A simple current model for short-channel IGFET and its application to circuit simulation',
IEEE Journal of Solid-State Circuits, Vol. SC-14, number 2, 1979

Hot-carrier effects

[10] K. Chen, S.A. Saller, I.A. Groves, D.B. Scott,
'Reliability effects on MOS transistors due to the hot-carrier injection',
IEEE Transactions on Electron Devices, Vol. ED-32, number 2, 1985

[10b] Y. Leblebici and S.M. Kang,
'Hot-Carrier Reliability of MOS VLSI circuits',
Kluwer Academic Publishers, 1993

Weak-inversion behaviour

[11] R.R. Troutman and S.N. Chakaravatri,
'Characteristics of insulated-gate field-effect transistors'
IEEE Transactions on Circuit Theory, vol. CT-20 pp 659-665, Nov. 1973

[12] E. Vittoz and J. Fellrath,
'CMOS analog integrated circuits based on weak-inversion operation',
IEEE Journal of Solid State Circuits, Vol. SC-12, pp 224-231, June 1977

[13] Yuan Tauer,
'The incredible shrinking transistor'
IEEE Spectrum, July 1999, pp 25-29

Transistor modelling

[14] H.C. de Graaf and F.M. Klaassen,
'Compact transistor modeling for circuit design'
Springer-Verlag, New York, 1990

[15] T.A. Fjeldly, et al.,
'Device Modeling Issues in Deep-Submicron MOSFETs',
Semiconductor International, pp 131-142, June 1996

[16] G. Massobrio and P. Antognetti,
'Semiconductor Device Modeling with Spice',
McGraw-Hill, Inc., 1993

2.11 Exercises

1. At 25°C the magnitude of an nMOS transistor's gain factor β is $240\,\mu A/V^2$ and its threshold voltage V_T is 0.4 V.

   a) Calculate the gain when the transistor is operating at 65°C.

   b) Calculate the threshold voltage for the temperature in a).

   c) What would be the consequences of this reduced threshold voltage for the stand-by current in an SRAM, for instance?

2. A reference transistor with an aspect ratio of $\frac{W}{L} = \frac{0.3\,\mu m}{0.25\,\mu m}$ has a threshold voltage $V_T = 0.5\,V$. Measurements on transistors of various dimensions reveal that a transistor with the same width but a length of $0.5\,\mu m$ has a threshold voltage which is 100 mV higher (assume $D_{l_1} = 0$).

   a) Calculate the threshold voltage of a transistor with an aspect ratio of $\frac{0.3\,\mu m}{1\,\mu m}$.

   b) What happens to the threshold voltage when the width W of this transistor is increased to $0.5\,\mu m$ if D_w=0.04 Vμm?

3. a) What is the effect on the gain factor β of a transistor with L=0.18 μm when the mobility is only influenced by velocity saturation caused by a very large horizontal electric field, $E_x = E_{xc}/2$?

   b) Calculate the drain-source voltage at which the relevant reduction in mobility occurs if $\theta_3 = 0.2\,V^{-1}$.

Chapter 3

Manufacture of MOS devices

3.1 Introduction

Until the mid-eighties, the nMOS silicon-gate process was the most commonly-used process for MOS LSI and VLSI circuits. Examples of these circuits include microprocessors, signal processors and DRAMs with storage capacities of 16 kbit, 64 kbit and 256 kbit. However, nearly all modern VLSI circuits are made in CMOS processes. CMOS circuits are explained in chapter 4; the technology used for their manufacture is discussed in this chapter.

Modern deep-submicron CMOS processes, with channel lengths below 0.5 μm, have emerged from the numerous manufacturing processes which have evolved since the introduction of the MOS transistors in integrated circuits. Differences between the processes were mainly characterised by the following features:

- The minimum feature sizes that can be produced.
- Gate oxide thickness.
- The number of interconnection levels.
- The type of substrate material. Alternatives include n-type and p-type silicon wafers, substrates with an epitaxial layer and isolating substrates.

- The choice of the gate material. Initially, the gate material was the aluminium implied in the acronym MOS (Metal Oxide Semiconductor). Molybdenum has also been used. Modern MOS processes, however, nearly all use *polycrystalline* silicon (polysilicon) as gate material. One of the main reasons is that a polysilicon gate facilitates the creation of self-aligned source and drain areas. Another reason for using polysilicon as gate material is that it allows accurate control of the formation of the gate oxide.
- The method to isolate transistors. The main distinction is between processes which use the so-called LOCOS isolation and processes which use Shallow-Trench Isolation (STI), see section 3.4.
- The type of transistors used: nMOS, pMOS, enhancement and/or depletion, etc.

Modern manufacturing processes consist of numerous photolithographic, etching, oxidation, deposition, implantation, diffusion and planarisation steps. These steps are frequently repeated throughout the process and they currently total a few hundred. This chapter starts with a brief description of each step. Most processes use masks to define the required patterns in all or most of the IC diffusion and interconnect layers. Modern CMOS manufacturing processes use between 20 and 30 masks. However, the initial discussion of IC manufacturing processes in this chapter focuses on a basic nMOS process with just five masks.

Subsequently, a basic CMOS process flow is briefly examined. Fundamental differences between various CMOS processes are then highlighted.

Finally, a sample deep-submicron CMOS process is explained. Many of the associated additional processing steps are an extension of those in the basic CMOS process flow. Therefore, only the most fundamental deviations from the conventional steps are explained. The quality and reliability of packaged dies are important issues in the IC manufacture industry. An insight into the associated tests concludes the chapter.

3.2 Lithography in MOS processes

The integration of a circuit requires a translation of its specifications to a description of the layers necessary for IC manufacture. Usually, these layers are represented in a *layout*. The generation of such a layout is

usually done via an interactive graphics display for handcrafted layouts, or by means of synthesis and place-and-route tools, as discussed in chapter 7. Figure 3.1 shows an example of a complex IC containing several synthesized functional blocks.

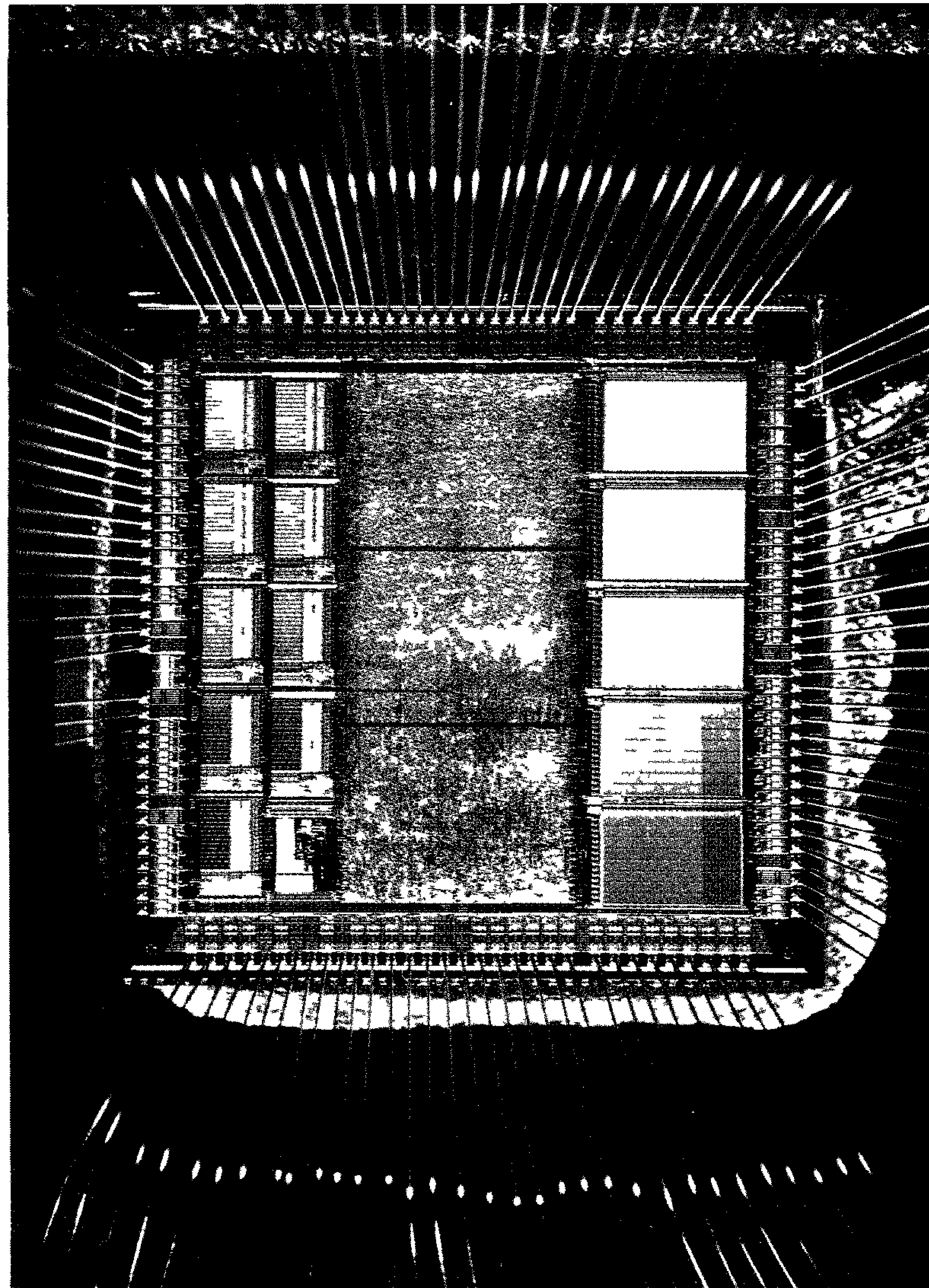


Figure 3.1: *Example of a complex signal processor chip, containing several synthesized functional blocks*

A complete design is subjected to functional, electrical and layout design rule checks. If these checks prove satisfactory, then the layout is stored in a computer file. A software program (post-processor) is used to convert this database to a series of commands. These commands control an *Electron-Beam Pattern Generator* (EBPG) or a *Laser-Beam Pattern Generator* (LBPG), which creates an image of each mask on a photographic plate called a *reticle*. In only a few years time, the minimum feature size of 2.5 μm on a 5x reticle for a 0.5 μm process has been scaled to 1 μm on a 4x reticle for a 0.25 μm process, see figure 3.2. The reticle pattern is thus demagnified as it passes through the projection optics.

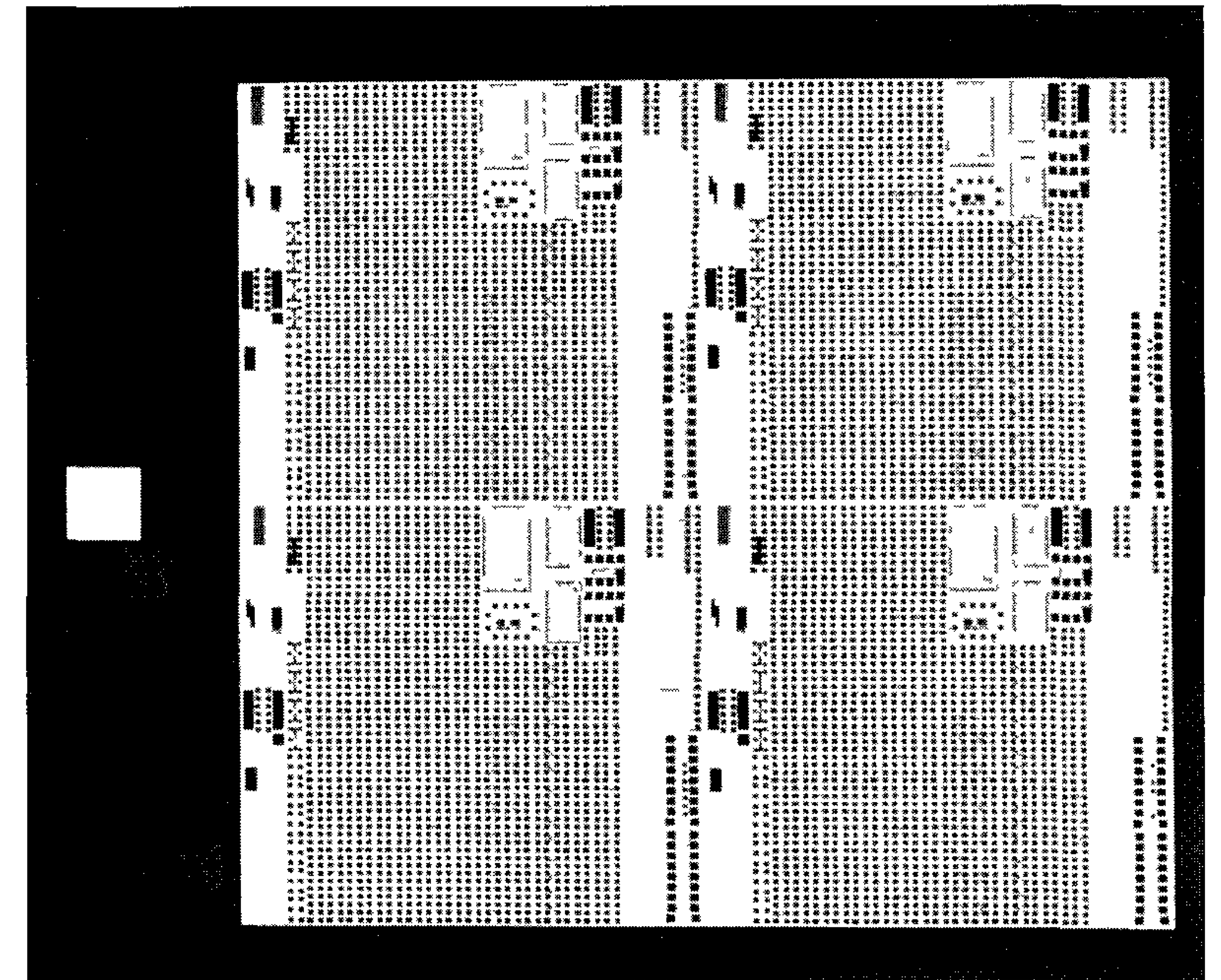


Figure 3.2: *Example of a 4x reticle*

Smaller feature sizes, as currently required in submicron and deep-submicron (< 0.5 μm channel lengths) processes, are obtained by using reduction steppers. These reduction steppers currently use *five-to-one (5x)* for minimum feature sizes up to 0.35 μm and *four-to-one (4x) reduction step-and-repeat* lithography for a 0.25 μm CMOS process. The reduction is achieved by means of a system of (very) complex lenses. Figure 3.3 shows a basic schematic of a stepper.

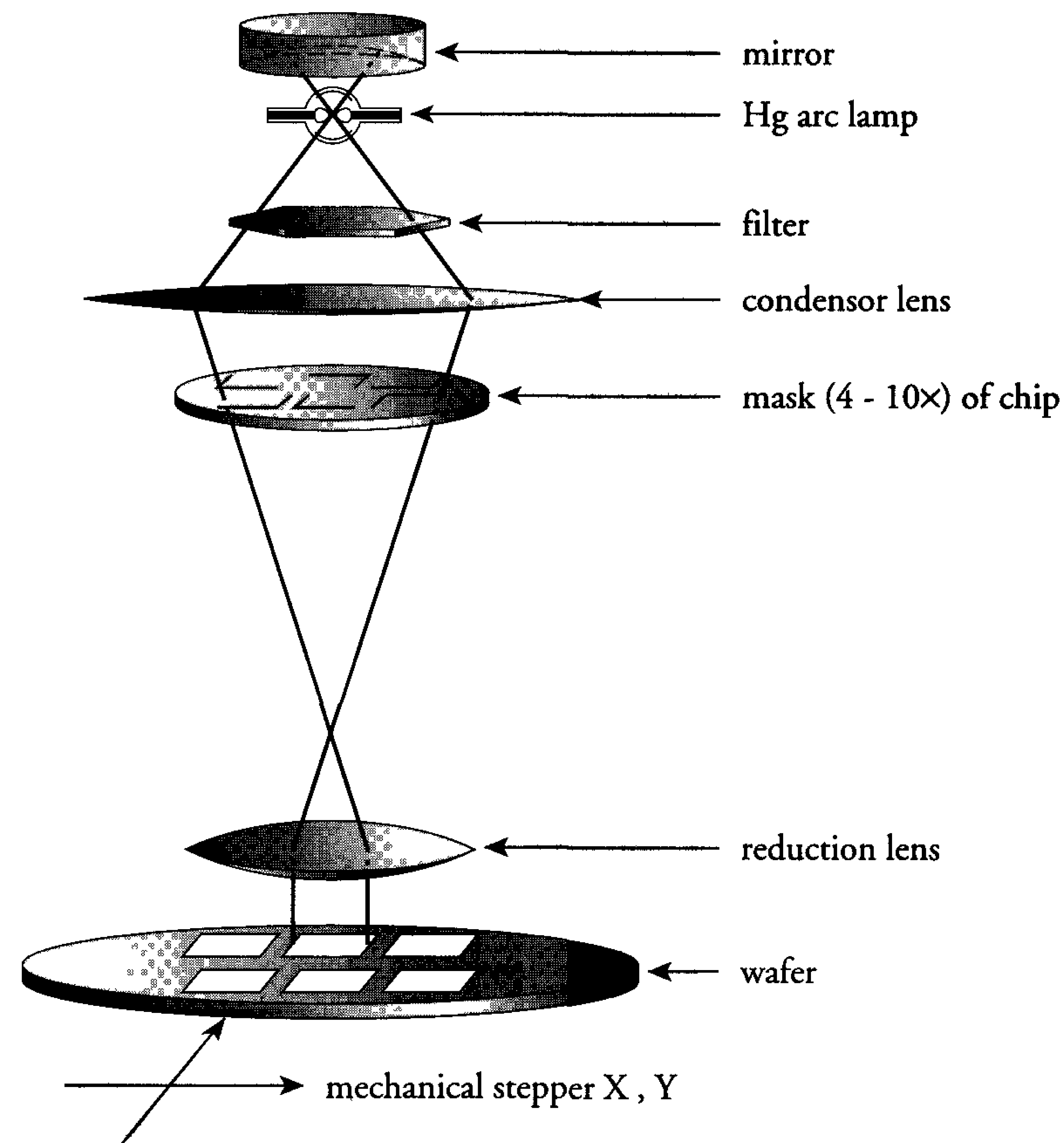


Figure 3.3: *Basic schematic of a stepper (source: American Chemical Society)*

Limitations of these projection lithography techniques are determined by the wavelength (λ) of the applied light source and the *Numerical Aperture* (NA) of the lens. The minimum resolution of the resulting projections is limited by diffraction and also depends on the properties of the photoresist. Better photoresists allow smaller minimum feature sizes. A generally used expression for the *minimum feature size* (maximum resolution) of the resulting projection is:

$$\text{minimum feature size } = k_1 \cdot \frac{\lambda}{NA}$$

where k_1 is a constant defined by the lithographic process.

Typical values for k_1 are between 0.4-0.7. The numerical aperture generally ranges from 0.3 to 0.7. Because of these values for NA and k_1, the minimum 'printable' feature size is almost equal to the wavelength of the light.

The currently most commonly-used light sources are mercury (Hg) lamps, which emit light at different wavelengths of the spectrum. These wavelengths are 436 nm for the g-line, 405 nm for the h-line, 365 nm for the i-line and 248 nm for the Deep-UV (DUV) line. Current 0.35 μm feature sizes are mostly printed on i-line (365 nm) steppers, while the 0.25 μm feature sizes are printed on the 248 nm DUV steppers. For these 248 nm steppers, either a mercury (Hg) light source or a krypton fluoride (KrF) light source can be used. The 248 nm DUV-line will be followed by the 193 nm DUV-line from an argon fluoride (ArF) and the 157 nm DUV-line from a fluorine laser source. The 193 nm and 157 nm steppers are expected to be used to print the 0.13 μm and possibly the 0.10 μm minimum feature sizes. However, the development of 193 nm DUV is probably too late for this feature size. According to Sematech officials, it might be possible to use high NA, 193 nm projection optical lithography to print 0.1 μm feature sizes as well, if it is combined with advanced mask technologies (including phase shift and Optical Proximity Correction (OPC) techniques) and resist technologies. It is expected that the traditional optical lithography will no longer be viable at shorter wavelengths ($\approx$ 0.07 μm), as optically transparent materials used for lens and mask will become completely absorbent at these wavelengths. Alternative lithography techniques will have to be very advanced and will not be attractive in terms of costs, although optical methods with advanced mask technologies are not cheap either. This is why the semiconductor industry is hoping that the use of optical lithography can be extended for many new process generations.

Assuming that non-optical lithography will first be introduced to image-critical masks in a 0.07 μm technology, the industry still has some time to choose the best of the expected alternatives: N:1 projection electron-beam lithography, 1:1 proximity X-ray lithography, projection ion lithography and all reflective Extreme-UV (EUV) lithographies. Because the choice is not clear yet, a detailed discussion of any one of these is beyond the scope of this book.

However, one alternative lithography has already been in use for quite some time to manufacture so-called multi-project wafers. As the manufacture of masks is much too expensive for small IC-design projects,

especially for prototyping or educational purposes, *Direct Slice Writing (DSW)* techniques are used. Such techniques use an *Electron-beam (E-beam) machine*, which writes the layout pattern directly onto a wafer resist layer, without using a mask. The resolution yielded by an E-beam machine is better than 0.07 μm, but at a lower throughput.

To summarise the evolution of the wafer stepper, table 3.1 presents several key parameters which reflect the improvements made over different generations of steppers [8,17].

Table 3.1: The evolution of the wafer stepper (source: [8,17])

Parameters	1977 GCA 4800 DSW	1995 i-line	2001 DUV: Step Scan
M:1	10x	5x	4x
Wavelength	g-line: 436 nm	i-line: 365 nm	DUV: 193 nm
Lens	0.28 NA	0.60 NA	0.85 NA
Resolution	1.25 μm	0.40 μm	0.13 μm
Field size	10 mm square	22 mm square	26 mm x 32 mm
Depth of focus	4.0 μm	1.0 μm	0.40 μm
Alignment (overlay)	±0.50 μm	±0.06 μm	±0.02 μm
Stage accuracy	100 nm	30 nm	5 nm
Lens distortion	250 nm	50 nm	< 25 nm
Wafer size	3, 4, 5 inch	5, 6, 8 inch	6, 8, 12 inch
Throughput	20 wph (4")	60 wph (6")	> 80 wph (8")
Cost	$300,000	$4,000,000	$12,000,000

Pattern imaging

The photolithographic steps involved in the transfer of a mask pattern to a wafer are explained with the aid of figure 3.4. Usually, the first step is oxidation and comprises the growth of a 30 to 50 nm thick silicon-dioxide (SiO_2) layer on the wafer. Subsequently, a nitride (Si_3N_4) layer is deposited. Next, this nitride layer is covered with a *photoresist layer*. The mask is used to selectively expose the photoresist layer to light. The photoresist is then developed, which leads to the removal of the exposed areas if the photoresist is positive. The resulting pattern in the resist acts as a barrier in the subsequent nitride etching step, in which the unprotected nitride is removed. Finally, the remaining resist is removed

and an image of the mask pattern remains in the nitride layer. This nitride pattern acts as a barrier for a subsequent processing step.

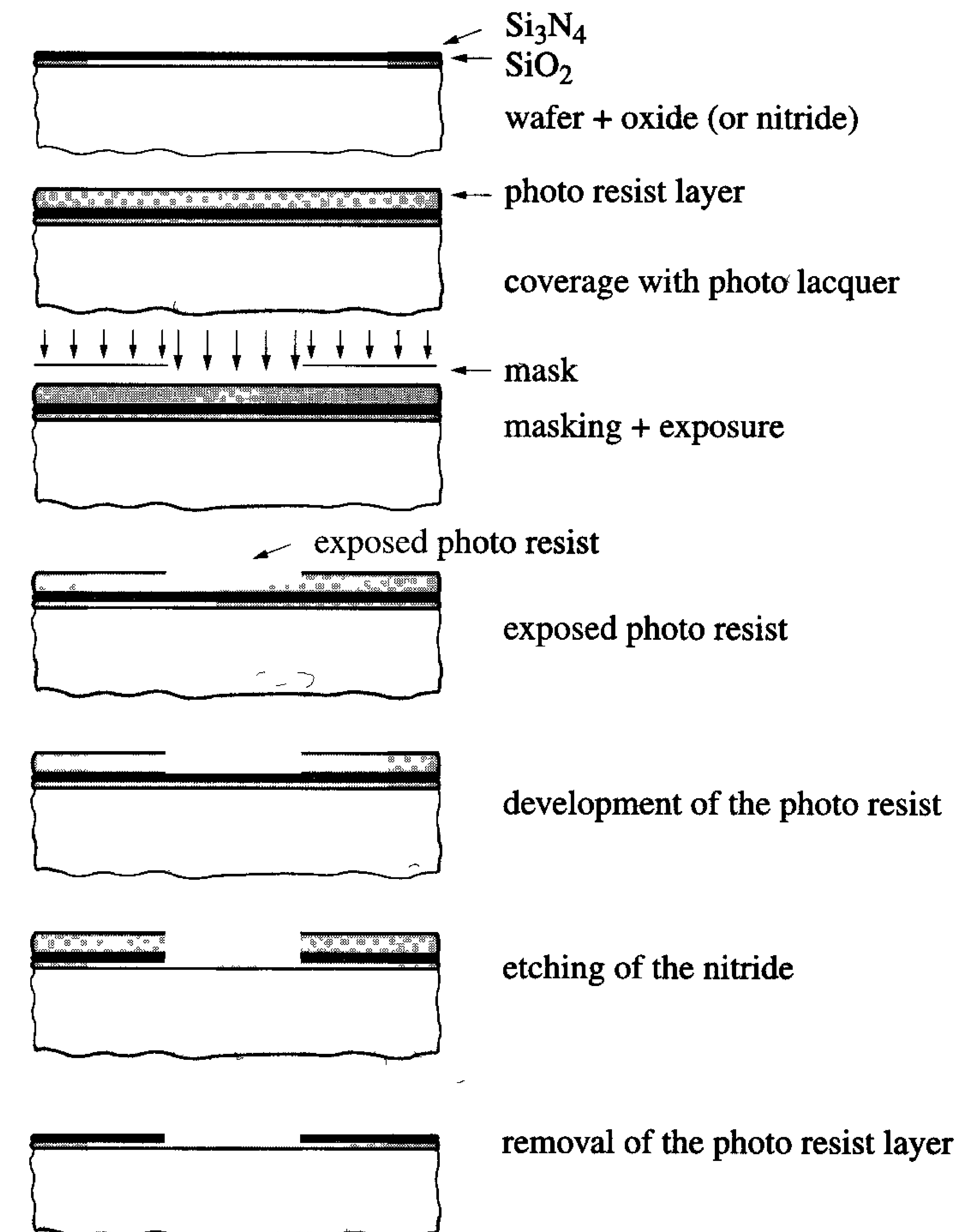


Figure 3.4: *Pattern transfer from mask to wafer*

Both positive and negative resists exist. The differences in physical properties of these resist materials result in inverting images, see figure 3.5.

The combination of pattern transfer and a processing step is repeated for all masks required to manufacture the IC. The types of layers used for the pattern transfer may differ from the silicon-dioxide and silicon-nitride layers described above.

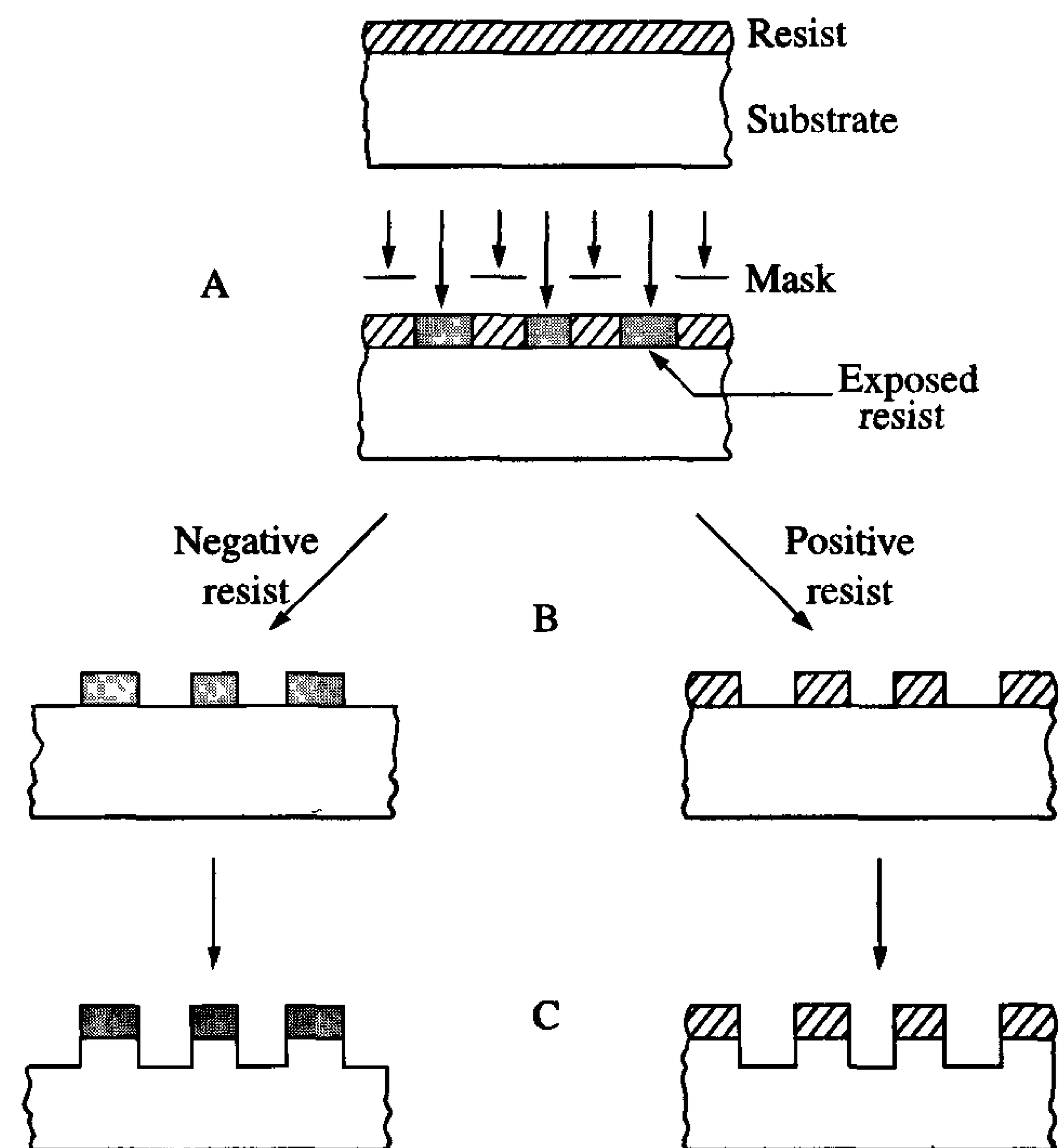


Figure 3.5: *The use of positive and negative resist for pattern imaging*

The principle, however, remains the same. The processing steps that follow pattern transfer may comprise etching, oxidation, implantation or diffusion and planarisation. Deposition is also an important processing step. These steps are described in detail in the following sections.

3.3 Etching

The previously-described photolithographic steps produce a pattern in a nitride or equivalent barrier layer. This pattern acts as a protection while its image is duplicated on its underlying layer by means of *etching* processes. There are several different etching techniques. With *wet etching*, the wafer is immersed in a chemical etching liquid. Dry etching methods may consist of both physical and chemical processes (anisotropic) or of a chemical process only (isotropic). The wet-etching methods are *isotropic*, i.e. the etching rate is the same in all directions. The associated *'under-etch'* problem illustrated in figure 3.6(a) becomes serious when the minimum line width of the etched layer approaches its thickness.

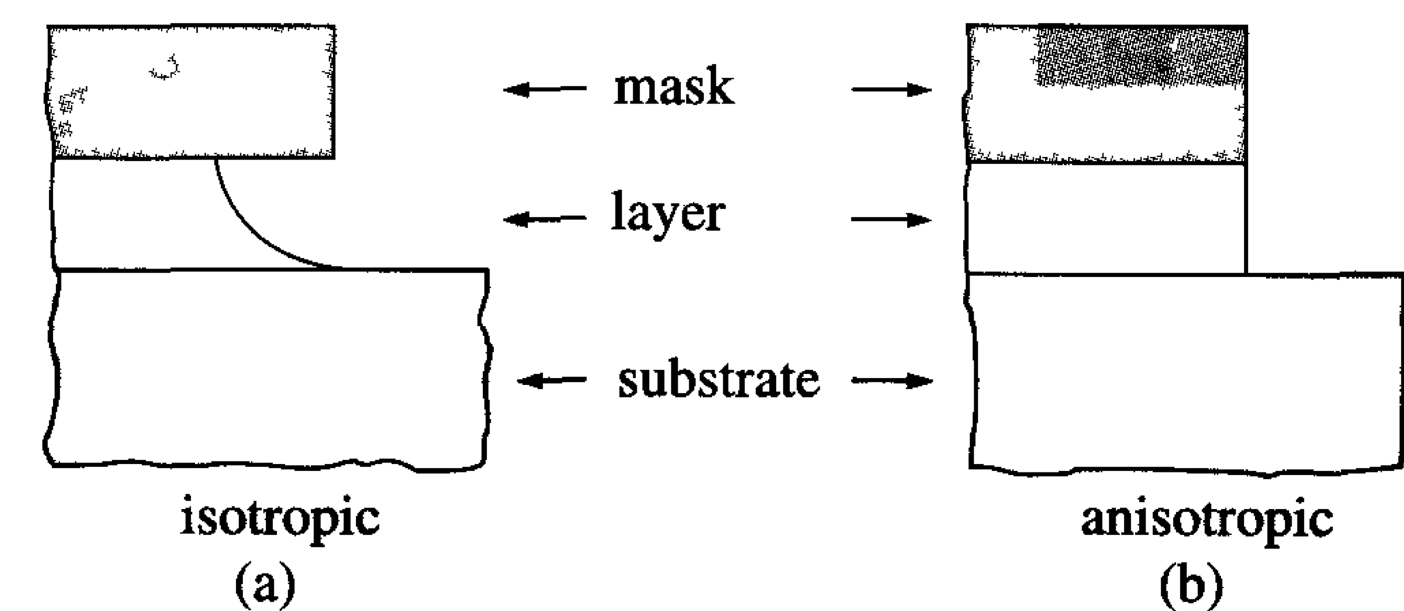


Figure 3.6: *The results of different etching methods*

Dry-etching methods, which use a plasma, allow *anisotropic* etching, i.e. the etching process is limited to one direction by the perpendicular trajectory of the ions used at the wafer surface. The result, shown in figure 3.6(b), is an accurate copy of the mask pattern on the underlying layer.

With *plasma etching* techniques, the wafers are immersed in a plasma containing chlorine or fluorine ions that etch e.g. Al and SiO_2, respectively. In *sputter etching* techniques, the wafer is bombarded by gas ions such as argon (Ar^+). As a result, the atoms at the wafer surface are physically dislodged.

Finally, a combination of plasma and sputter etching techniques is used in *Reactive Ion Etching* (RIE). Satisfactory etching processes have been developed for most materials that are currently used in IC manufacturing processes. New process generations, however, require improved selectivity, uniformity, reproducibility and process control. Selectivity can be improved by the compound of the gaseous plasma or by the creation of polymers at the underlying layer. The use of an additional carbonaceous substance such as CHF_3 during etching enhances its anisotropic properties. The use of this substance creates a thin layer close to the side wall of a contact hole, for example, which improves the anisotropy of the etching process. A second advantage is that carbon reacts with oxygen; actually, it consumes oxygen. It therefore increases the selectivity of the etching process because, when used in the etching of a contact-to-silicon, the reaction is stopped immediately on arrival at the silicon surface. Carbon does not react with silicon.

For critical anisotropic etching steps, both low-pressure etching techniques and *High-Density Plasma (HDP)* techniques are used. In HDP, energy is coupled into the plasma inductively to increase the number of electrons. HDP is operated at low (some mtorr) pressure. This in turn

results in a higher plasma density and a higher degree of ionisation. HDP is used to provide high-aspect ratios.

The focus on new etching techniques does not preclude further development of existing techniques such as high-pressure etching and RIE.

Many process steps use plasma or sputter-etching techniques, in which charged particles are collected on conducting surface materials (polysilicon, metals). These techniques can create significant electrical fields across the thin gate oxides; this is called the *antenna effect*. The gate oxide can be stressed to such an extent that the transistor's reliability can no longer be guaranteed. The antenna effect can also cause a V_T-shift, which affects matching of transistors in analogue functions. It is becoming industry practice to introduce additional "antenna design rules" to limit the ratio of antenna area to gate oxide area. Also, protection diodes are used to shunt the gate.

3.4 Thermal oxidation

The dielectrics used in the manufacture of deep-submicron CMOS circuits must fulfil several important requirements [9]:

- high breakdown voltage
- low dielectric constant
- no built-in charge
- good adhesion to other process materials
- low defect density (no pinholes)
- easy to be etched
- permeable to hydrogen.

One of the materials that incorporates most of these properties is silicon dioxide (SiO_2). SiO_2 can be created by different processes: thermal *oxidation* or deposition. A *thermal oxide* was used to isolate the transistor areas in conventional MOS ICs. In these isolation areas, the oxide must be relatively thick to allow low capacitive values for signals (tracks) which cross these areas. This *thick oxide* was created by exposing the monocrystalline silicon substrate to pure oxygen or water vapour at a high temperature of 900°C to 1200°C. The oxygen and water vapour

molecules can easily diffuse through the resulting silicon dioxide at these temperatures. The following respective chemical reactions occur when the oxygen and water vapour reach the silicon surface:

Dry oxidation : Si (solid) + O_2 (vapour) $\longrightarrow$ SiO_2 (solid)

Wet oxidation : Si (solid) + $2H_2O$ (vapour) $\longrightarrow$ SiO_2 (solid) + 2 H_2

The *LOCOS* (Local Oxidation of Silicon) process is an oxidation technique which has found universal acceptance in MOS processes with gate lengths down to 0.5 μm. Silicon is substantially consumed at the wafer surface during this process. The resulting silicon-dioxide layer extends about 46% below the original wafer surface and about 54% above it. The exact percentages are determined by the concentration of the oxide, which contains about $2.3 \cdot 10^{22}$ atoms/cm^3, while silicon contains about $5 \cdot 10^{22}$ atoms/cm^3. A disadvantage of the LOCOS process is the associated rounded thick oxide edge. This *bird's beak* is shown in figure 3.7(a).

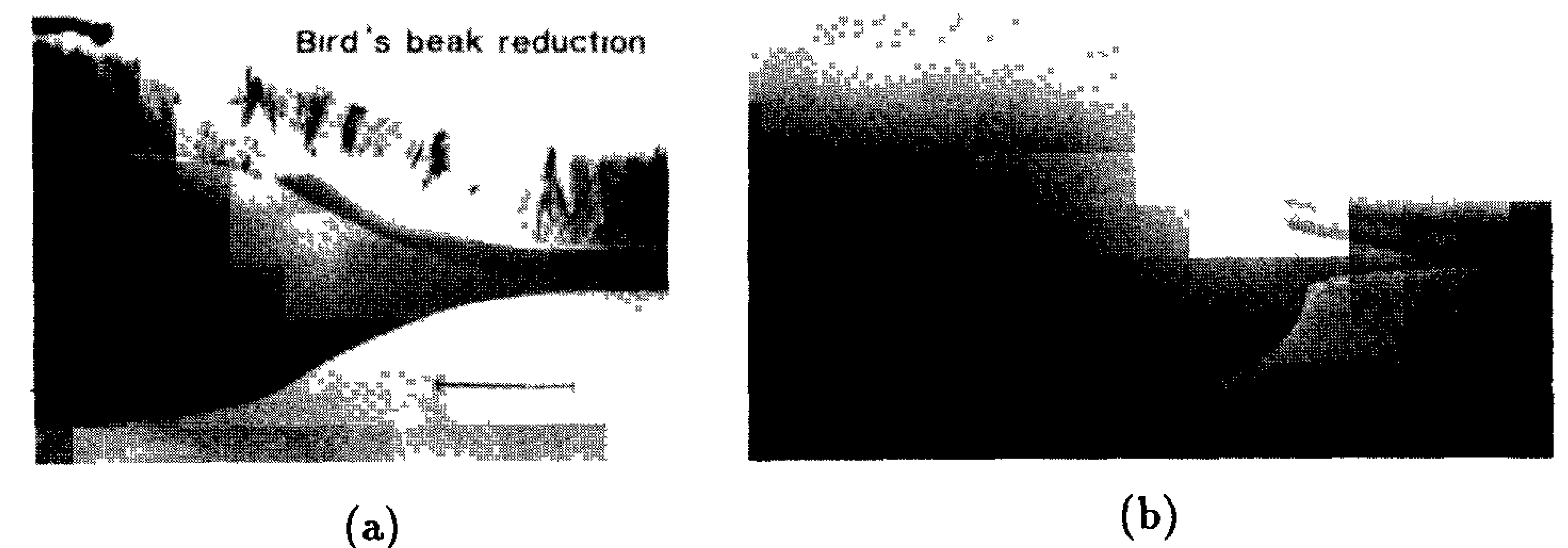


Figure 3.7: *Comparison of (a) a conventional LOCOS process and (b) a new oxide formation process which yields a suppressed bird's beak*

The formation of the bird's beak causes a loss of geometric control, which becomes considerable as transistor sizes shrink. Intensive research efforts aimed at suppression of bird's beak formation have resulted in lengths of just 0.1-0.15 μm for an oxide thickness of 0.5 μm. Such a bird's beak is shown in figure 3.7(b).

Even with a suppressed bird's beak, the use of LOCOS is limited to the isolation of sub-0.25 μm transistors. Several advanced LOCOS isolation techniques, however, are currently being investigated for their

potentials and limitations for further use in deep-submicron technologies. One uses polysilicon to encapsulate the LOCOS during its formation: *Polysilicon-Encapsulated LOCOS* (PE-LOCOS). Another example is the *Locally Sealed LOCOS* (LS-LOCOS), which shows excellent bird's beak control and little field oxide thinning, by using a local nitride interface sealing. LOCOS isolation techniques use thermally-grown S_iO_2.

An important alternative to these LOCOS techniques, already used in 0.35 μm CMOS technologies, is the *Shallow-Trench Isolation (STI)*. STI uses deposited dielectrics to fill trenches which are etched in the silicon between active areas. The use of STI for deep-submicron technologies is discussed later in this chapter (section 3.8.3).

Another important application of thermally grown oxide is the oxide layer between a transistor gate and the substrate. This *'gate oxide'* must be of high quality and very reliable. Defects such as pinholes and oxide charges have a negative effect on electrical performance and transistor lifetime. The gate oxide requires continuous special attention as its thickness is constantly being reduced with each new process generation.

The *gate oxide* thickness scales according to table 3.2, which reflects the year of volume shipment.

Table 3.2: Gate oxide scaling over a decade (the indicated year reflects volume shipment)

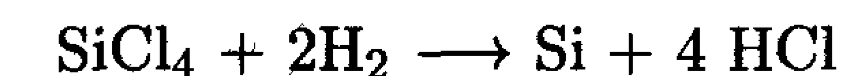
Year	Channel length	Gate oxide thickness
1993	0.8 μm	15 nm
1996	0.5 μm	12 nm
1998	0.35 μm	7 nm
2001	0.25 μm	≈4.5 nm
2003	0.18 μm	≈3.5 nm
2005	0.12 μm	≈2.0 nm

The use of dielectric SiO_2 layers below about 2 nm thickness causes gate oxide direct *tunnelling*, resulting in currents which may exceed a level of 1 A/cm^2. At these thin gate oxides, boron out-diffusion from the p^+ polysilicon gate also starts penetrating the gate oxide, and can lead to auto-doping of the channel as a result of diffusion. Moreover, the combination of thinner gate oxide and increased channel dopes also causes depletion of the bottom region of the gate material. This is called *gate depletion*. As a result of these effects, the currently-used *double-flavoured* polysilicon gate material will eventually fade away (n^+ doped gate for nMOS transistors and p^+ doped gate for pMOS transistors). A lot of R&D resources are required to find acceptable alternatives for both the gate oxide and the gate material itself.

3.5 Deposition

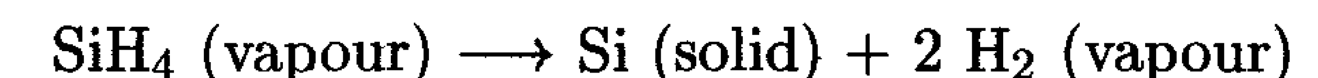
The *deposition* of thin layers of dielectrical material, polysilicon and metal is an important facet of IC production.

The growth of an *epitaxial film* (layer) is an example of the deposition of a thin (0.25-20 μm) layer of single crystal material on top of a single crystal substrate (or wafer). If the deposited layer is the same material as the substrate, it is called *homo-epitaxy* or epi-layer for short. Silicon on sapphire is an example of *hetero-epitaxy*, in which the deposited and substrate materials differ [10]. Epitaxial deposition is created by a *Chemical Vapour Deposition (CVD)* process. This is a process during which vapour-phase reactants are transported to and react with the substrate surface, thereby creating a film and some by-products. These by-products are then removed from the surface. Normally, the actual film created by a CVD process is the result of a sequence of chemical reactions. However, a different overall reaction can generally be given for each of the silicon sources. The hydrogen reduction of silicon tetrachloride ($SiCl_4$), for example, can be represented as:



$$SiCl_4 + 2H_2 \longrightarrow Si + 4\ HCl$$

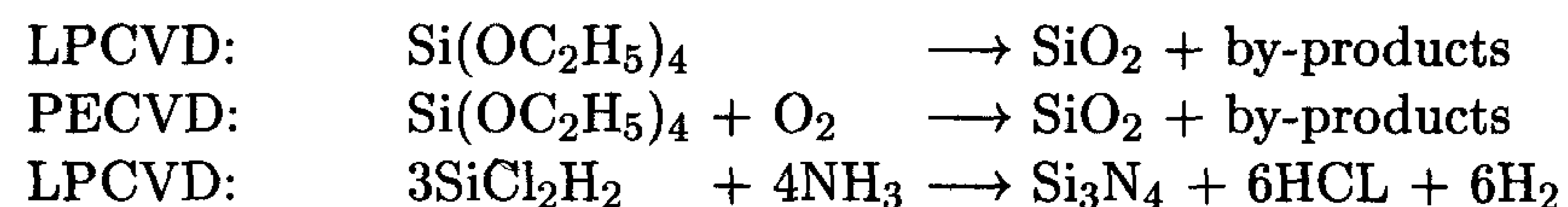
Several parameters determine the growth rate of a film, including the source material and deposition temperature. Usually, high temperatures (> 1000 °C) are used for the depositions because the growth rate is then less dependent on the temperature and thus shows fewer thickness variations. The overall reaction for the deposition of polysilicon is:



$$SiH_4\ \text{(vapour)} \longrightarrow Si\ \text{(solid)} + 2\ H_2\ \text{(vapour)}$$

This reaction can take place at lower temperatures, because SiH_4 decomposes at a higher rate. The creation of dielectric layers during IC manufacture is also performed by some form of CVD process. The most commonly-used dielectric materials are silicon dioxide (SiO_2) and silicon

nitride (Si_3N_4). In an Atmospheric-Pressure CVD *(APCVD)* process, the material is deposited by gas-phase reactions. This deposition generally results in overhangs and a poor step coverage (figure 3.8). APCVD is currently used to deposit Boron PhosphoSilicate Glass *(BPSG)* epitaxial layers and form the scratch-protection layer (PSG). BPSG is a dielectric which is deposited on top of polysilicon (between polysilicon and first metal). BPSG contains boron and phosphorus for a better flow (spread) of the dielectric. The phosphorus also serves to improve internal passivation. The following reactions apply for the deposition of SiO_2 and Si_3N_4, respectively:

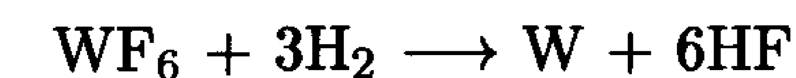


$$\begin{array}{lll} \text{LPCVD:} & Si(OC_2H_5)_4 & \longrightarrow SiO_2 + \text{by-products} \\ \text{PECVD:} & Si(OC_2H_5)_4 + O_2 & \longrightarrow SiO_2 + \text{by-products} \\ \text{LPCVD:} & 3SiCl_2H_2 \quad + 4NH_3 & \longrightarrow Si_3N_4 + 6HCL + 6H_2 \end{array}$$

Two versions of CVD have been introduced by the above reactions: LPCVD and PECVD. *LPCVD* is a low-pressure CVD process, usually performed in a vacuum chamber at medium vacuum (0.25-2.0 torr) and at temperatures between 550 and 750 °C. Under these conditions, the vapour-phase reactions are suppressed, while the decomposition now occurs at the surface, leading to a much better step coverage. In the previously-discussed CVD process, the chemical reactions are initiated and sustained only by thermal energy. *PECVD* is a plasma-enhanced CVD process. A *plasma* is defined to be a partially ionised gas which contains ions, electrons and neutrals. The plasma is generated by applying an RF field to a low-pressure gas, thereby creating free electrons within the discharge regions [10]. The electrons gain sufficient energy so that they collide with gas molecules, thereby causing gas-phase dissociation and ionisation of the reactant gases. At room temperature, a plasma therefore already contains high-energy electrons. Thus, even at low temperatures, a PECVD process can generate reactive particles; it therefore has a higher deposition rate than other CVD processes.

If we compare the previous reactions to depositing SiO_2, we see that the LPCVD which occurs at high temperature therefore needs no additional oxygen, while the PECVD process needs additional oxygen because the oxygen cannot be dissociated from the *TEOS* (tetra ethylorthosilicate: $Si(OC_2H_5)_4$) at low temperatures. A Sub-Atmospheric CVD *(SACVD)* process occurs at temperatures around 700 to 800 °C. Because of the high pressure ($\approx 1/2$ atmosphere instead of a few torr), the deposition speed will be higher, resulting in a higher throughput. This form of CVD is particularly used for BPSG.

Metal layers are deposited by both physical and chemical methods. The physical methods are *evaporation* and *sputtering*. In the sputtering technique, an aluminium target is bombarded with argon ions and a flux of aluminium flows from the target to the wafer surface. CVD methods form the chemical alternative for the deposition of metals. Tungsten (W), for example, yields the following CVD reaction:



$$WF_6 + 3H_2 \longrightarrow W + 6HF$$

The choice of deposition method is determined by a number of factors, of which *step coverage* is the most important. Figure 3.8 shows an example of bad aluminium step coverage on a contact hole. Such a step coverage can dramatically reduce the lifetime of an IC. It also causes problems during further processing steps and the associated temperature variations can lead to voids in the aluminium.

Moreover, the local narrowings cannot withstand high current densities. *Current densities* of $\approx 10^5$ A/cm^2 are not exceptional in modern integrated circuits. Excessive current densities in metal tracks cause *electromigration*. This leads to the physical destruction of metal tracks and is another phenomenon that reduces the reliability of ICs. This topic is examined more closely in chapter 9.

Figure 3.8: *Example of poor step coverage*

3.6 Diffusion and ion implantation

Diffusion and *ion implantation* are the two commonly-used methods to force impurities or dopants into the silicon.

Diffusion

Diffusion is the process by which the impurities are spread as a result of the existing gradient in the concentration of the chemical. Diffusion is often a two-step process.

The first step is called *pre-deposition* and comprises the deposition of a high concentration of the required impurity. The impurities penetrate some tenths of a micrometre into the silicon, generally at temperatures between 700 to 900°C. Assuming that the impurities flow in one direction, then the flux is expressed as:

$$J = -D \cdot \frac{\delta C(x,t)}{\delta x}$$

where D represents the *diffusion coefficient* of the impurity in [cm^2/s] and $\frac{\delta C}{\delta x}$ is the impurity concentration gradient.

As the diffusion strongly depends on temperature, each different diffusion process requires individual calibration for different processing conditions. During the diffusion process, silicon atoms in the lattice are then substituted by impurity atoms.

The second step is called *drive-in diffusion*. This high-temperature ($> 1000\,^\circ$C) step decreases the surface impurity concentration, forces the impurity deeper into the wafer, creates a better homogeneous distribution of the impurities and activates the dopants. As a result of the increased requirements of accurate doping and doping profiles, diffusion techniques are losing favour and ion implantation has become the most popular method for introducing impurities into silicon.

Ion Implantation

The ion implantation process is quite different from the diffusion process. It takes place in an *ion implanter*, which comprises a vacuum chamber and an ion source that can supply phosphorus, arsenic or boron ions,

for example. The silicon wafers are placed in the vacuum chamber and the ions are accelerated to the silicon under the influence of electric and magnetic fields. The *penetration depth* in the silicon depends on the ion energy. This is determined by the mass and electrical charge of the ion and the value of the accelerating voltage. Ion implanters are equipped with a mass separator (analysing magnet), which ensures that only ions of the correct mass and charge can reach the silicon wafer. Ion implantation is characterised by the following four parameters:

- The type of ion. Generally, this is phosphorus, arsenic or boron. The mass and electrical charge of the ion are important.
- The accelerating voltage (V), which varies from tens to hundreds of kilovolts.
- The current strength (I), which lies between 0.1μA and 1mA.
- The implantation duration (t), which is in the order of tens of seconds per wafer. The total charge $Q = I \cdot t$ determines the number of ions that will enter the silicon. Typical doses range from 10^{11}-10^{16} atoms/cm^2.

Variables V, I and t can be measured with very high accuracy. This makes ion implantation much more reproducible for doping silicon than classical diffusion techniques. In addition, V and I can be varied as a function of t to produce a large variety of dope profiles that are not possible with diffusion. The maximum impurity concentration is almost always at the surface when diffusion techniques are used.

The ion implantation technique, however, can be used to selectively create profiles with peaks below the wafer surface. The concentration of impurities decreases toward the wafer surface in these *'retrograde profiles'*. The most important material that is used to mask ion implanting is photoresist. Ion implantation causes serious damage (disorder) in the crystal lattice of the target. In addition, only a fraction of the implanted ions occupies a silicon atom location. The other part does not occupy lattice sites. The *interstitial dope atoms* are electrically inactive and do not operate as donors or acceptors. A subsequent thermal step, at temperatures between 800 to 1000°C, is used to recover the crystal structure. This *annealing process* causes the vast majority of the dopants to become electrically active on the lattice sites.

Ion implantation adds flexibility and increased process control to CMOS manufacture. It is superior to chemical deposition techniques

for the control of impurities ranging from 10^{14} to 10^{18} atoms/cm^3. The heart of an ion implanter is formed by an ion source and a 90° analysing magnet. Because the ion beam is a mixture of different fractions of molecules and atoms of the source material, the 90° analysing magnet causes only the selected ions to reach the resolving aperture, see figure 3.9 and [10].

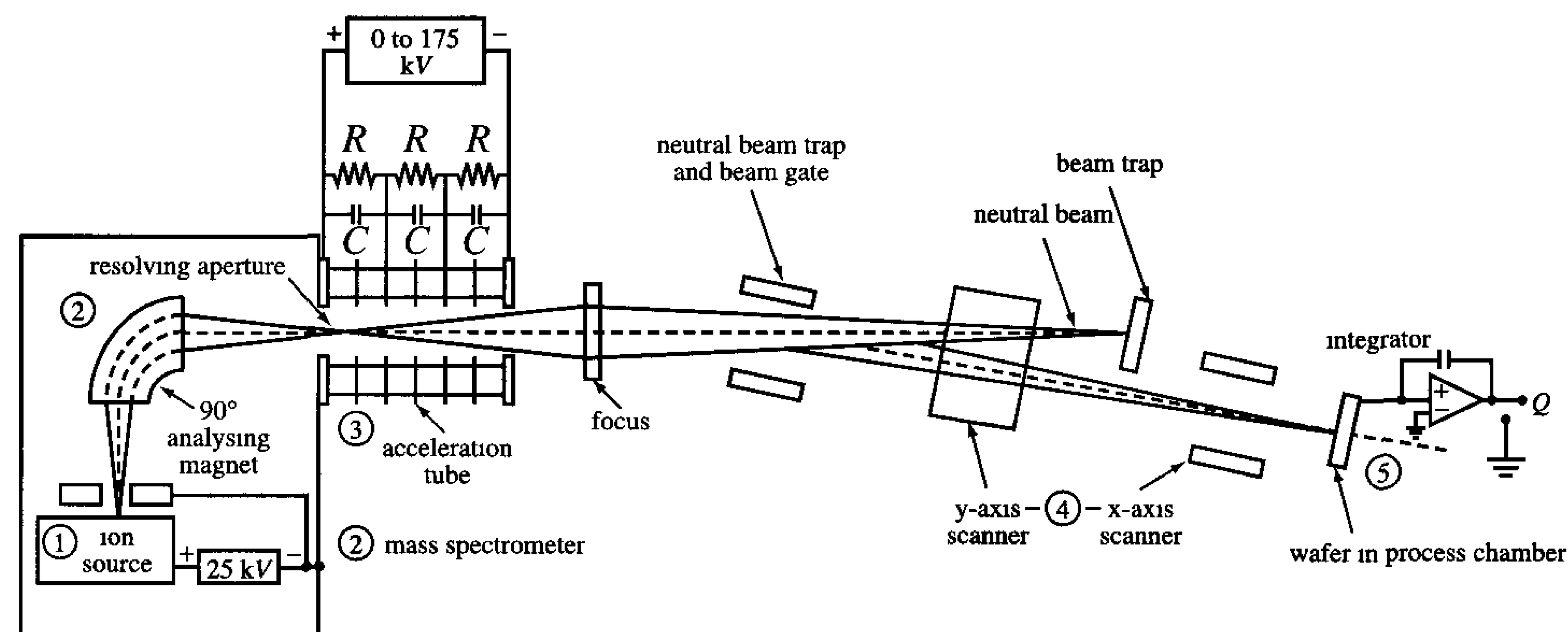


Figure 3.9: *Basic set-up of an ion implanter (Source: [10])*

Examples of the use of ion implantation are:

- threshold voltage adjustment
- retrograde-well implantation
- channel-stop implantation
- source/drain formation (0.15-0.4 μm shallow regions)
- etc.

Disadvantages of ion implantation include the following:

- lateral distribution of impurities is not zero
- throughput is lower than in diffusion process
- it uses complex implanters, which requires well-trained operators
- initial cost of equipment: 4 to 6 M$.

The use of ion implantation in the formation of source/drain regions becomes increasingly challenging as these junctions become very shallow in scaled processes. The dope concentration does not increase with scaling. Only the energy during implantation must be adjusted to create those shallow junctions. *Silicidation* of sources and drains becomes a problem in that silicide can penetrate through the shallow junctions. This is called *junction spiking*. Unsilicided sources and drains show a factor five increase in sheet and contact resistance, affecting the electrical properties of the transistors.

3.7 Planarisation

The increase in the number of processing steps, combined with a decrease in feature sizes, results in an increasingly uneven surface. Therefore, all submicron and deep-submicron processes use several planarisation steps. These steps flatten or 'planarise' the surface before the next processing step is performed.

In conventional CMOS processes, *planarisation* was used during the back-end of the process, i.e. in between the formation of successive metal layers to flatten the surface before the next metal layer was defined. In such a *Spin-On-Glass (SOG)* formation, the surface was coated with a liquid at room temperature. After this, the wafer was rotated (spun), such that the liquid flowed all over the wafer to equalise the surface. Next, the wafer was cured to form a hard silicate or siloxane film. To prevent cracking, phosphorus was often incorporated in the film. The resulting dielectric layer was planarised to a certain extent. An advantage of SOG is that very small gaps are easy to fill. However, with SOG, the surface is locally, but not globally, planarised, see figure 3.10. On locally rough areas (A), the surface is reasonably planarised. A disadvantage of SOG is that it is a contaminated material.

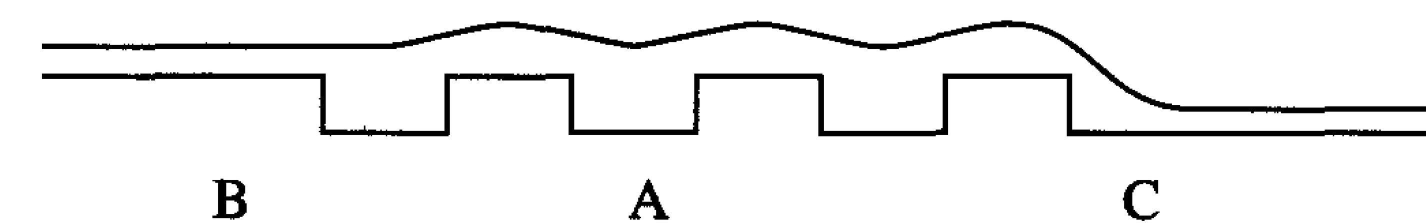


Figure 3.10: *SOG planarisation results*

On top of closed areas (B) and open areas (C), the dielectric thickness will be less than in the small open parts in rough areas (A). In a multilevel metal chip, this effect would be much worse and would lead to

etching problems and problems with the *Depth Of Focus (DOF)* of the stepper. In most advanced technologies of 0.25 μm and below, a very good alternative planarisation technique is used: *Chemical Mechanical Polishing (CMP)*.

CMP is based on the combination of mechanical action and the simultaneous use of a chemical liquid (slurry) and actually polishes the surface, see figure 3.11.

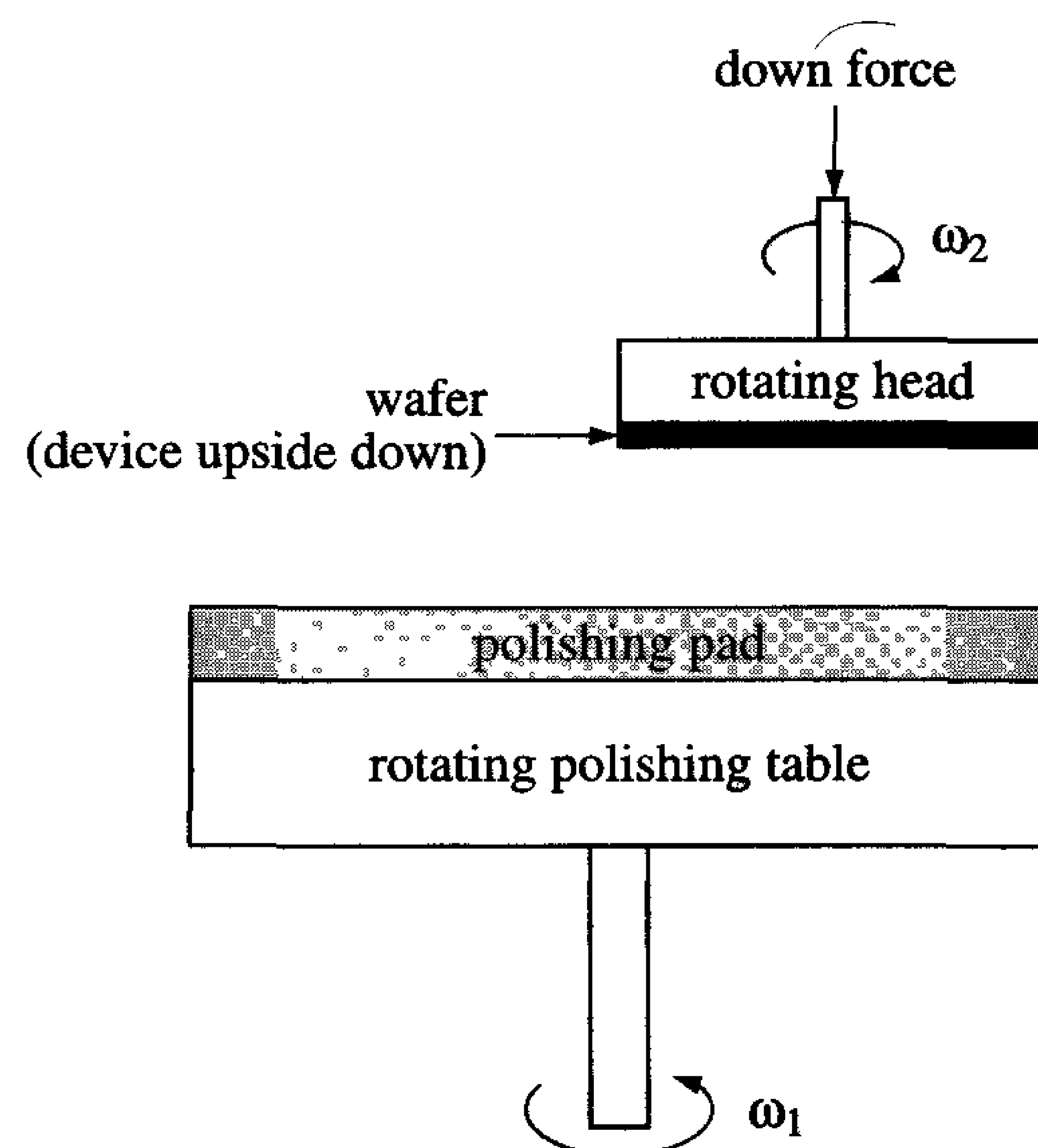


Figure 3.11: *Schematic overview of the CMP polishing process*

The *slurry* contains polishing particles (e.g. silica or alumina) and an etching substance (KOH or NH_4OH (ammonia)). A polishing pad together with the slurry planarises the wafer surface. Because CMP is based on a mechanical action, it is extremely well suited for the planarisation of rough areas, i.e. after the creation and oxide filling of trenches (STI; section 3.8.3) and during the metallisation (back-end) part of a multi-layer metal process.

Figure 3.12 shows the use of CMP in combination with STI formation.

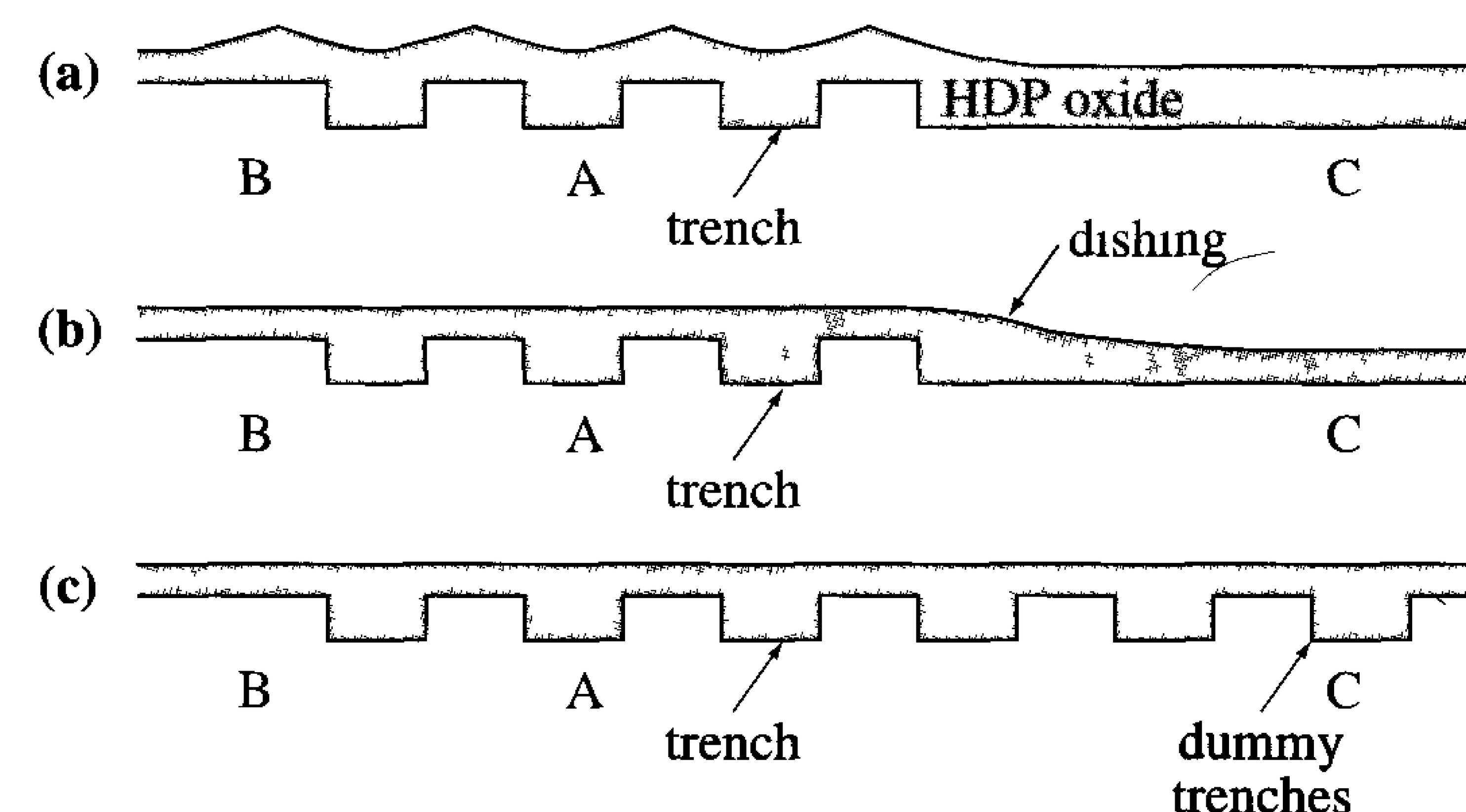


Figure 3.12: *The use of CMP in combination with STI formation: (a) after filling of the trenches, before CMP. (b) after CMP (c) after CMP and with the use of dummy trenches in open areas and dry etching of the large closed areas.*

Figure 3.12(a) shows the roughness of the surface after filling the trenches with High-Density Plasma (HDP) oxide. This process uses ion sputter etching simultaneously with the deposition of oxide. This results in a very good distribution of the deposited material, particularly at tight spaces, which are difficult to fill. Figure 3.12(b) shows the situation after the use of CMP.

However, if we look at the global surface, we see differences in the height of the surface, caused by *dishing*, for example. Large open field-isolation areas will therefore be 'filled' with dummy trenches, see figure 3.12(c). An inverse ACTIVE mask is used to etch large closed (active) areas, e.g. areas that contain large transistors. These areas are dry-etched separately, because CMP 'etches' closed areas more than areas with a high density of trenches (A). In this way, a global and local high quality planarisation can be obtained.

The use of CMP for the back-end metallisation process is much simpler than in combination with STI formation. As shown before, the CMP technique has limitations with respect to global planarisation. An important parameter is the planarisation length (L_{plan}), see figure 3.13.

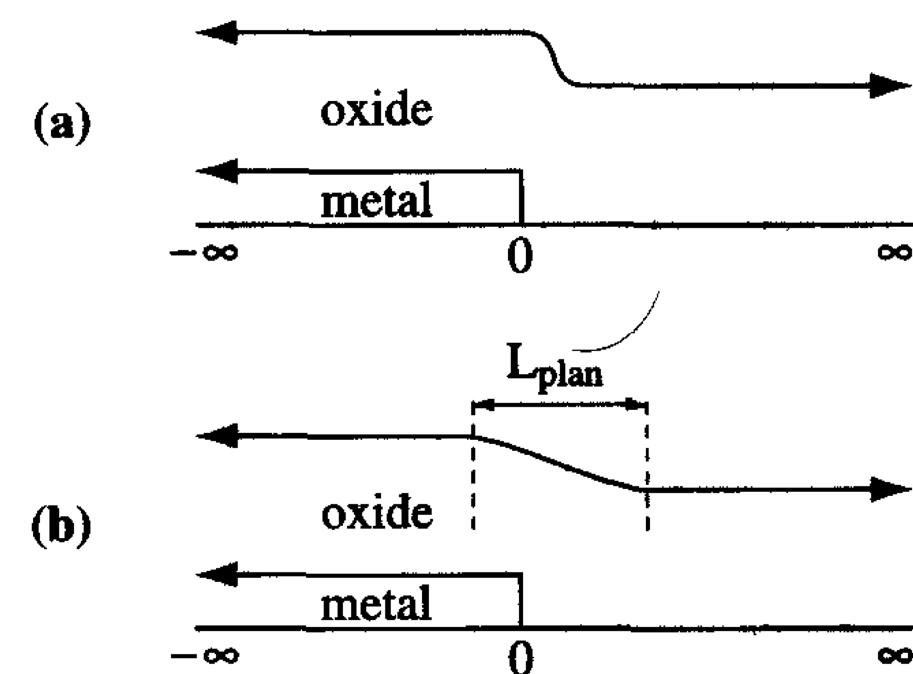


Figure 3.13: *After CMP, the original metal step is spread over a planarisation length L_{plan}*

Figure 3.13(a) shows the situation after oxide deposition on top of a certain metal pattern which occupies only half of a large area. After polishing, the remaining oxide thickness on top of the metal is equal to that at the non-metal area, except for a certain planarisation length L_{plan}, see figure 3.13(b), in which we have a smooth step. L_{plan} is determined by the polishing pad stiffness and the mechanical pressure. It is difficult to influence this parameter, which typically measures 2 to 3 mm.

Now consider the same metal pattern as in figure 3.13, but with an additional isolated metal track at more than roughly half the planarisation length. After oxide deposition, the cross-section looks like figure 3.14(a). CMP removes the thin oxide layer on top of the isolated track very quickly. As a result, the remaining oxide on top of this isolated track is much less than on the large metal area, see figure 3.14(b). This will cause a problem in multi-metal layer technologies.

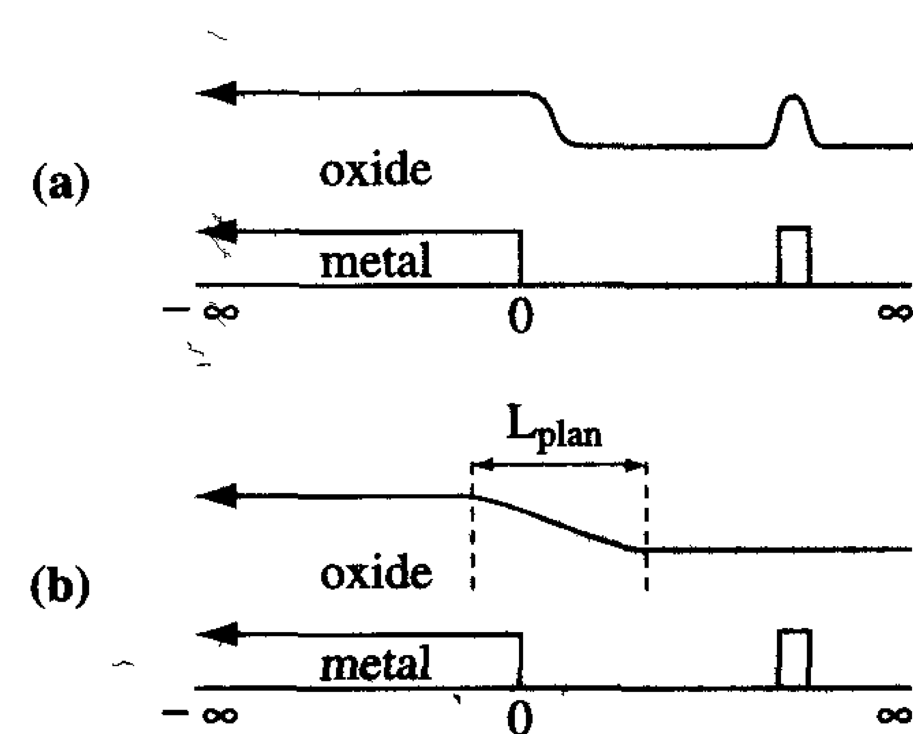


Figure 3.14: *After CMP, the oxide thickness on top of an isolated metal track is much less than on top of a densely filled metal pattern*

If several of such isolated tracks are stacked in successive metal masks, the isolated pattern will be out of focus during exposure. Figure 3.15 shows this effect. This will particularly be a problem in (deep-)submicron processes, which typically contain five or more metal layers.

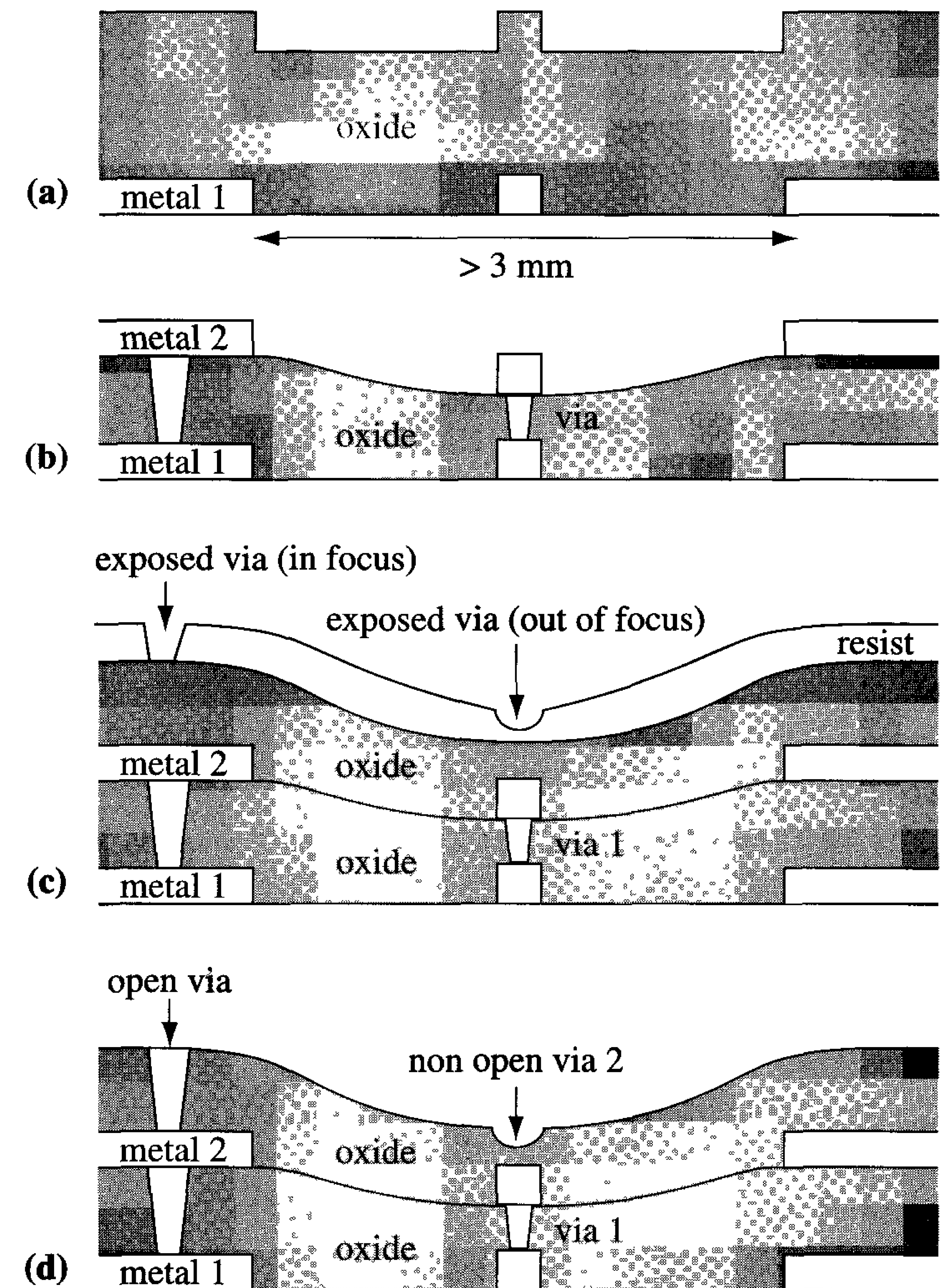
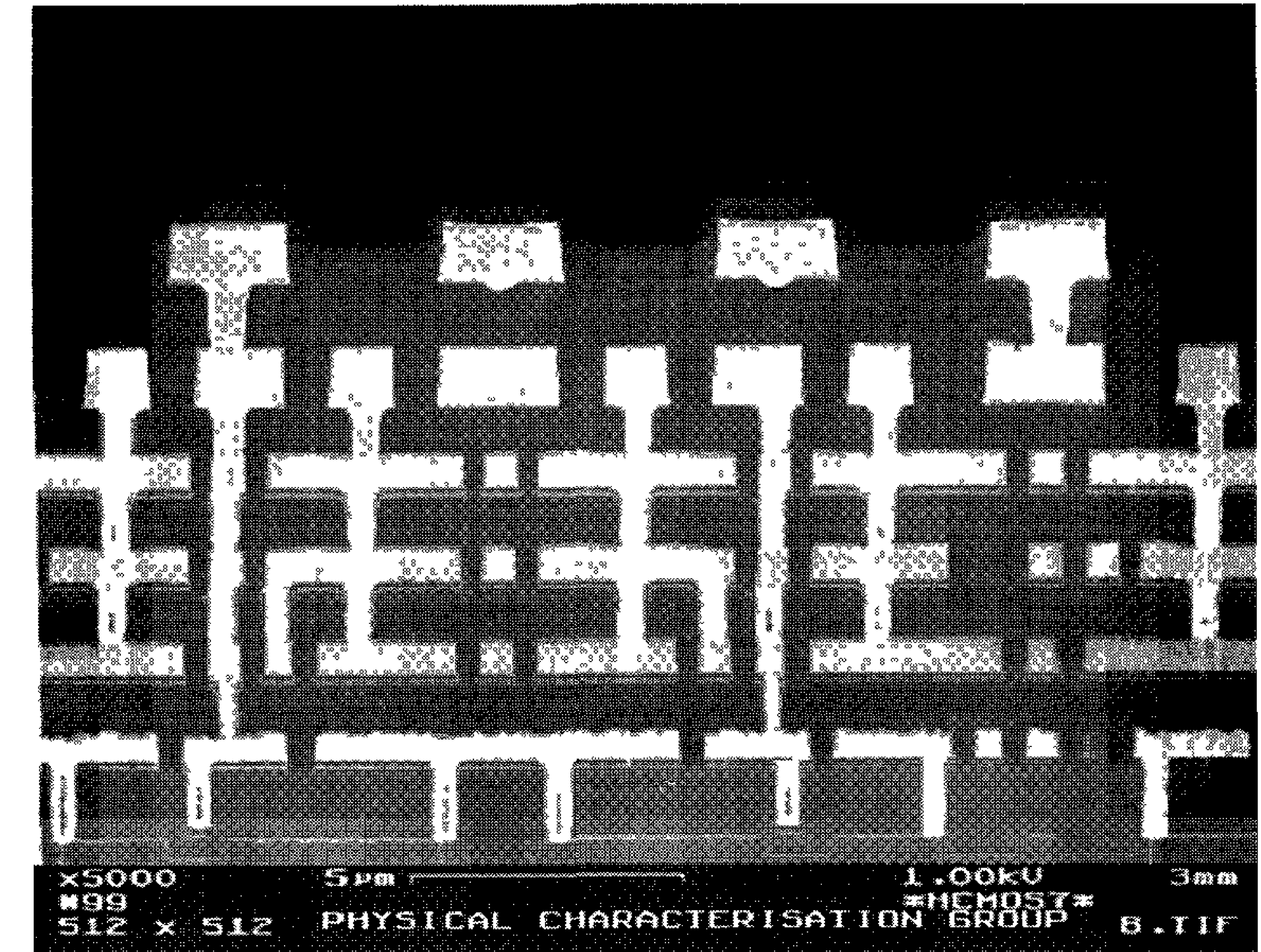


Figure 3.15: *(a) Oxide deposition after METAL1 patterning. (b) After oxide CMP, VIA1 etch and fill and METAL2 patterning. The length scale of the open area is much larger than 3 mm. (c) When topography differences accumulate to unacceptable values, vias are not well defined in the resist at all locations on the wafer because of out-of-focus situations. (d) These badly exposed vias will cause failing electrical contacts after dry etch.*

Measures to prevent planarisation problems in the back-end metallisation process include the creation of dummy metal patterns in scarcely-filled areas. The idea is to create metal patterns with as uniform a density as possible. These dummy metal patterns, sometimes also called *tiles*, should be automatically defined during chip finishing. The use of tiles improves the quality of global planarisation and also results in a better charge distribution (reduced *antenna effect*) during back-end processing (deposition and etching of the successive metal layers). The shape of the individual tiles should be chosen such that it hardly affects the performance and signal integrity of a logic block.

A disadvantage of CMP is the mechanical wear of the polishing pad. As a result, the speed of polishing is reduced and, sometimes after each wafer, a diamond-brush step is performed to recondition the pad. After about 500 wafers, the polishing pad must be completely replaced by a new one. Figure 3.16 shows the result of the CMP planarisation technique in a multi-metal layer process.



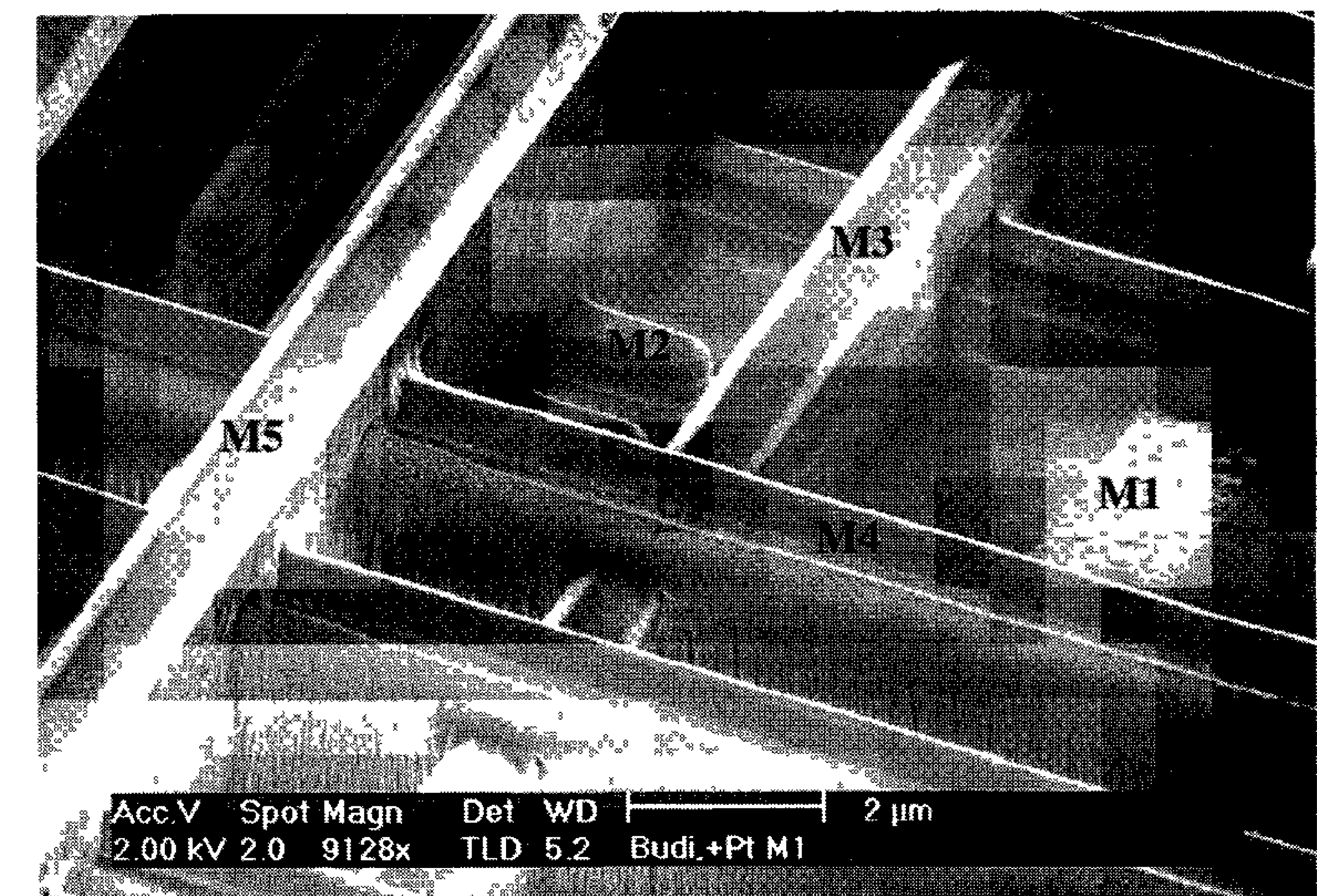


Figure 3.16: *Cross-section of a multi-metal layer process with CMP planarisation and a top view of the different metal layers in a CMP planarised chip*

3.8 Basic MOS technologies

Sections 3.2 to 3.7 illustrate that MOS processes mainly consist of several basic actions that are repeated. In modern CMOS processes, the total number of actions has increased to several hundreds.

In this section, a basic nMOS process with just five masks is discussed. A good understanding of this *silicon-gate nMOS* process enables a smooth transition to the complex modern CMOS processes. With the exception of some new steps, these CMOS processes are just an extension of the basic nMOS process presented here. A good insight into both technology types is a prerequisite when comparing the advantages and disadvantages of nMOS and CMOS.

Finally, a deep-submicron CMOS process is presented and the associated fundamentally new steps are discussed. The section is concluded with a quantitative discussion of the most dominant CMOS technologies.

3.8.1 The basic silicon-gate nMOS process

An *nMOS process* which uses a mere five masks is explained with the aid of figure 3.17. First, an oxide is grown on the base silicon wafer. Next, the oxidised silicon wafer is coated with a silicon nitride (Si_3N_4) layer, as shown in figure 3.17(a).

The first mask is the ACTIVE mask, which is used to define nitride areas corresponding to substrate regions where transistors should be formed. After the nitride is etched, boron is implanted through the resulting holes to produce the channel stopper, discussed in section 1.8 and indicated in figure 3.17(b). The wafer is then oxidised to produce the LOCOS areas in figure 3.17(c). The resulting thick oxide only exists at places that were not covered by the nitride. The channel stopper is thus automatically present everywhere beneath the LOCOS oxide. This is a great advantage of the LOCOS process. The removal of the remaining nitride reveals the areas in which transistors will be created. Now, the oxide is removed by a wet HF dip. The next step is the growth of a *thin oxide in these areas.*

The thickness of this oxide varies between 15 and 25 nm in most MOS processes. The threshold voltage adjustment implantation which follows this oxidation damages the thin oxide. The implantation is therefore done through this *sacrificial gate oxide.* Low-energy impurity atoms such as iron (Fe) and/or copper (Cu) from the ion implanter may be caught in and/or masked by the sacrificial gate oxide during the implantation.

This sacrificial gate oxide is subsequently removed and the actual gate oxide is grown. In some processes, however, impurities are implanted through the gate oxide, e.g. during a threshold voltage (correction) implant. The properties of a MOS transistor are largely determined by the gate oxide. Gate oxidation is therefore one of the most critical processing steps. Its thickness is between 3 and 10 nm (see table 3.2).

After this, a polysilicon layer of about 0.1 to 0.4 μm thickness is deposited. A subsequent phosphorus diffusion, used to dope the polysilicon, is followed by photolithographic and etching steps, which yield polysilicon of the required pattern on the wafer. The POLY mask is the second mask step in the process and is used to define the pattern in the polysilicon layer. This step corresponds to figure 3.17(d). The polysilicon is used both as MOS transistor gate material, where it lies on thin oxide, and as an interconnection layer, where it lies on thick oxide (LOCOS). The *sheet resistance* of polysilicon interconnections lies between 20 and 50 $\Omega/\square$. Polysilicon can therefore only be used for very short interconnections (inside library cells).

Phosphorus (P) or arsenic (As) are mainly used to create the source and drain areas. The sheet resistance of these areas is about the same as that of polysilicon. The edges of the n^+ areas are defined by the LOCOS and the polysilicon gate. Source and drain areas are thus not defined by a mask but are *self-aligned,* according to the location of the gate. The overlap of the gate on the source and drain areas is therefore determined by the *lateral diffusion* of the source and drain under the gate. The length of the lateral diffusion is about 60% of the diffusion depth of the drain and source.

Currently, lower doped drain extensions are used which show a lateral diffusion of about 40% of their depth, see also section 3.8.3. With a *drain extension* of 0.1 μm, the lateral diffusion is only about 0.04 μm in a 0.25 μm process. The *effective transistor channel length* is therefore equal to the polysilicon width minus twice the lateral diffusion.

The wafer is then covered with a new oxide layer, deposited by an LPCVD step. The resulting SILOX layer indicated in figure 3.17(e) is about 0.4 to 0.8 μm thick. The CONTACT mask is the third mask step in the process and is used to define contact holes in the SILOX layer, see also figure 3.17(e). The metal layer is then deposited by means of sputtering, see section 3.5. The METAL mask is the fourth mask in this sample process. It is used to define the pattern in the aluminium or tungsten layer.

Basically, the processing is now completed, see figure 3.17(f). However, as a final step, the entire wafer is covered with a plasma-nitride *passivation* layer. This *scratch-protection* layer protects the integrated circuit from external influences. Figure 3.17(f) shows the situation before deposition of the scratch protection. With a final mask step, the scratch protection is etched away in the bonding pads to be able to make wiring connections from the chip to the package. This mask and the associated processing steps are not included in the figure.

In summary, the mask sequence for the considered basic silicon-gate nMOS process is as follows:

1. ACTIVE	definition of active areas
2. POLY	polysilicon pattern definition
3. CONTACT	definition of contact holes between aluminium and monocrystalline silicon or polysilicon
4. METAL	interconnection pattern definition in aluminium.

Finally, the NITRIDE mask is used to etch openings in the nitride passivation layer, to be able to connect bonding pads with package leads.

Note: The temperatures used for the source and drain diffusion exceed 900°C. Aluminium evaporates at these temperatures. Self-aligned source/drain formation is therefore impossible in an aluminium-gate process. Molybdenum gates have also been experimented with. However, they have never been industrially applied.

The silicon-gate nMOS process has the following properties:

- Small gate-source and gate-drain overlap capacitances, caused by the self-aligned implantations.
- A relatively low number of masks, i.e. basically five to six.
- Three interconnection layers, i.e. n^+ diffusion, polysilicon and aluminium. However, intersections of n^+ and polysilicon interconnections are not possible as these result in the formation of a transistor. Chapter 4 presents a basic summary on the properties of nMOS circuits.

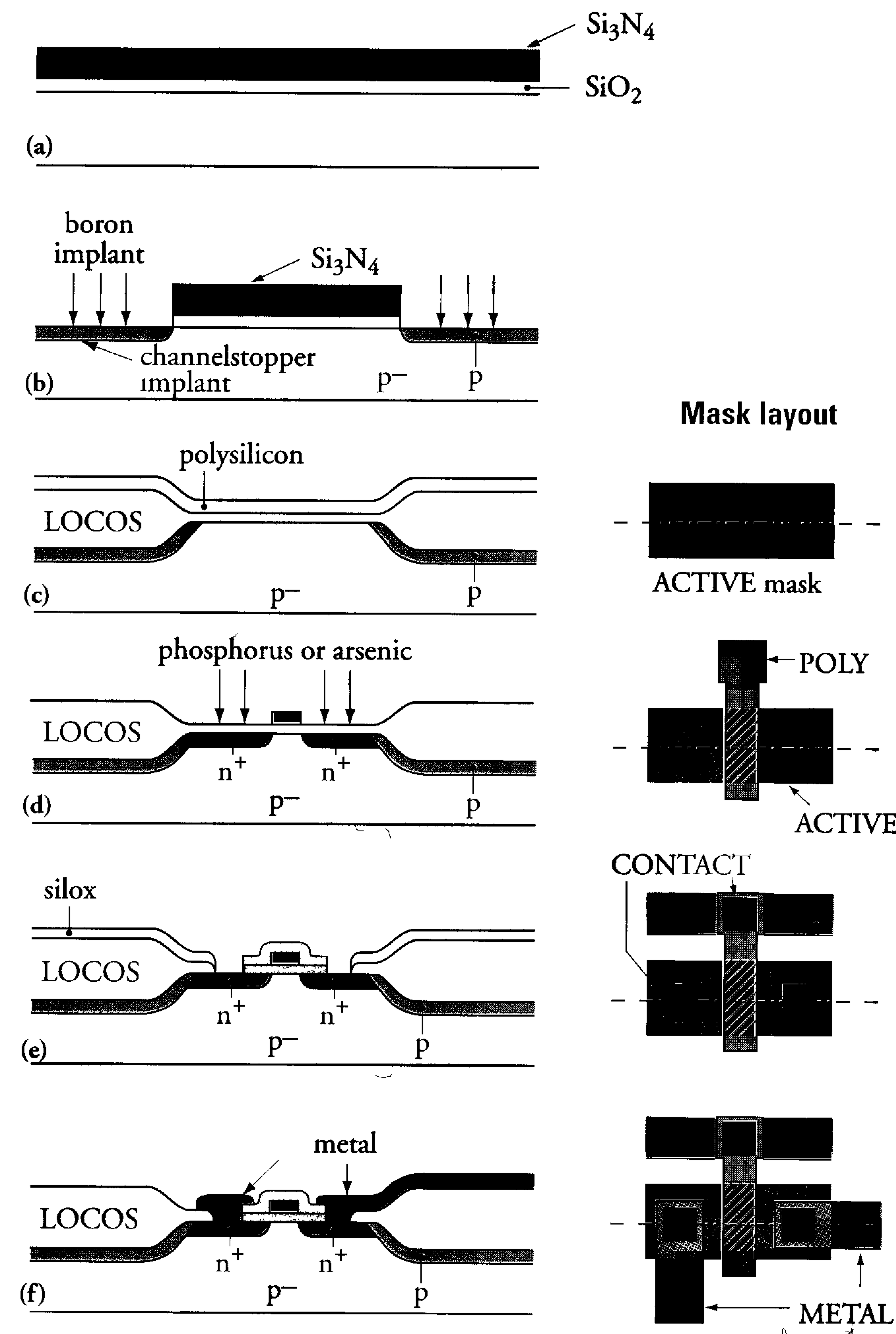


Figure 3.17: *The basic silicon-gate nMOS process*

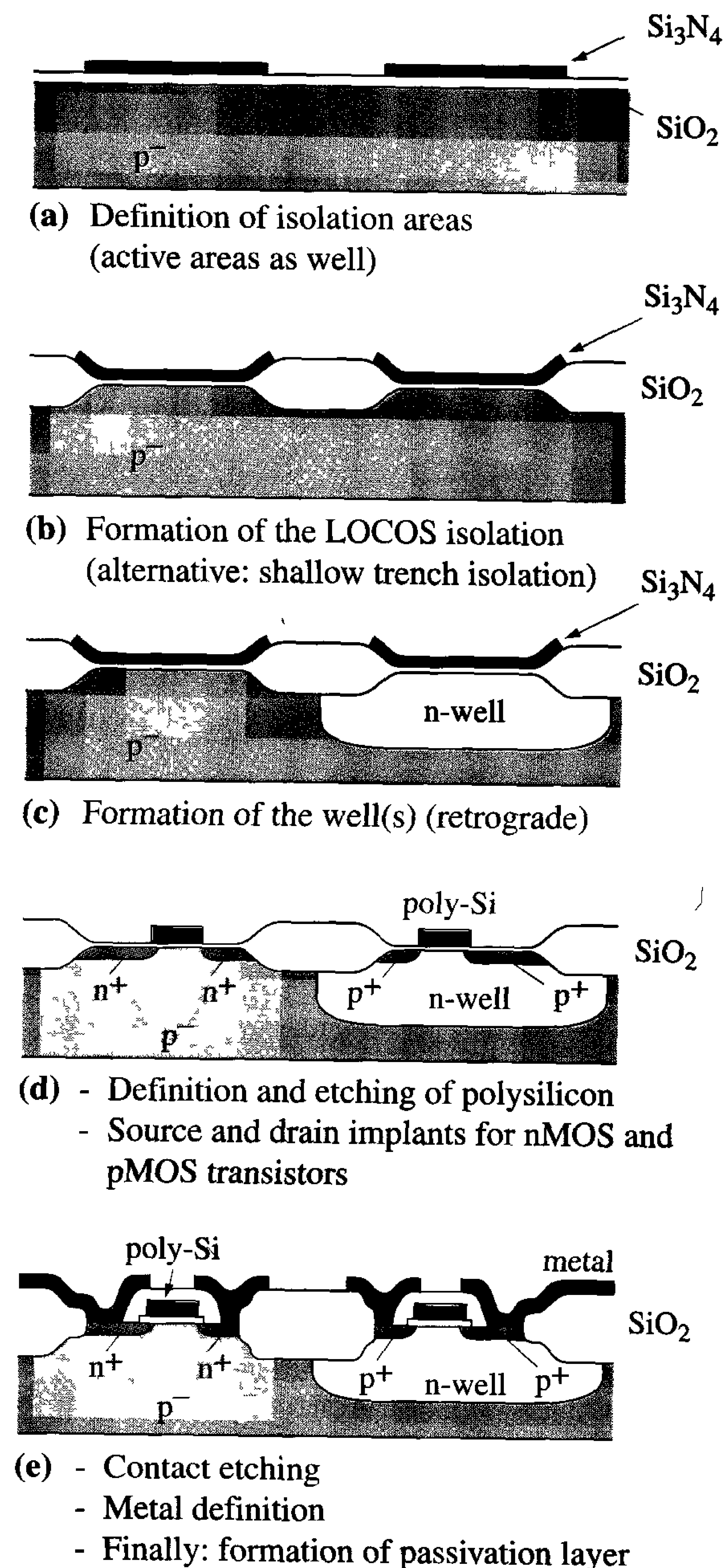


Figure 3.18: *The basic CMOS process*

3.8.2 The basic Complementary MOS (CMOS) process

CMOS circuits and technologies are more complex than their nMOS counterparts. In addition, a static CMOS circuit contains more transistors than its nMOS equivalent and occupies a larger area in the same process generation. However, CMOS circuits dissipate less power than their nMOS equivalents. This is an important consideration when circuit complexity is limited by the 1 W maximum power dissipation associated with cheap plastic IC packages. In fact, reduced dissipation is the main reason for using CMOS instead of nMOS.

Both n-type and p-type transistors are integrated in CMOS processes. Figure 3.18 illustrates the flow of a simple CMOS process with an *n-well*, or *n-tub*, in which pMOS transistors are implemented. This process serves as an example for the many existing CMOS technologies.

The basic CMOS process begins with the oxidation, to some tens of nanometres, of a monocrystalline p-type silicon wafer. A layer of silicon nitride (Si_3N_4) is then deposited on the wafer. This is followed by a photoresist layer. A mask is used to produce a pattern in the photoresist layer corresponding to *active areas*. Circuit elements will be created in these areas.

The defined pattern determines which silicon nitride remains during a subsequent etching step. The photoresist is then completely removed, as shown in figure 3.18(a). LOCOS oxide is then grown by exposing the wafer to oxygen at a high temperature. This oxide will not be grown on the exposed Si_3N_4 areas. The LOCOS oxide separates active areas, see figure 3.18(b) for an indication of the result. Instead of LOCOS, STI is used in deep-submicron processes to separate active areas (see next subsection). A new photoresist layer is then deposited and the p-type transistor areas are 'opened' during photolithographic steps. The n-well is implanted (mostly phosphorous) in these areas, as shown in figure 3.18(c). Initially, the implanted ions collect at the silicon surface but they diffuse more deeply during a subsequent high temperature step. A layer of polysilicon is then deposited on the wafer, which now consists of n-type n-well areas with a limited depth of 1-2 μm and p-type substrate areas.

Polysilicon doping reveals either n-type polysilicon for both nMOS and pMOS transistor gates, or *double-flavoured polysilicon* (n-type and p-type polysilicon for nMOS and pMOS transistor gates, respectively). This is also sometimes referred to as n^+/p^+ *dual polysilicon*.

A photolithographic step follows and the polysilicon pattern is etched.

The resulting polysilicon is used for short interconnections and for transistor gates.

Separate masks are used for the self-aligned source/drain implantations for the nMOS and pMOS transistors in the substrate and n-well, respectively. The result is shown in figure 3.18(d).

The first step in the creation of interconnections between the different transistor areas is to deposit an SiO_2 layer on the wafer. Contact holes are etched in this layer to allow connections to the gates, drains and sources of the transistors. A metal layer is then deposited, in which the final interconnect pattern is created by means of photolithographic and etching steps. Figure 3.18(e) shows the final result.

Modern CMOS processes use 20 to 30 masks. Basically, these processes are all extensions of the simple CMOS process described above. VLSI and memory processes now use channel (gate) lengths of 0.13 to 0.5 μm and offer several levels of polysilicon and/or metal. These multiple interconnection layers facilitate higher *circuit densities*. The next section discusses a state-of-the-art deep-submicron CMOS process.

3.8.3 An advanced deep-submicron CMOS process

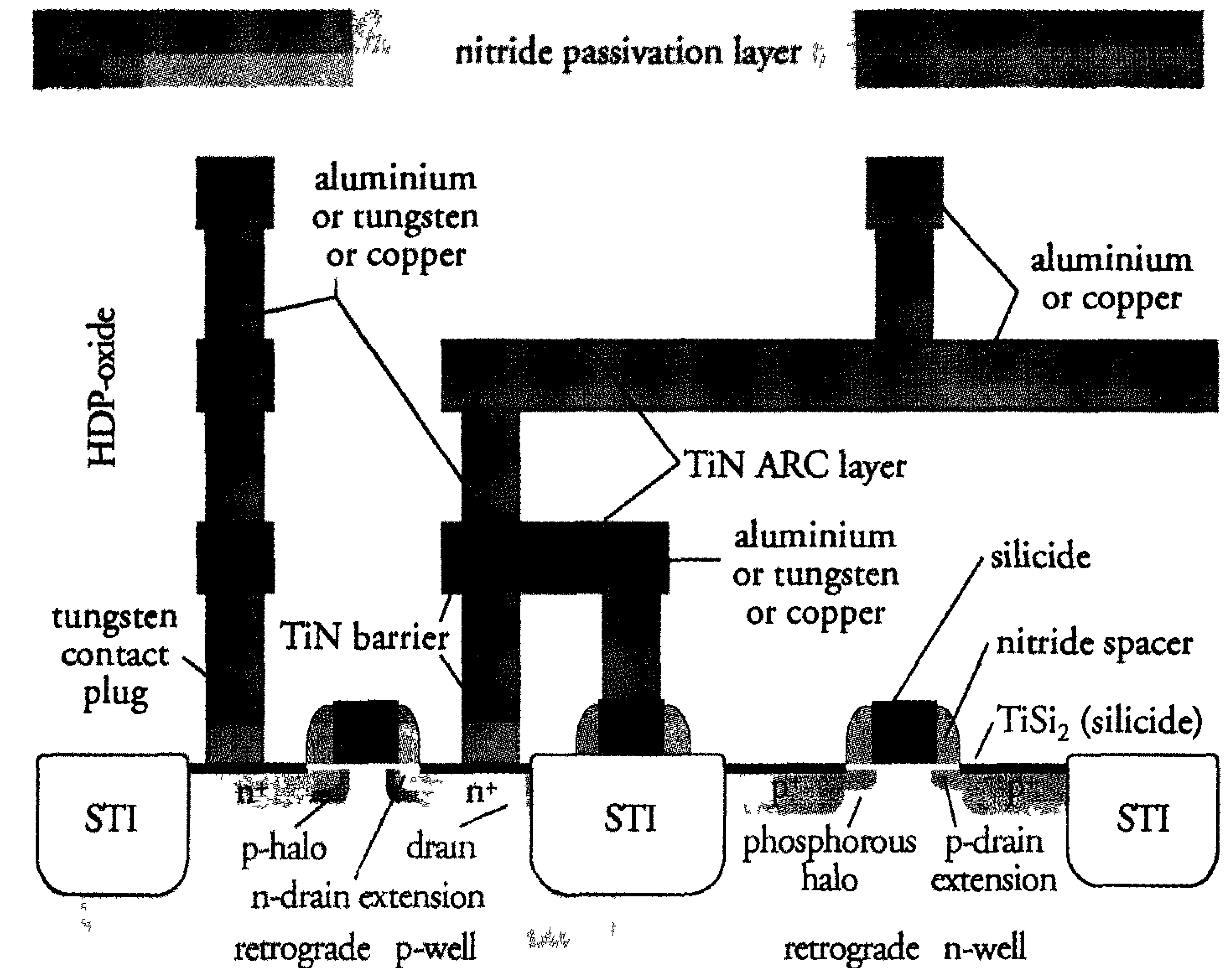


Figure 3.19: *An advanced deep-submicron process*

Compared to the basic CMOS process discussed before, an advanced deep-submicron CMOS process, with channel lengths of 0.25 μm and below, incorporates several major different processing steps. These differences will now be discussed in some detail.

Shallow-trench isolation

Actually, LOCOS is thick SiO_2 that is thermally grown between the active areas. In contrast, *Shallow-Trench Isolation (STI)* is implemented at significantly lower temperatures, preventing any warpage and stress problems associated with a high-temperature step. The STI process starts with a thermally-grown oxide with a thickness between 10 nm to 14 nm. This is followed by an LPCVD deposition of 100 nm to 160 nm

nitride. Next, the active areas are masked and a dry etch step is applied to create the trenches, which have a typical depth between 300 nm and 500 nm. The corners at the bottom and the top of the trench are rounded by a thermally-grown oxide layer (between 20 nm and 50 nm) along the side walls of the trench, see figure 3.20.

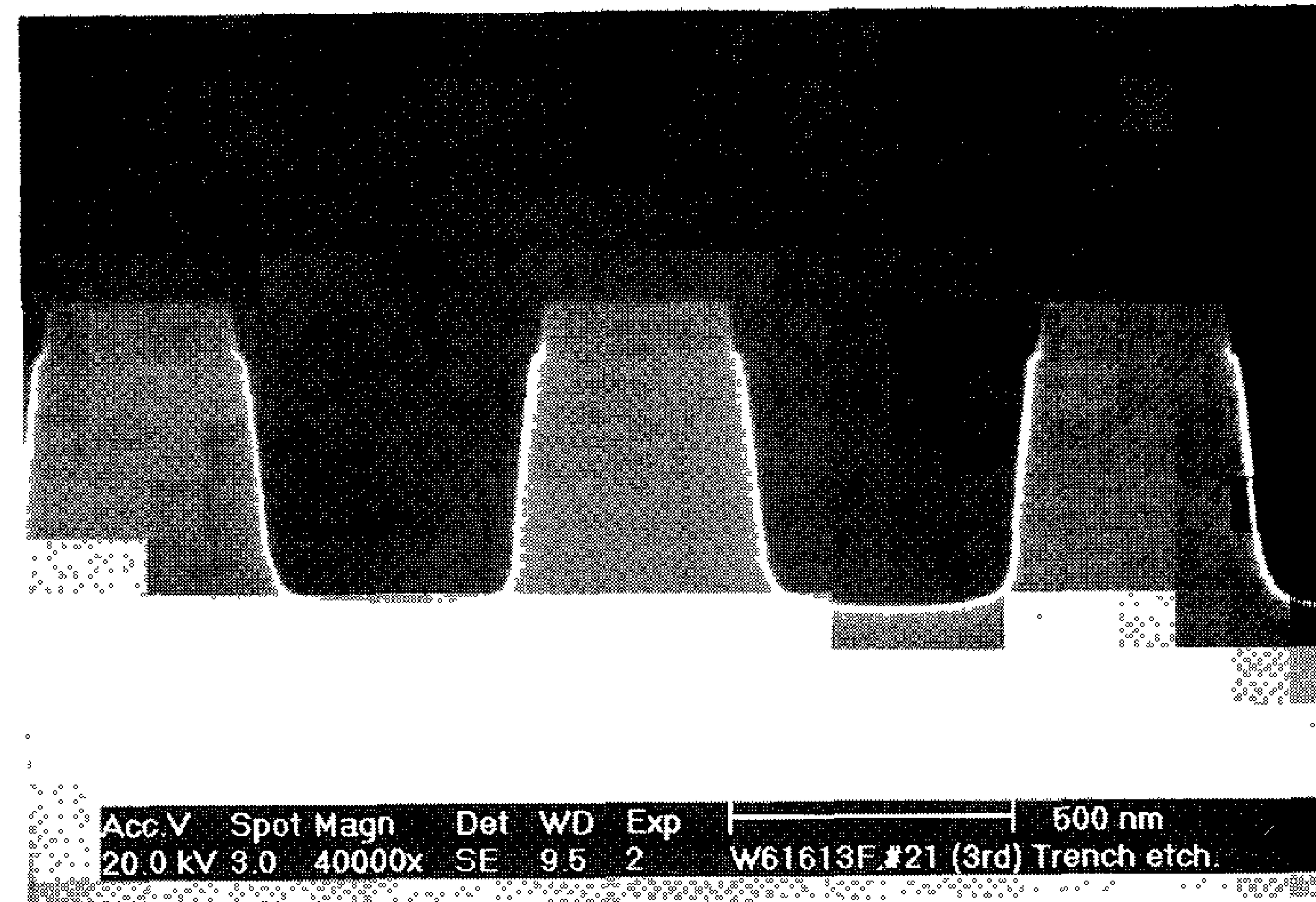


Figure 3.20: *Cross-section after etching the trenches in the silicon*

After removing the resist, a thick oxide High-Density Plasma (HDP), typically 700 nm to 1100 nm, is deposited. HDP is capable of filling the high aspect ratio of the trenches, which includes the pad oxide and nitride layer thicknesses. As shown in figure 3.21, the step coverage of the oxide is dependent on the geometry of the active area mask.

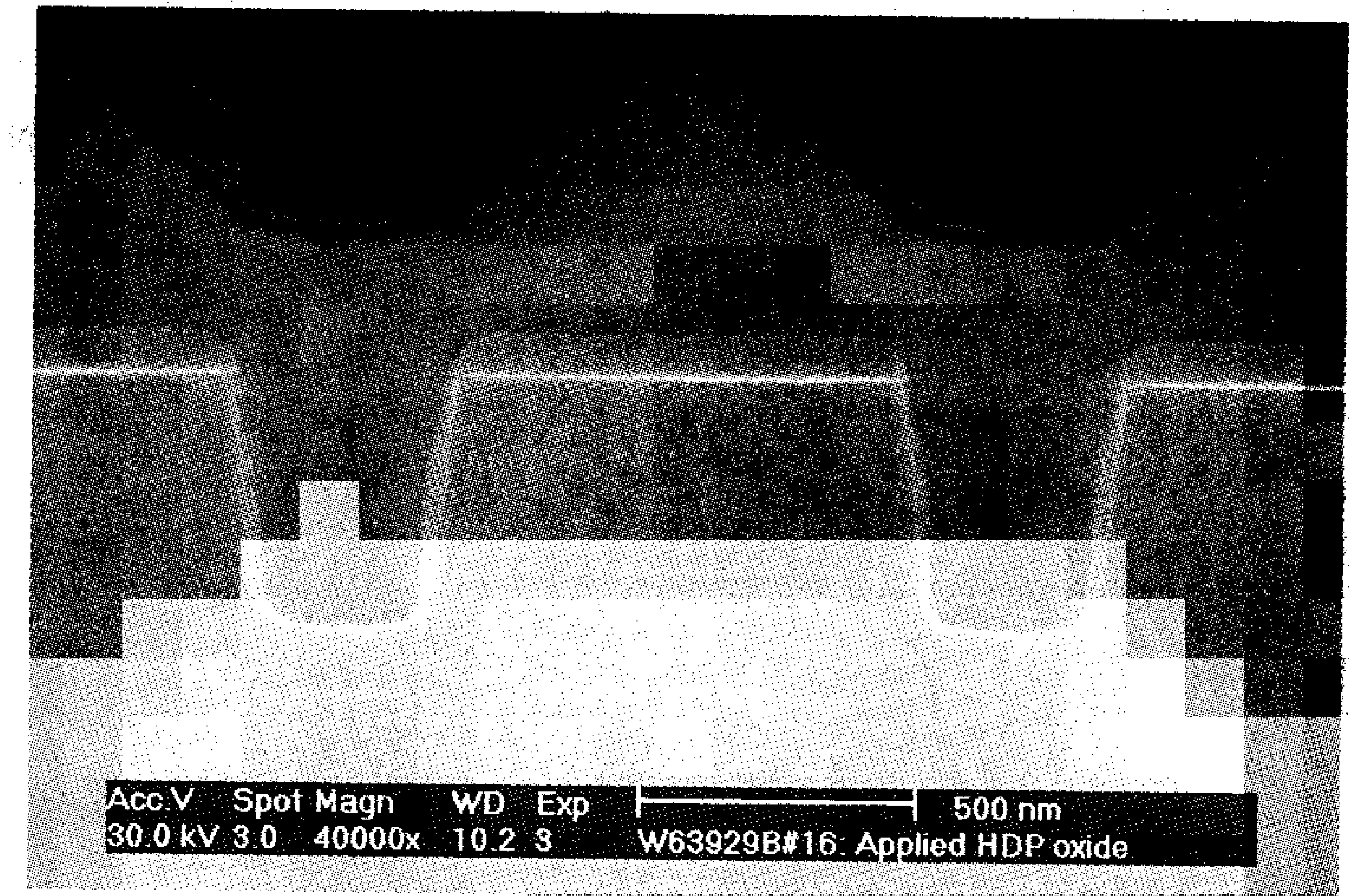


Figure 3.21: *STI process cross-section after thick oxide deposition*

In dense areas, the oxide level is well above the silicon nitride, while the oxide thickness equals the deposited oxide thickness in large open areas. The remaining topology is planarised using CMP, see section 3.7. The nitride layer is used for end-point detection, see figure 3.22.

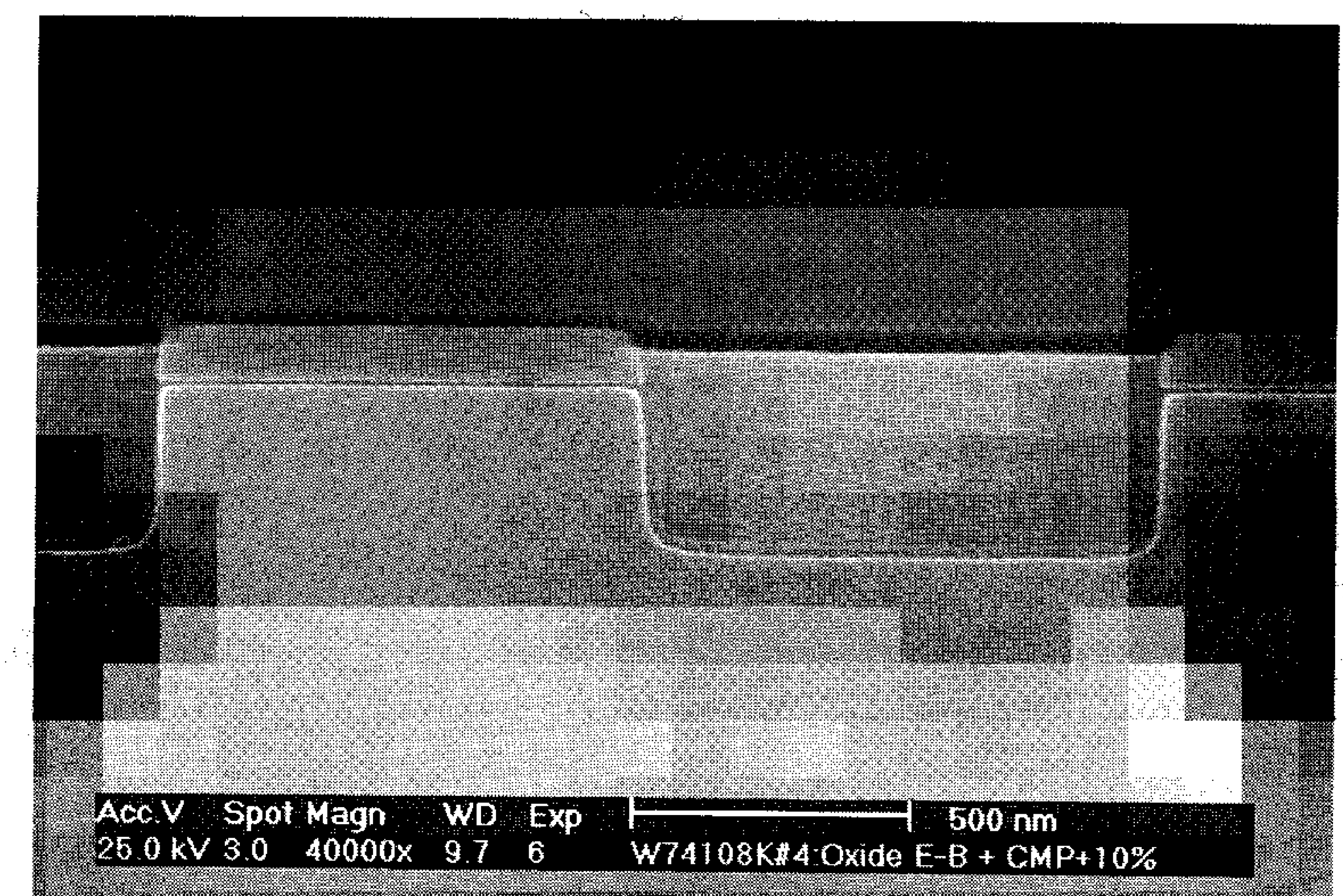


Figure 3.22: *SEM cross-section after CMP*

Next, the nitride masking layer is removed, using a wet etch and subsequently sacrificial oxide, and gate oxide is grown, polysilicon is deposited, etc. Figure 3.23 shows a cross-section through the width of the device. The thermal gate oxide between the polysilicon layer and the monocrystalline silicon substrate has a thickness of 4 nm.

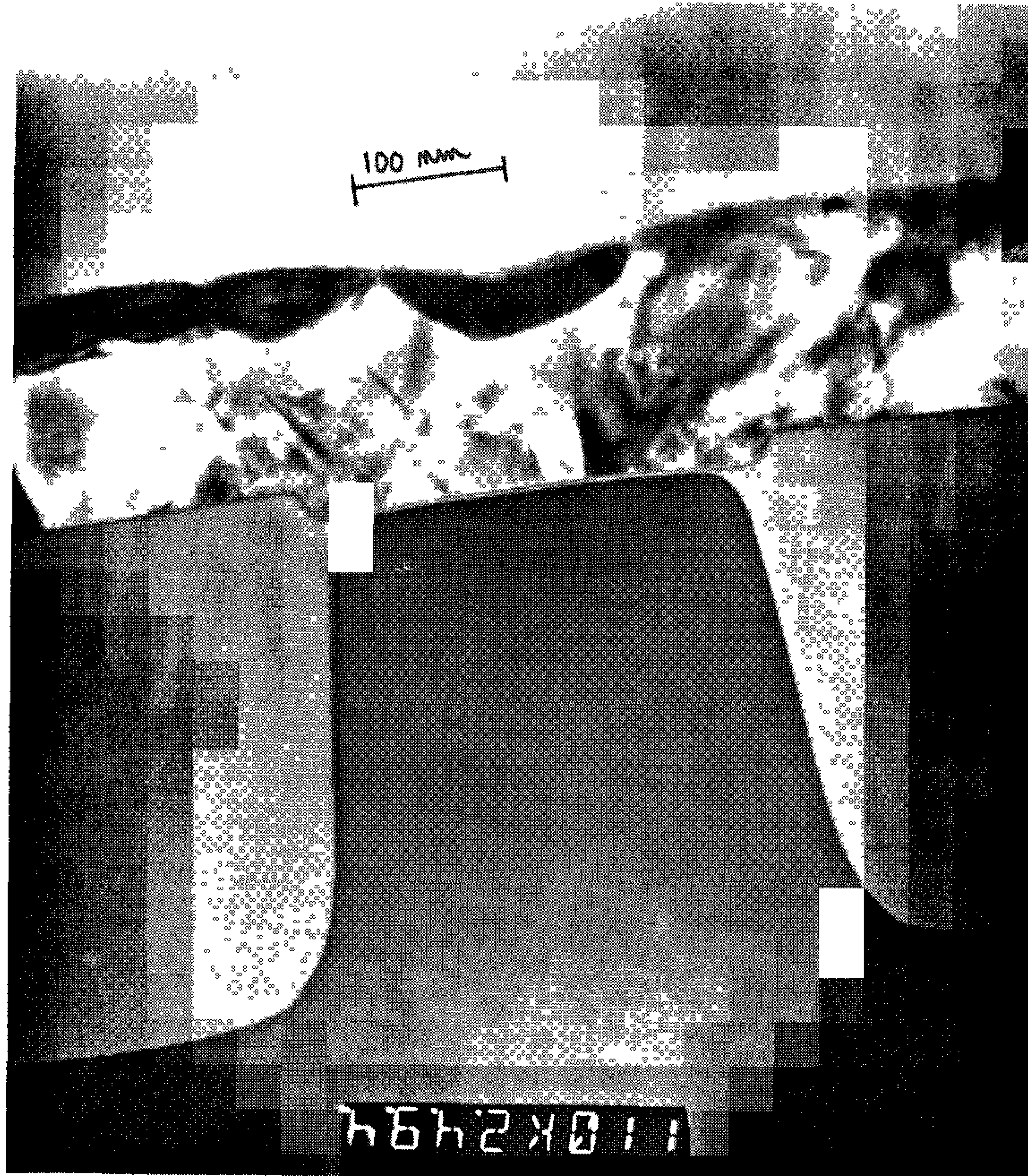


Figure 3.23: *TEM cross-section through the width of the device*

In this way, device widths of 0.2 μm are well defined. Figure 3.24 shows a comparison between LOCOS and STI field isolation techniques. It is clear that the STI is much more accurately defined and better suited for deep-submicron field oxide definition.

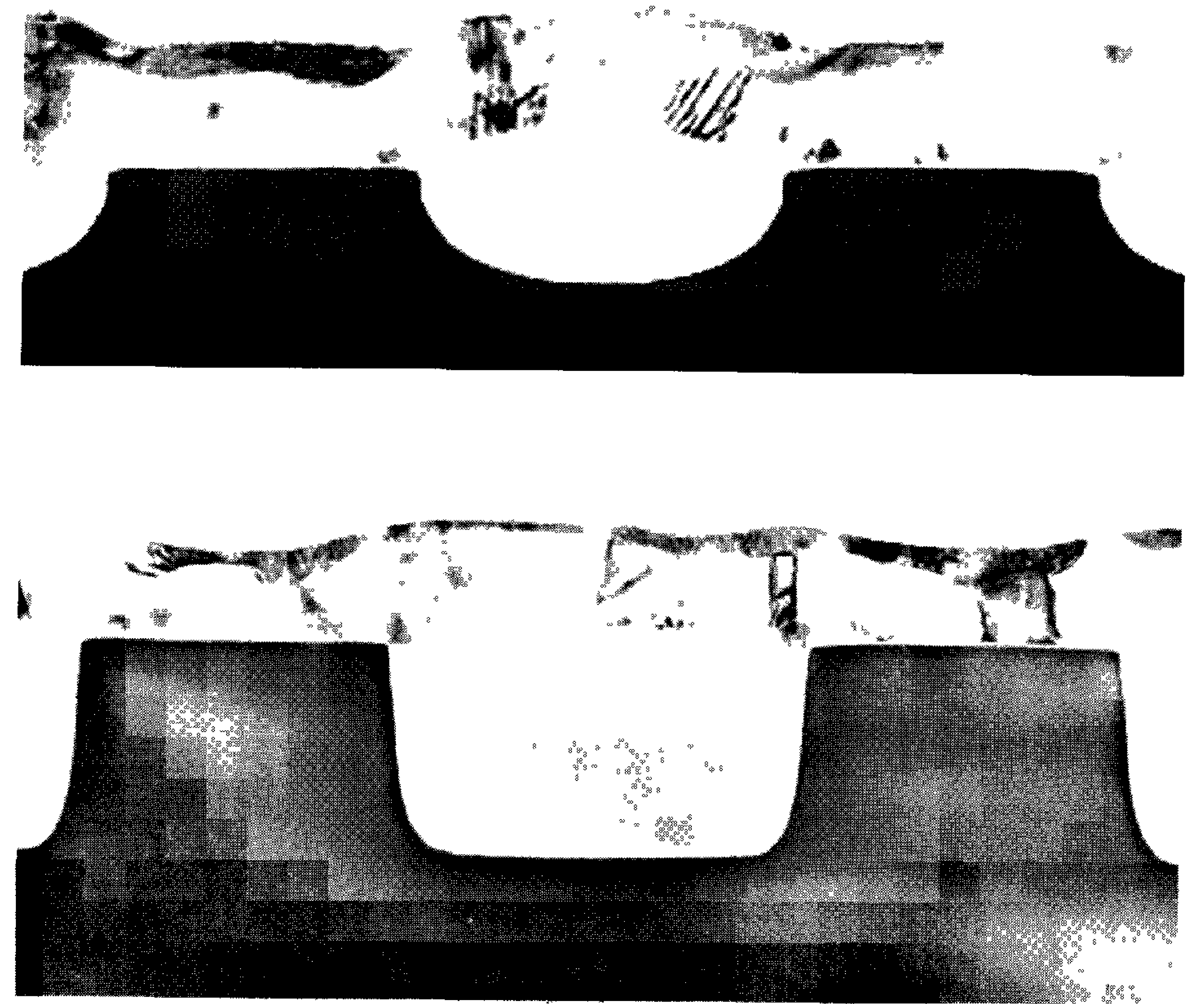


Figure 3.24: *Comparison between LOCOS (top) and STI field isolation (bottom) techniques*

Retrograde-well formation

A *retrograde-well* process (figure 3.19) uses both n-wells and p-wells, and is also called a twin-well process. These wells form the substrate for p-type and n-type devices, respectively. High-energy implantation of the wells yields doping profiles with maxima at about 0.6 μm beneath the wafer surface in active areas. The maximum dope level beneath thick

oxide areas (STI areas) is only a short distance below the bottom of these oxides. The implantation therefore acts as a very effective *channel stopper* for parasitic devices in these areas.

Only a limited temperature is required to drive the well implants to appropriate depths, which results in limited lateral diffusion. Consequently, the wells can be accurately defined and their separation from source and drain areas of their own type (e.g. n-well to n^+ source/drain regions and p-well to p^+ source/drain regions) can be relatively small. This is the most important reason for applying retrograde-well processing.

Each well can be optimised to yield the highest performance for both types of transistors. This can be done by minimising source/drain junction capacitances and body effect or by using an *'anti-punch-through'* (APT) implant. Another advantage is the associated feasible symmetrical electrical behaviour. In addition, the two wells are basically each other's complement and can be formed by defining only a single mask during the design. Finally, another significant advantage of twin-well CMOS processes is formed by the better scaling properties, which facilitate the rapid transfer of a design from one process generation to another. Scaling is extensively discussed in chapter 11.

Optimizing technologies for high-speed digital designs degrades analogue circuit performance of long-channel devices. Careful optimisation of the front-end process (including the wells) is required to improve mixed analogue/digital circuit performance [11].

Drain extension

The *hot-carrier effect*, as discussed in chapter 2, only manifests itself when carriers acquire more kinetic energy than about 3.5 eV. In 2.5 V processes and below, the hot-carrier effect is almost impossible (energy equals $q \cdot V = 2.5\,\text{eV}$ in a 2.5 V process). Carriers can only acquire such energies after a lot of collisions in the pinch-off region. As the pinch-off regions become very narrow for deep-submicron technologies, this is very unlikely to happen.

The LDD (chapter 2) implants, as used in processes of 0,35 μm and larger, are thus replaced by a more highly doped source/drain extension. This source and drain extension is produced in the same way as the LDD. However, the dope ($\approx 5 \cdot 10^{18}/\text{cm}^3$) is about a factor ten higher than usually applied in an LDD, and results in a lower series resistance. This source/drain extension implant is less deep than the actual source/drain

junctions, which allows a better control of the channel length and reduces the short-channel effects. Actually, such an extension acts as a hard mini-drain. In some cases in literature, only one implant is used to create the drain. This is then without extension implant, and called *Highly-Doped Drain (HDD)*. The phosphorous halo with increased dope in the channel around the drain, reduces the depletion layer thickness and suppresses short-channel effects such as punch-through.

Silicides, polycides and salicides

Silicides may be formed by the use of $TiSi_2$, WSi_2, $CoSi_2$ or other metal silicides. When, for example, a titanium film is deposited directly on a silicon surface, after the definition of the polysilicon and the formation of the source/drain junctions, the titanium and the silicon react to form a silicide layer during a subsequent heating step. Titanium (and some other metals) react with exposed polysilicon (resulting in *polycide*) and source/drain regions to form $TiSi_2$ silicide. A layer of titanium nitride (TiN) is formed simultaneously on the silicon dioxide. This will be selectively etched away. Silicidation yields low-ohmic silicide top layers in polysilicon and source/drain regions to reduce RC delays by about a factor five and improve circuit performance. Because the silicidation step is maskless, it is also called *self-aligned silicide* or *salicide*.

Ti/TiN film

Titanium (Ti) is used in the contact holes to remove oxides and to create a better contact with the underlying silicide. A *titanium nitride* (TiN) film is used in the contacts, as well as on the top of the PETEOS, because of its good adhesive properties. When the tungsten is being etched away with a plasma, TiN is used as an etch stop. The TiN is also responsible for an increased resistance of the contact plugs.

Anti-Reflective Coating (ARC)

Reflections during exposure of a metal mask may cause local narrowing in the resist pattern and, consequently, in the underlying metal pattern, which is to be defined. A titanium nitride film is often deposited on top of the metal layer and serves as an *Anti-Reflective Coating (ARC)*. This film is highly absorbent at the exposure wavelength. It absorbs most ($\approx$ 75%) of the radiation that penetrates the resist. It also suppresses scattering from topographical features.

Contact (re)fill

In many processes, particularly those which include planarisation steps, oxide thickness may vary significantly. Deep contact holes with high aspect ratios require special techniques to guarantee good filling of such contacts. This *contact filling* is often done by tungsten, called (tungsten) plugs, pillars or studs.

Damascene metal patterning

In most current processes, metal patterning is done by depositing a metal layer, followed by a dry etching step to etch the metal away according to a mask pattern. In the damascene process, metal patterns are created by etching trenches in the dielectric, overfilling these trenches with metal and then polishing the overfill away using CMP, until the polishing pad lands on the dielectric, see figure 3.25.

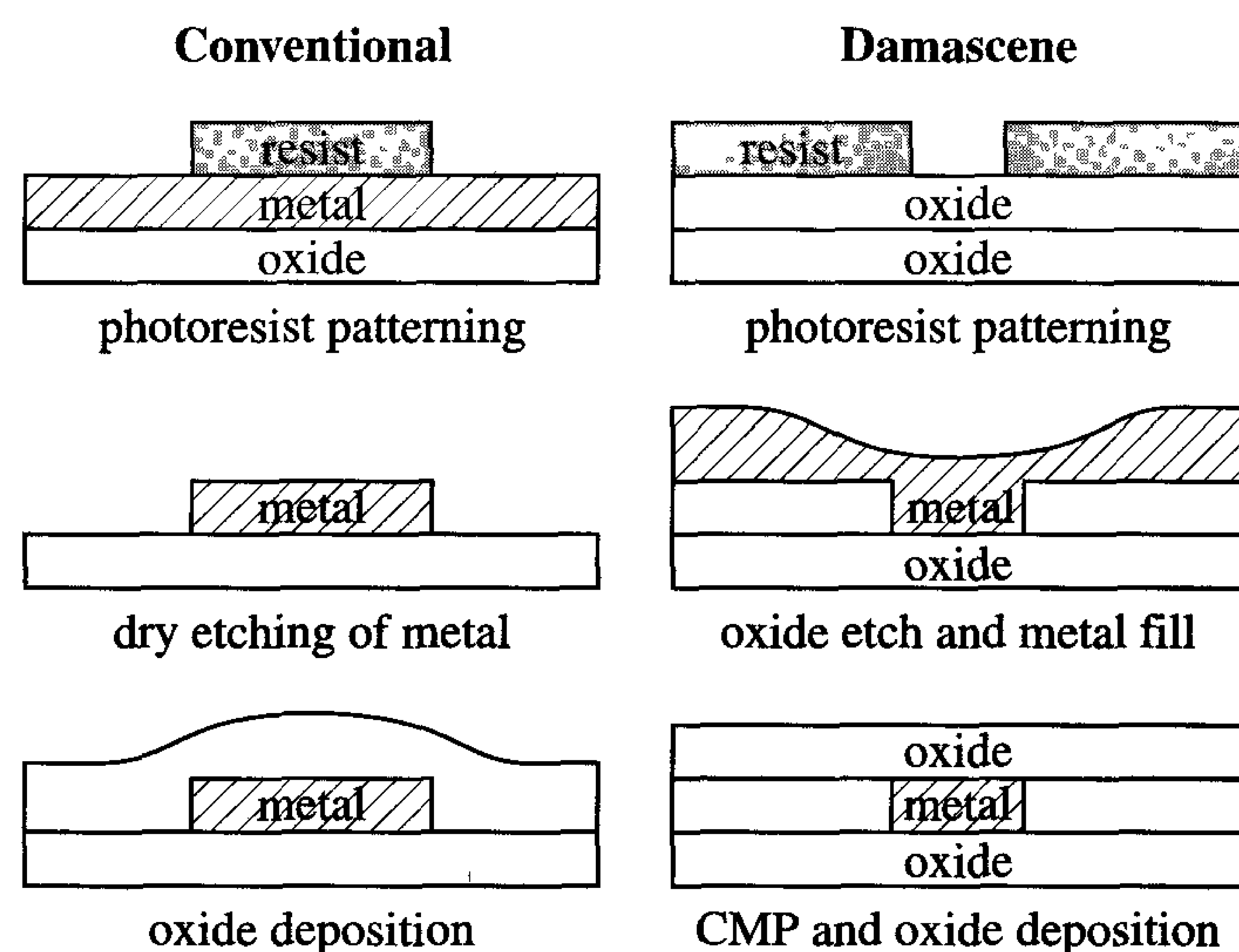


Figure 3.25: *Comparison of conventional and damascene metal patterning*

Currently, *damascene patterning* is also used, particularly in 0.18 μm and below, to form tungsten wires. In a *dual-damascene* process, plugs (studs, pillars) and wires are deposited simultaneously. This process replaces the deposition of the plug and its etching, thereby reducing processing costs. The damascene process can also be used to pattern

copper, which currently cannot be etched like aluminium in plasma reactors. The copper will create too many by-products which remain on the surface and cannot be removed. In the damascene process, the overfill of trenches with copper can be polished with CMP such that copper only remains in the wire trenches, see chapter 11. The use of copper instead of aluminium for interconnection results in a reduction of the interconnection resistivity by 25 to 30%. In combination with the use of low-ϵ dielectrics, the speed can be doubled and the power can be halved. Copper can also withstand higher current densities (reduced chance of electromigration, see also section 11.3).

3.8.4 Silicon-on-insulator CMOS (SOI-CMOS) process

The previously-discussed CMOS processes show relatively large source/ drain capacitances. This can be avoided with the *SOI-CMOS* process illustrated in figure 3.26. The complete isolation of nMOS and pMOS transistors associated with this process removes the possibility of latch-up.

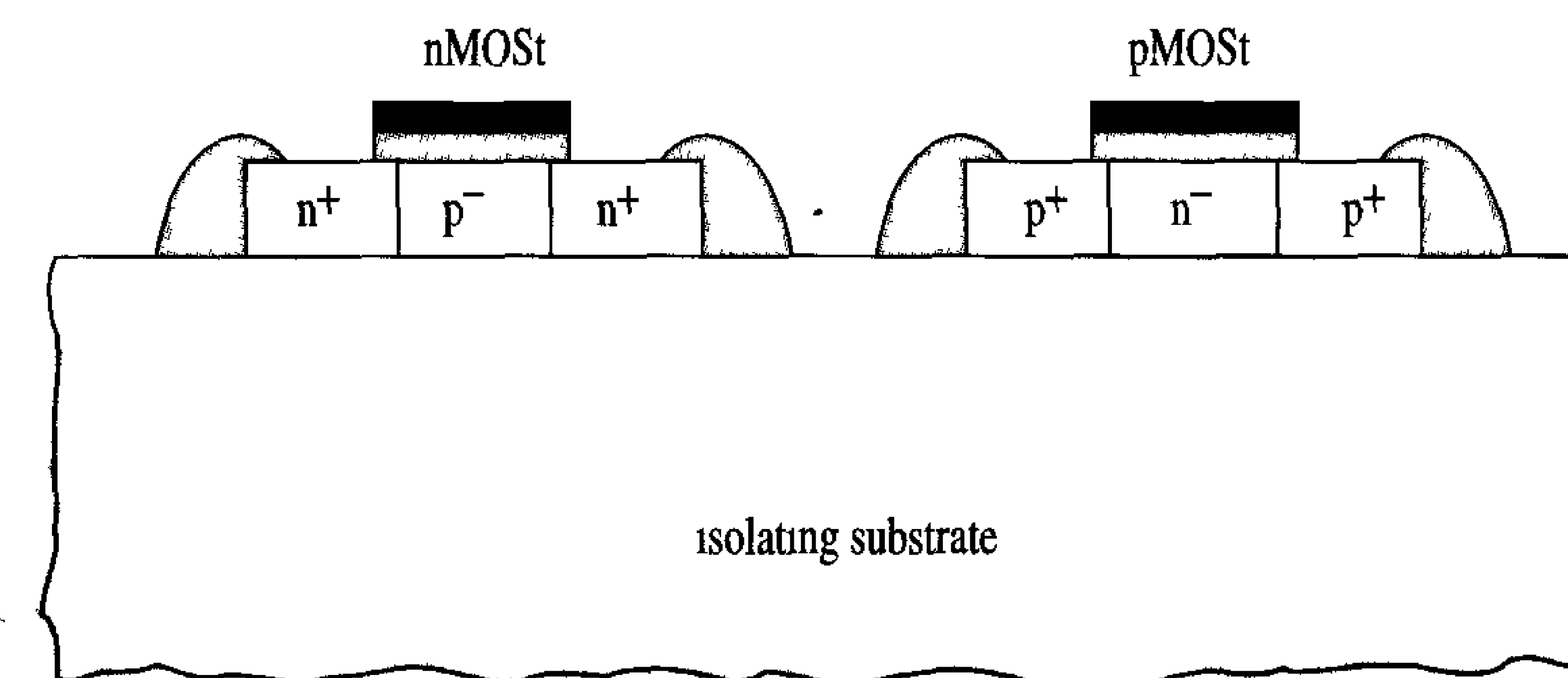


Figure 3.26: *Cross-section of a basic SOI-CMOS process*

Neither the nMOS nor pMOS transistor channels require over-compensating impurity dopes. Very small body effects and source/drain capacitances are therefore possible for both types of transistor. In addition, the n^+ and p^+ diffusion regions do not have bottom junctions. Consequently, the parasitic capacitances are much less than those of the previously-discussed CMOS processes. This makes the SOI-CMOS process particularly suitable for high-speed and/or low-power circuits. Murphy's law, however, ensures that there are also several disadvantages associated with SOI-CMOS processes. The absence of substrate diodes,

for example, complicates the protection of inputs and outputs against the *ESD* pulses discussed in chapter 9.

Sapphire is used as the isolating substrate in SOI-CMOS processes, despite the fact that it is substantially more expensive than silicon. The *SIMOX* (*'Separation by IMplantation of OXygen'*) process provides a cheap alternative for these *silicon-on-sapphire* or 'SOS-CMOS' processes. Several modern SOI-CMOS processes are based on SIMOX. These processes use a retrograde implantation of oxide atoms to obtain a highly concentrated oxygen layer beneath the surface of a bare silicon wafer. The resulting damage to the wafer's crystalline structure is corrected in an annealing step. The result is shown in figure 3.27.

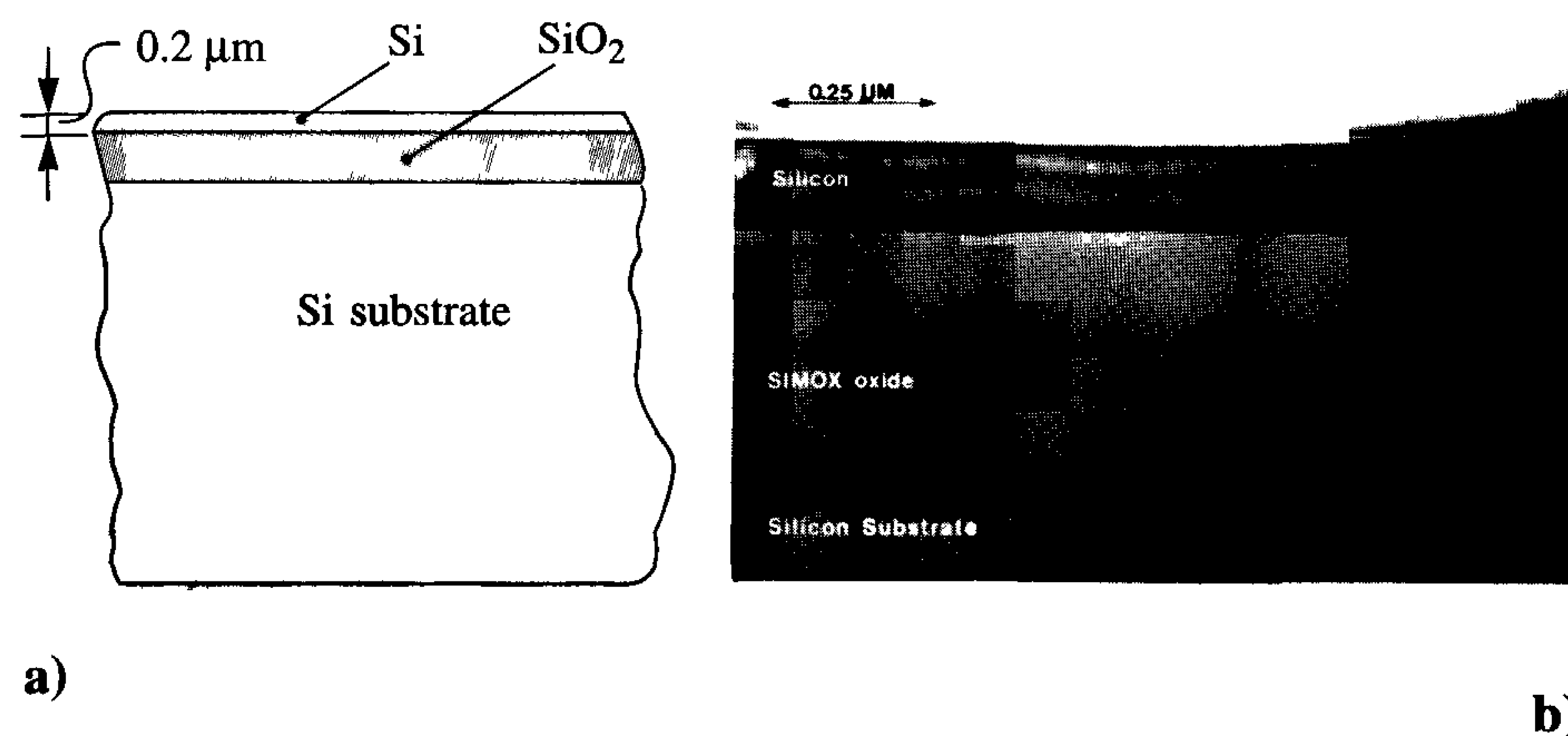


Figure 3.27: *(a) Cross-section of a SIMOX wafer and (b) SEM photograph of such a cross-section*

SIMOX wafers are delivered with an SiO_2 layer between 300 and 400 nm thickness, which lies at least 100 to 200 nm below the wafer surface. This is done to reduce the consequences of damage on the wafer surface. Fully depleted devices can be realised by reducing the thickness of the top layer to 0.1 μm, for example, during processing.

A general disadvantage of SIMOX (or SOI) is that the electrical isolation of the active channel region is isolated from the main body. This leads to 'floating body' effects, in which the body potential varies with the transistor voltages. If the body of each transistor must be connected to the same electrical node, the required area per transistor would dramatically increase. If the SOI process is properly controlled, the floating body effects can be reduced. Several SOI devices have been presented

in literature [12]. However, SOI technology is still not very mature because the yield on wafers in the same lot can vary by a factor of 3 to 4. If key problems are solved, SOI can become an important alternative to bulk silicon technology. The semiconductor industry recently shows a growing interest in SOI [13], since it may have advantages over bulk CMOS. For advanced low-voltage CMOS ($\leq$1 V) system-on-chip designs with digital, analogue and RF parts, SOI is expected to offer a better performance than bulk CMOS technology [14,15]. SOI is said to deliver more speed at the same power consumption, or to consume less power at the same speed. Furthermore, SOI realises better isolation between digital, analogue and RF parts on the IC. Those circuits will therefore be less affected by substrate noise. Additionally an SOI transistor has lower parasitic capacitances and consequently exhibits a better RF performance. For sub-50 nm CMOS technologies the substrate doping in conventional CMOS will become too high and SOI technology may become the most attractive alternative.

An extension of SOI, *Silicon On Anything (SOA)* technologies is also being developed. In this way, 'conventional CMOS' processing is performed on normal bulk (or epitaxial) silicon. After process finishing, the wafer is glued upside down to a substrate, such as glass. Subsequently, the bulk silicon is etched away, such that the transistors, the metals and the thick top glass layer remain. The component looks like an upside down circuit on a glass substrate. This version of SOA is also called *silicon on glass* [16].

3.9 Conclusions

This chapter presents a summary of different MOS processes and their component steps. The summary starts with a basic nMOS process and proceeds to a complex deep-submicron CMOS process with more than twenty masks.

The processing requirements for different types of circuits can be quite diverse. RAMs, for example, require a technology that allows very high bit densities. CMOS static RAMs therefore require tight n^+-diffusion to n-well spacings. This can be achieved when a retrograde-well implantation is used to minimise lateral well diffusion.

The continuous drive for smaller feature sizes has produced processes in which complex high-density circuits can be integrated. A state-of-the-art, deep-submicron technology is therefore discussed here as well. Finally, several trends are discussed which focus on future technology requirements. Chapters 9 and 11 focus on the physical and electrical design consequences of the continuous scaling process.

3.10 References

General

[1] Richard C. Jaeger,
'Introduction to Microelectronics Fabrication',
Modular series on solid state devices, volume V, Addison-Wesley, 1988

Technical Publications:

[2] IEEE Transactions on Electron Devices

[3] IEEE Transactions on Semiconductor Manufacturing

Conferences:

[4] International Electron Device Meeting (IEDM)

[5] European Solid State Device Conference (ESSDERC)

Technology and production processes:

[6] S.M. Sze,
'VLSI technology',
McGraw Hill, New York, 1983

[7] L.C. Parillo, et al.,
'Twin-Tub CMOS - A technology for VLSI circuits',
IEEE IEDM Conference, pp 752-755, Washington, 1980

[7a] S.M. Sze,
'Modern Semiconductor Device Physics',
John Wiley & Sons, 1997

[7b] James R. Sheats, Bruce W. Smith,
'MICROLITHOGRAPHY, Science and Technology',
Marcel Dekker Inc., 1998

References used in the text

[8] Levinson and Arnold,
'Handbook of Microlithography, Micromachining and Microfabrication', Vol.1, Microlithography'
SPIE International Society for Optical Engineering, 1997

[9] Dipankar Pramanik,
'Challenges for intermetal dielectrics',
Future Fab International, 1997

[10] S. Wolf and R.N. Tauber,
'Silicon Processing for the VLSI Era',
Volume-1, Process Technology, Lattice Press, 1986

[11] R.F.M. Roes, et al.,
'Implications of pocket optimisation on analog performance in deep sub-micron CMOS',
ESSDERC, digest of technical papers, 1999, pp 176-179.

[12] Harold J. Hovel,
'Status and prospects for SOI materials',
Future Fab International, 1997

[13] IEEE International Solid-State Circuits Conference, 1999, Joint Session:
'SOI Microprocessors and Memory',
ISSCC, Digest of Technical Papers, 1999, pp 426-439.

[14] T. Buchholtz, et al.,
'A 660 MHz 64b SOI Processor with Cu Interconnects',
ISSCC, Digest of Technical Papers, February 2000

[15] J.L. Pelloie, et al.,
'SOI Technology Performance and Modelling',
ISSCC, Digest of Technical Papers, 1999, pp 428-429.

[16] R. Dekker, et al.,
'An Ultra Low-Power RF Bipolar Technology on Glass',
IEDM, Digest of Technical Papers, 1997, pp 921-923.

[17] C. Juffermans, Philips Research Labs
'Private Communication',
January 2000

3.11 Exercises

1. Why is the formation of the gate oxide a very important and accurate process step?
2. Briefly explain the major differences between the diffusion process and the ion-implantation process. What are the corresponding advantages and disadvantages?
3. What are the possible consequences of an aluminium track with a bad step coverage?
4. Describe the main differences between the formation of LOCOS and STI.
5. What are the major advantages of self-aligned sources and drains?
6. Why is planarisation increasingly important in modern deep-submicron technologies?
7. Assume that the sixth metal layer in a $0.25\,\mu$m CMOS process is optional. In which designs would you use the sixth metal and why? What is/are the advantage(s)/disadvantage(s) of using the sixth metal layer?
8. Why was copper not used earlier in the metallisation part of a CMOS process?
9. What are the disadvantages of plasma etching?
10. What are 'tiles', as meant in the manufacture of a deep-submicron chip? Why may they be needed in such a design?
11. For which type of circuits would SOI be particularly beneficial in terms of speed and power?

Chapter 4

CMOS circuits

4.1 Introduction

Although it existed in the seventies, it took until the mid-eighties before CMOS became the leading technology for VLSI circuits. Prior to that time, only a few circuits were designed in CMOS. These early designs were generally limited to analogue circuits and digital circuits that dissipated little power. Examples include chips for calculators, watches and remote controls. CMOS offers both n-type and p-type MOS transistors. This renders CMOS processes the most complex of all MOS technologies. Initially, this meant that CMOS circuits were more costly than their nMOS equivalents.

The majority carriers in pMOS and nMOS transistors are holes and electrons, respectively. The mobility of holes is about three times lower than electron mobility. This makes pMOS circuits significantly slower than nMOS circuits of equal chip area. The continuous drive for increased integrated circuit performance therefore led to the early disappearance of pMOS technologies. The demand for higher packing densities and performance led to an increase in the complexity of nMOS processes.

In particular, the quest for a lower τD product (power delay product) necessitated the availability of several different transistor threshold voltages in a single nMOS process. These included a few enhancement threshold voltages ($V_T > 0$) and different depletion threshold voltages ($V_T < 0$). Even threshold voltages of zero volts had to be available. These threshold voltages were provided at the cost of additional masks and extra processing steps, which rapidly elevated the complexity of

nMOS processes to about the level of CMOS processes. A few advantages afforded by CMOS processes therefore led to their domination of the MOS IC world.

Modern manufacturing processes make it possible to integrate increasingly complex circuits and even complete systems on a single chip. The resulting volume of transistors per chip may reach tens of millions. The associated power dissipation can easily exceed the critical 1 W maximum limit for cheap plastic IC packages. Circuits that are manufactured in CMOS processes generally consume less than half and mostly only a fifth of the power dissipated by an nMOS equivalent. Moreover, CMOS circuits have better noise margins. These advantages have led to the use of CMOS for the integration of most modern VLSI circuits. These include memories, digital signal processors, microprocessors, speech synthesizers, data communication chips and complete Systems On Chip (SOC).

The various CMOS processes and their characteristic properties are extensively treated in section 3.8. This chapter starts with a discussion on basic nMOS circuits to be able to understand CMOS circuit properties more easily. Basic design principles and problems associated with CMOS are subjects of this chapter. Several different types of both static and dynamic CMOS circuits are discussed. One of the significant handicaps associated with CMOS circuits is the 'latch-up' problem. This subject is discussed in chapter 9, together with clock routing strategies and other timing issues. The chapter ends with a section on CMOS layout design. A layout design method is illustrated by means of an example.

Finally, it should be noted that many examples are based on an n-well CMOS process. Initially, this process was chosen because of its compatibility with the conventional nMOS process. In addition, many dynamic CMOS circuits are 'nMOS-mostly'. Currently, most processes are twin-well CMOS processes, in which the nMOS and pMOS transistors can both be realised with optimum performance.

4.2 The basic nMOS inverter

4.2.1 Introduction

Generally, the electrical properties of a static nMOS circuit are completely determined by its DC behaviour and transient response. These

will be explained with the aid of one of the most elementary MOS circuits, i.e. the inverter.

Figure 4.1 shows schematics of an inverter and its different types of *'load elements'*.

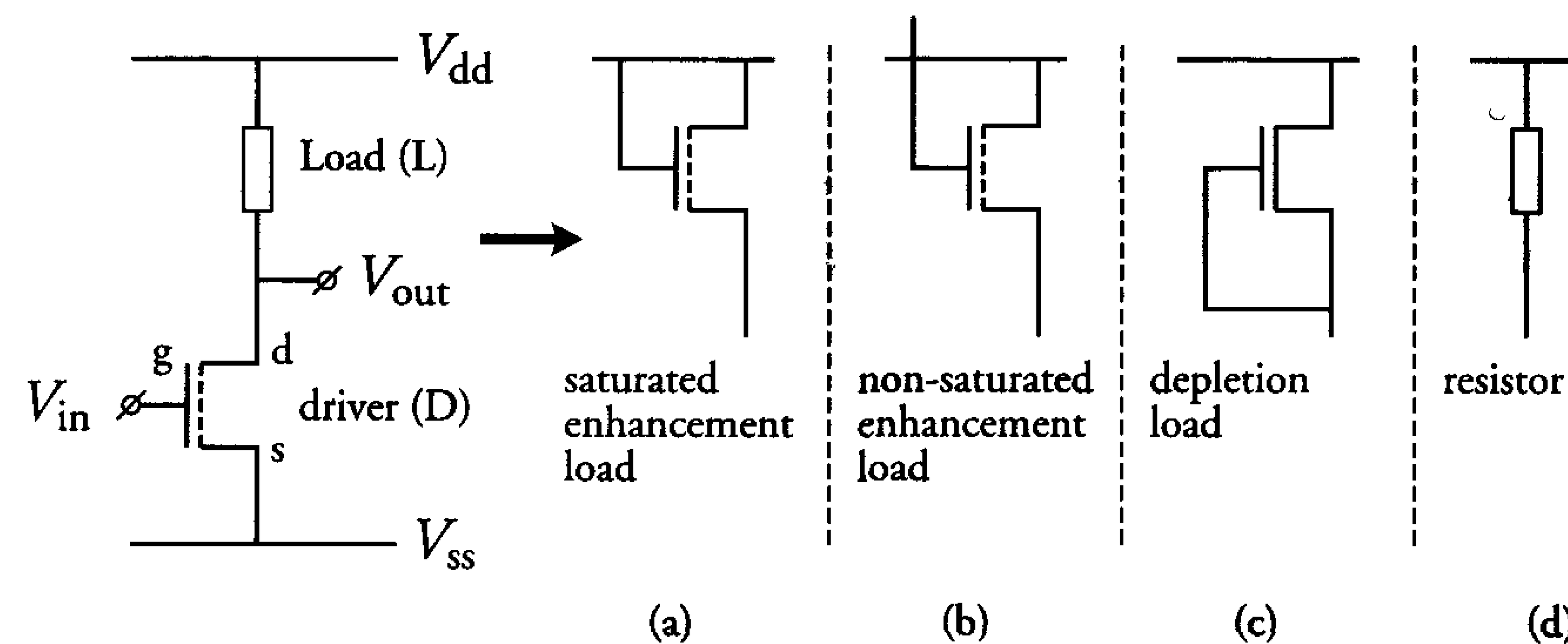


Figure 4.1: *An inverter and its different types of load elements*

The inverter's DC behaviour and transient response are discussed for its different types of load elements. The discussions are based on formulae (1.17) which express the current in a transistor as follows:

$$\text{Linear region}: \quad I_{ds} = \beta(V_{gs} - V_T - V_{ds}/2)V_{ds} \quad (V_{ds} < V_{gs} - V_T)$$

$$\text{Saturation region}: \quad I_{ds} = \beta/2(V_{gs} - V_T)^2 \quad (V_{ds} \geq V_{gs} - V_T)$$

$$\text{Where}: \quad V_T = V_x + k\sqrt{V_{sb} + 2\phi_f}$$

Two criteria are important when determining the dimensions of transistors in MOS logic gates:

- The location of the operating points. These are the output voltages V_L and V_H, which correspond to the logic values '0' and '1', respectively. Output voltage V_L, for example, must be a *'noise margin'* less than the threshold voltage V_{T_D} of the n-type enhancement driver transistor. The noise margin ensures that subsequent logic gates always interpret V_L correctly. V_{T_D} is about 0.5 V and a noise margin of about 0.25 V is normally used. This implies that $V_L \leq 0.25$ V in nMOS circuit design.

- The transient response. This implicitly refers to the rise and fall times associated with changes in the output's logic levels.

In the next sections, these criteria are discussed for the four types of inverters shown in figure 4.1.

4.2.2 The DC behaviour

The DC behaviour of inverters with different types of load elements are explained separately below with the aid of figure 4.2. This figure shows the *'driver transistor'* characteristic $I_{ds} = f(V_{ds})|_{V_{gs}=V_H}$ together with the *'load lines'* of the different load elements in figure 4.1. The shapes of the load lines are characteristic of the respective load elements.

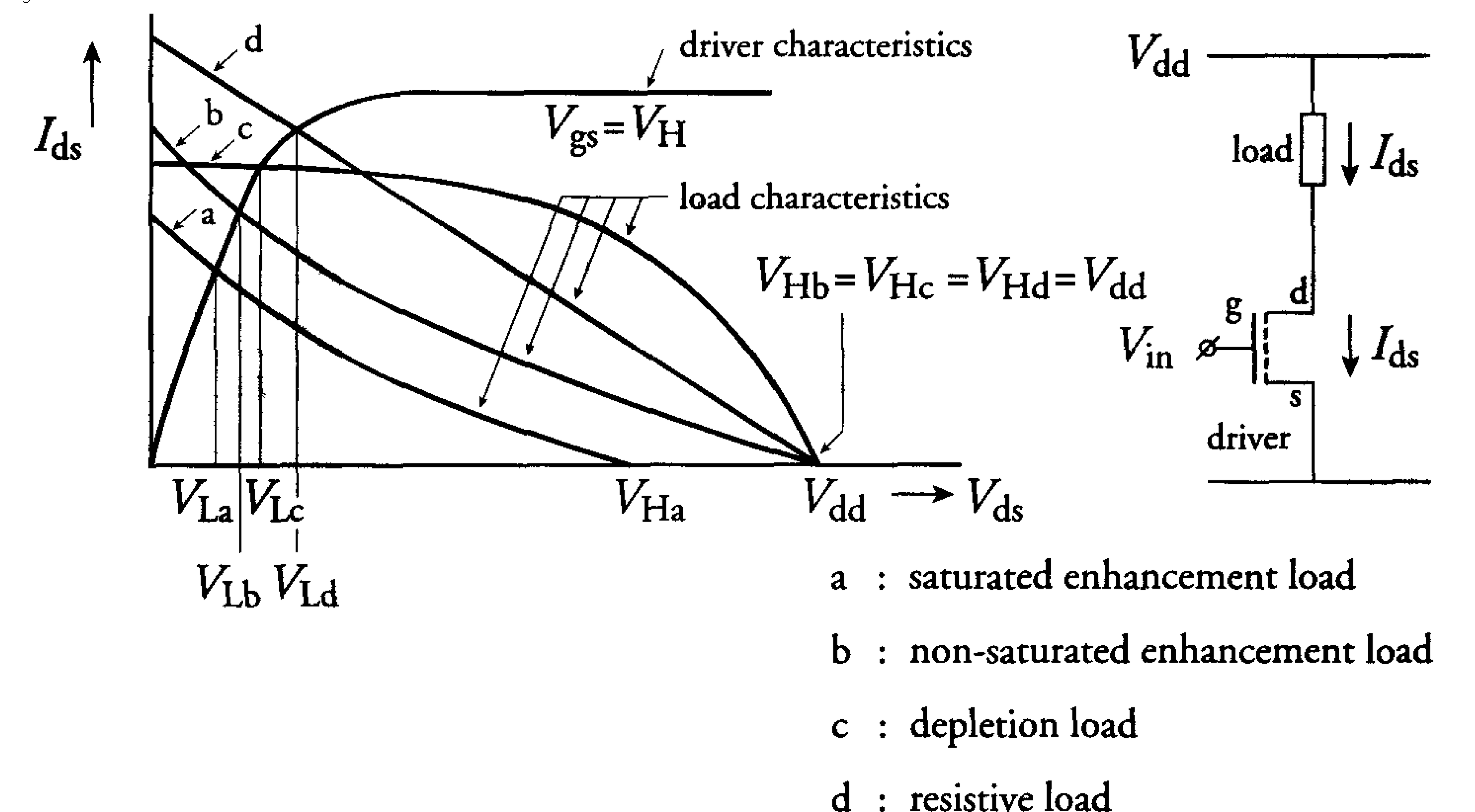


Figure 4.2: *Inverter characteristics for different types of load elements*

The output voltage of an inverter is 'low' ($V_{out} = V_L$) if its input voltage is 'high' ($V_{in} = V_H$) and vice versa. The output low level values corresponding to the different load elements are determined by the intersection of the driver characteristic and the relevant load line. These values are indicated by V_{La}, V_{Lb}, etc. in figure 4.2. The indicated positions are chosen for clarity and are not typical for the various load elements. The point of intersection between a load line and the driver characteristic is in fact chosen by the designer. For inverters that use transistors as

load elements, this point is determined by the *'aspect ratio'* A, which is expressed as follows:

$$A = \frac{(\frac{W}{L})_D}{(\frac{W}{L})_L}$$

Achieving a correct 'low' level in static nMOS logic clearly requires a minimum ratio between the driver and load transistor sizes. This type of circuit is therefore called *ratioed logic.*

Saturated enhancement load transistor

The DC behaviour of an inverter with a *saturated enhancement load* transistor is explained with the aid of figure 4.3, which shows a schematic diagram of the inverter. The load line and four driver characteristics, for different values of V_{in}, are also shown.

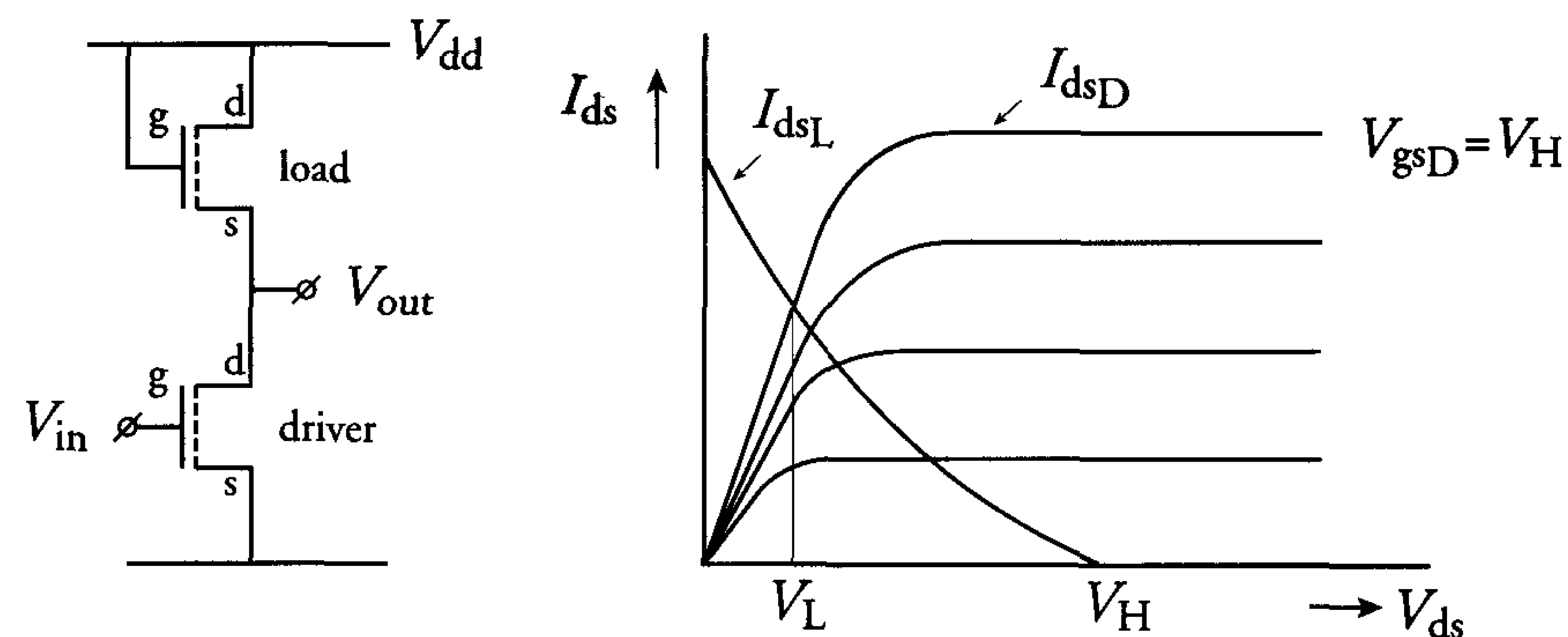


Figure 4.3: *An inverter with a saturated enhancement load transistor*

The minimum drain-source voltage of the load transistor is a threshold voltage, i.e. as $V_{ds}=V_{gs}$, $V_{ds_L} > V_{gs_L} - V_{T_L}$ always applies. Therefore, the load transistor always operates in the saturation region. The application of formulae (1.17) yields the following expression for the current in the load transistor:

$$I_{ds_L} = \frac{\beta_L}{2}(V_{ds_L} - V_{T_L})^2$$

The DC operation of an inverter with a saturated enhancement load transistor is described as follows:

- If $V_{in} = V_L < V_{T_D}$, then the driver transistor is 'off' and $I_{ds_D} = I_{ds_L} = 0$.
  According to the above expression for I_{ds_L}, the output voltage is then: $V_{out}=V_H=V_{dd}-V_{T_L}$.
- If $V_{in}=V_H \gg V_{T_D}$ then $V_{out}= V_L$. The driver current I_{ds_D} and the load transistor current I_{ds_L} will then be equal:

$$I_{ds_D} = I_{ds_L}$$

$$\Rightarrow \underbrace{\beta_D \cdot \left(V_H - V_{T_D} - \frac{V_L}{2}\right) \cdot V_L}_{\text{driver transistor in linear region}} = \underbrace{\frac{\beta_L}{2}((V_{dd} - V_L) - V_{T_L})^2}_{\text{load transistor always saturated}}$$

Assuming $V_L \ll V_{dd}$ and $V_L/2 \ll V_H - V_{T_D}$ yields:

$$(\frac{W}{L})_D \cdot (V_H - V_{T_D})V_L = (\frac{W}{L})_L \cdot \frac{1}{2} \cdot (V_{dd} - V_{T_L})^2$$

With $V_{dd}-V_{T_L}= V_H$, this reduces to the following expression for the aspect ratio A of this inverter:

$$A = \frac{(\frac{W}{L})_D}{(\frac{W}{L})_L} \geq \frac{V_H^2}{2(V_H - V_{T_D})V_L} \tag{4.1}$$

The use of a saturated enhancement load transistor is disadvantaged by the associated *'threshold loss'*, which produces a high level V_H, and this is only $V_{dd}-V_{T_L}$ rather than V_{dd}. The corresponding relatively low input voltage applied to a subsequent logic gate results in a lower speed. The use of a non-saturated enhancement or depletion load transistor overcomes this problem and produces a V_H equal to V_{dd}.

The non-saturated enhancement load transistor

An inverter with a *non-saturated enhancement load* transistor is illustrated in figure 4.4.

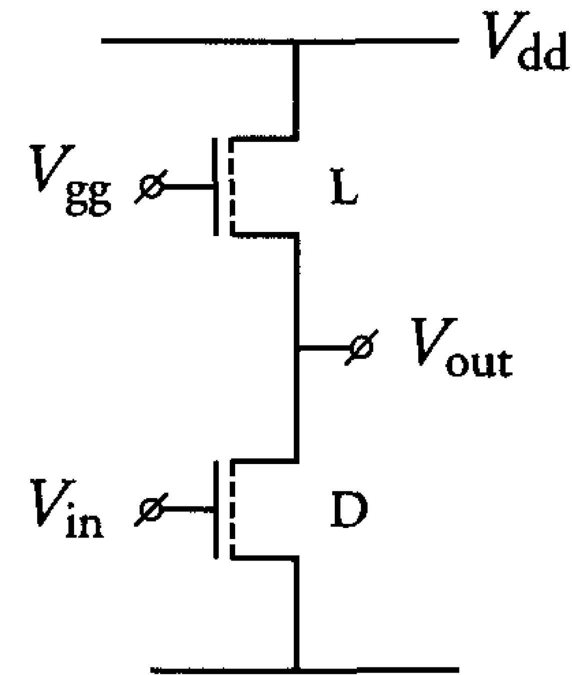


Figure 4.4: *An inverter with a non-saturated enhancement load transistor*

The gate of the load transistor is connected to an extra supply voltage V_{gg} instead of the supply voltage V_{dd}. The extra supply voltage is large enough to ensure that the load transistor always operates in the non-saturated region, i.e. $V_{gg} > V_{dd} + V_{T_L}$.

The DC operation of the above inverter is described as follows:

- $V_{in} = V_L < V_{T_D} \Rightarrow I_{ds_L} = 0\,\text{mA}$ and $V_{out} = V_H = V_{dd}$.
- $V_{in} = V_H \gg V_{T_D} \Rightarrow V_{out} = V_L$.

The driver now operates in the linear region. The driver and load transistor currents are equal:

$$\begin{aligned} I_{ds_D} &= (\frac{W}{L})_D \cdot \beta_\square \left(V_H - V_{T_D} - \frac{V_L}{2}\right) V_L \\ &= (\frac{W}{L})_L \cdot \beta_\square \left(V_{gg} - V_L - V_{T_L} - \frac{V_{dd} - V_L}{2}\right)(V_{dd} - V_L) \\ &= I_{ds_L} \end{aligned}$$

Assuming $V_L \ll V_{dd}$, $\frac{V_L}{2} \ll V_H - V_{T_D}$ and $V_{gg} - V_{T_L} \gg V_L$ yields the following expression for the inverter's aspect ratio A:

$$A = \frac{(\frac{W}{L})_D}{(\frac{W}{L})_L} = \frac{\left(V_{gg} - V_{T_L} - \frac{V_{dd}}{2}\right) \cdot V_{dd}}{(V_H - V_{T_D}) \cdot V_L}$$

Since $V_H - V_{T_D} < V_{dd}$, the aspect ratio A is expressed as follows:

$$A = \frac{(\frac{W}{L})_D}{(\frac{W}{L})_L} \geq \frac{V_{gg} - V_{T_L} - \frac{V_{dd}}{2}}{V_L} \tag{4.2}$$

The use of a non-saturated enhancement transistor as load element has the following advantages:

- High V_H ($=V_{dd}$);
- Large noise margin;
- Fast logic.

The most significant disadvantage is the extra supply voltage required V_{gg} ($V_{gg} \geq V_{dd} + V_{T_L}$), which may necessitate an extra pin on the chip package. Alternatively, V_{gg} can be electronically generated on the chip. This results in a *'bootstrapped load'* element, as shown in figure 4.5.

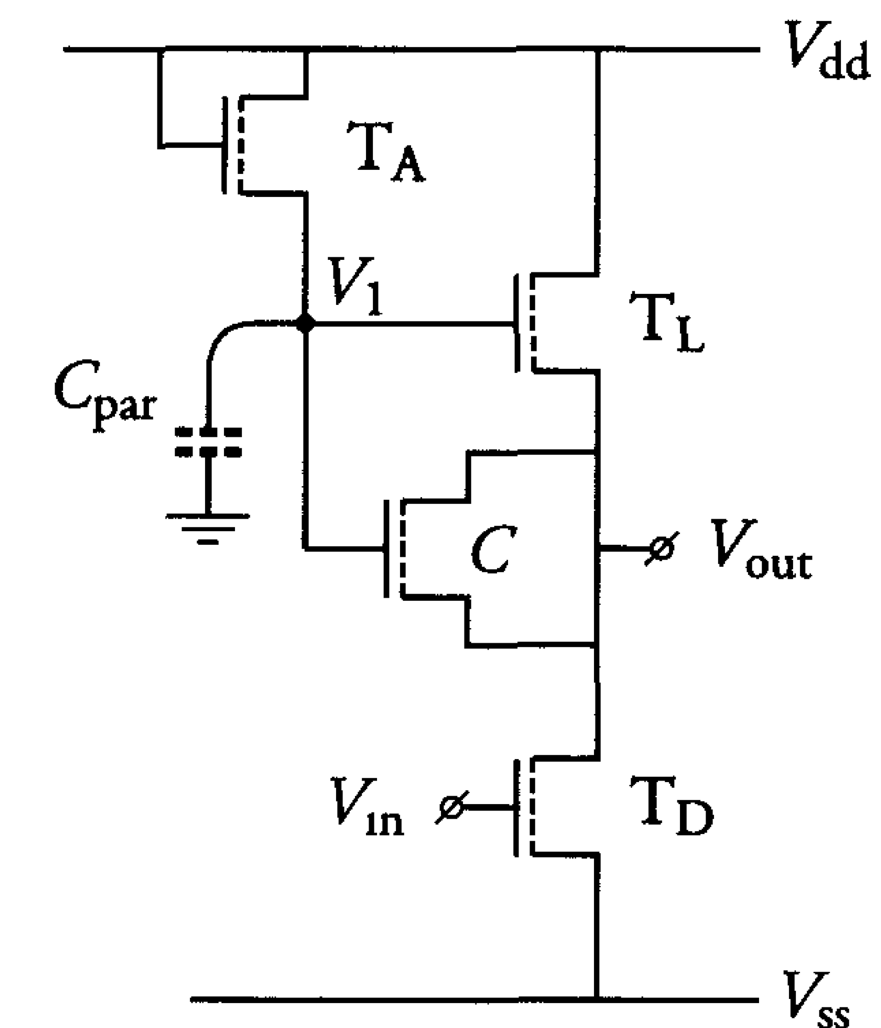


Figure 4.5: *The bootstrapped inverter*

The DC operation of the bootstrapped inverter is explained as follows:

- If $V_{in} = V_H$, then $V_{out} = V_L$ and $V_1 = V_{dd} - V_{T_A}$. The MOS *'bootstrap' capacitance* C therefore charges.
- When V_{in} switches from V_H to V_L, then V_{out} increases by ΔV_{out} and V_1 increases by ΔV_1. The magnitude of ΔV_1 is determined by the values of the bootstrap capacitance C and the parasitic capacitance C_{par} such that:

$$\Delta V_1 = \frac{C}{C + C_{par}} \cdot \Delta V_{out}$$

This means that V_1 immediately passes the V_{dd}–V_{T_A} level and transistor T_A therefore no longer conducts. The voltage V_1 can then further increase to a voltage greater than V_{dd}. The maximum value of V_1 is determined by the capacitance ratio:

$$a = \frac{C}{C + C_{par}}$$

The value of a required to produce a 'high' output voltage is: $V_H = V_{dd}$ and is derived as follows:
$V_H = V_{dd}$ when $V_1 \geq V_{dd} + V_{T_L}$.
$\Delta V_1 = a \cdot \Delta V_{out}$ and $V_1 = V_{dd} - V_{T_A} + a \cdot \Delta V_{out}$.

The load transistor T_L must remain in the linear operating region. The following equation therefore applies:
$V_1 - V_{T_L} > V_{dd}$

$$\Rightarrow V_{dd} - V_{T_A} - V_{T_L} + a \cdot \Delta V_{out} > V_{dd}$$

$$\Rightarrow \Delta V_{out} > \frac{V_{T_A} + V_{T_L}}{a}$$

The output high level must be equal to the supply voltage, i.e. $V_{out} = V_H = V_{dd}$. Therefore, $\Delta V_{out} = V_{dd} - V_L$. Assuming $V_{T_A} \approx V_{T_L}$ yields the following expression for a:

$$a > \frac{2V_{T_L}}{V_{dd} - V_L} \qquad (4.3)$$

- If $V_{in} = V_H$, then $V_{out} = V_L$ and the gate voltage of the load transistor T_L is $V_{dd} - V_{T_A} \approx V_{dd} - V_{T_L}$. Load transistor T_L therefore operates in the saturation region when $V_{out} = V_L$. The aspect ratio A of the bootstrapped inverter is therefore identical to that given in equation (4.1) for the inverter with a saturated enhancement load transistor.

The bootstrapped inverter has the following advantages:

1. There is no threshold loss when the bootstrap capacitance C is correctly dimensioned.
2. There is no extra supply voltage required.

The depletion load transistor

The manufacture of depletion transistors requires an extra mask (DI) and additional processing steps. There are, however, considerable advantages associated with the use of a depletion transistor as load element. These include the following:

- The output high level equals V_{dd}, i.e. $V_H = V_{dd}$;
- There is no extra supply voltage required;
- Circuit complexity is minimal and bootstrapping is unnecessary;
- Noise margins are high.

For these reasons, most nMOS processes are *'E/D technologies'* and contain both enhancement and depletion transistors. Some manufacturers even include depletion transistors in their CMOS technologies.

Figure 4.6 shows an inverter with a *depletion load* transistor.

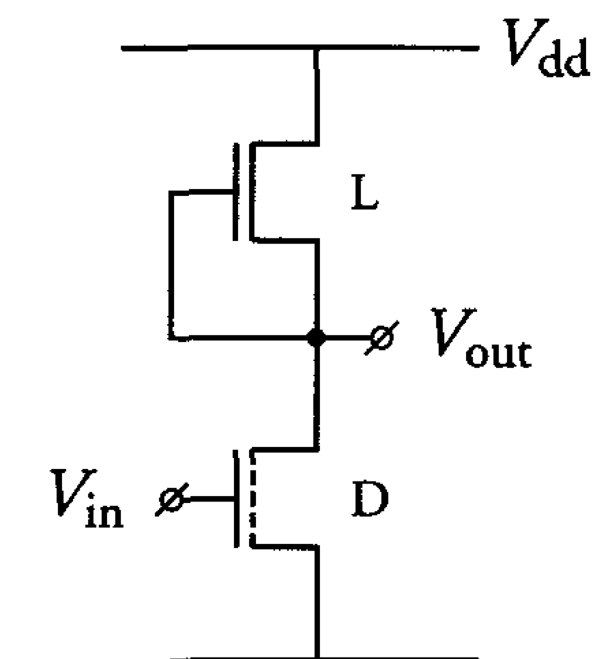


Figure 4.6: *An inverter with a depletion load transistor*

The DC operation of the inverter with a depletion load transistor is described as follows:

- The depletion load transistor has a negative threshold voltage which is usually between −1 V and −3 V. Therefore, $V_{out} = V_H = V_{dd}$ when $V_{in} = V_L < V_{T_D}$.
- When $V_{in} = V_H > V_{T_D}$, then $V_{out} = V_L$ and $V_{gs_L} (= 0\,V) < V_{ds_L} + V_{T_L}$. In this case, the load transistor operates in the saturation region

while the driver transistor operates in the triode region. Equating the currents in the load and driver transistors yields:

$$I_{ds_D} = I_{ds_L}$$

$$\Rightarrow (\frac{W}{L})_D \cdot \beta_\square \cdot \left(V_H - V_{T_D} - \frac{V_L}{2}\right) \cdot V_L = (\frac{W}{L})_L \cdot \frac{\beta_\square}{2} \cdot V_{T_L}{}^2$$

If $\frac{V_L}{2} \ll V_H - V_{T_D}$, then the aspect ratio A of the depletion-load inverter can be expressed as follows:

$$A = \frac{(\frac{W}{L})_D}{(\frac{W}{L})_L} \geq \frac{V_{T_L}{}^2}{2V_L \cdot (V_H - V_{T_D})} \tag{4.4}$$

The resistive load

VLSI circuits may consist of millions of logic gates which may dissipate no more than 0.1 to 1 μW each. A supply voltage of 2.5 V therefore requires a *resistive load* of several MΩ per logic gate. Both diffusion and polysilicon have a *sheet resistance* of about 25 $\Omega/\square$. Realisation of a 2.5 MΩ resistance in a 0.25 μm wide polysilicon track therefore requires a length of 25 mm. At the cost of extra processing complexity, however, large resistances can be realised on small chip areas. For random-access memories (RAMs), the disadvantages of complex processing are justified by very large production quantities (1.5 billion 4M-DRAMs in 1996). The addition of a second polysilicon layer with very high resistivity in many static RAM processes facilitates the realisation of memory cells that are considerably smaller than the *full-CMOS* cells. The use of resistive load elements is therefore mainly limited to application in static memories and it is not normally encountered in VLSI circuits.

Figure 4.7 shows an inverter with a resistance as load element.

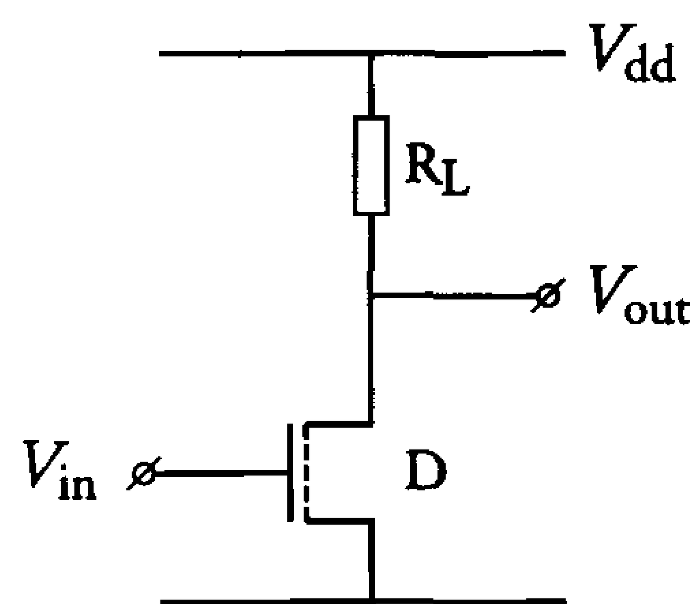


Figure 4.7: *An inverter with a resistive load*

The DC operation of the above inverter is described as follows:

- When $V_{in} = V_L < V_{T_D}$, then $V_{out}=V_H=V_{dd}$.
- When $V_{in} = V_H > V_{T_D}$, then $V_{out}=V_L$. The driver transistor is then operating in the linear region. Equating the currents in the driver transistor and load resistance yields:

$$I_{ds_D} = (\frac{W}{L})_D \cdot \beta_\square \left(V_H - V_{T_D} - \frac{V_L}{2}\right) V_L = \frac{V_{dd} - V_L}{R_L} = I_R$$

$$\Rightarrow (\frac{W}{L})_D = \frac{V_{dd} - V_L}{\beta_\square R_L \left(V_{dd} - V_{T_D} - \frac{V_L}{2}\right) V_L}$$

Assuming $V_{dd} - V_{T_D} - \frac{V_L}{2} \leq V_{dd} - V_L$ yields the following expression for the aspect ratio A of the driver transistor:

$$A = (\frac{W}{L})_D \geq \frac{1}{\beta_\square R_L V_L} \tag{4.5}$$

4.2.3 The transient response

The outputs of logic gates in digital MOS circuits are loaded by interconnection tracks and by the gates of other MOS transistors. These output loads behave like pure capacitances. The *transient behaviour* of nMOS circuits can therefore simply be represented by the charging and discharging of capacitances. Section 1.10 contains a complete description of the capacitances associated with the MOS transistor. The capacitances presented to the output transistors of a logic gate include the junction capacitances of their own reverse-biased source and/or drain p-n junctions. These capacitances and the gate capacitances of MOS transistors are non-linear. For a first-order approximation, however, the capacitances are assumed to be constant. Formulae (1.17) are then used to derive dimensioning formulae which apply to the transient responses of nMOS inverters with different types of load. The charging of a *capacitive load* through each of the load elements in figure 4.1 is first considered. Subsequently, the discharging of a capacitance through a driver transistor is discussed. A comparison of the transient behaviour of inverters with different types of load elements concludes this section.

Charging a capacitance through a saturated enhancement transistor

Figure 4.8 shows a capacitance C which is charged through a *saturated enhancement transistor.*

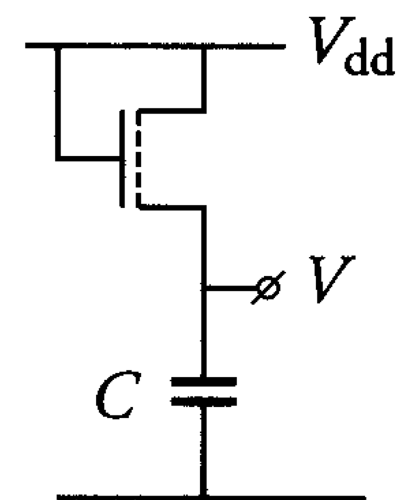


Figure 4.8: *Capacitor charging through a saturated enhancement transistor*

The voltage on the charged capacitor is V and the current that charges the capacitor is equal to the transistor current:

$$C \cdot \frac{dV}{dt} = \frac{\beta_L}{2} \cdot (V_{dd} - V_T - V)^2$$

From the solution of this differential equation, the following can be derived for β_L, with $V = 0.6 \cdot V_{dd}$, $V_L = 0.1 \cdot V_{dd}$, $V_H = 0.7 \cdot V_{dd}$ and $V_T = 0.2 \cdot V_{dd}$:

$$\beta_L = \frac{16 \cdot C}{V_{dd} \cdot t} \tag{4.6}$$

This equation calculates the required β_L for charging a capacitance C in a time t to a voltage level equal to $0.6V_{dd}$.

Charging a capacitance through a non-saturated enhancement transistor

Figure 4.9 shows a capacitance C that is charged through a *non-saturated enhancement transistor.*

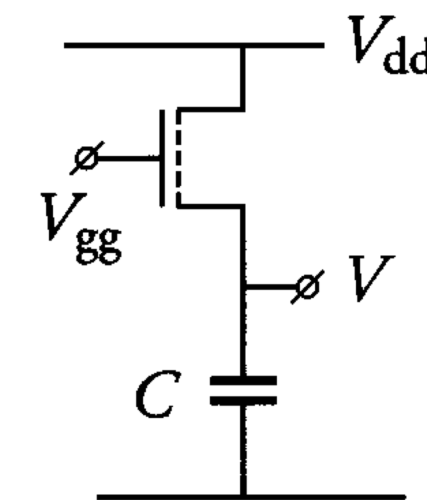


Figure 4.9: *Capacitor charging through a non-saturated enhancement transistor*

As in the previous case, formulae that describe the charging of the capacitive load are derived by equating the charging current and the transistor current:

$$C \cdot \frac{dV}{dt} = \beta_L \cdot (V_{gs_L} - V_T - V_{ds_L}/2) \cdot V_{ds_L}$$

A boundary condition of this differential equation is $V = V_L$ at $t = 0$. The solution yields the following expression for the gain factor β_L of a non-saturated enhancement transistor which will charge a capacitance C from $V_L = 0.1 \cdot V_{dd}$ to a voltage $V = 0.9 \cdot V_{dd}$ in time t ($V_T = 0.2 \cdot V_{dd}$):

$$\beta_L = \frac{6 \cdot C}{V_{dd} \cdot t}$$

for $V_{gg} = 1.4 \cdot V_{dd}$.

Charging a capacitance through a depletion transistor

Figure 4.10 shows a capacitance C, which is charged through a *depletion transistor.*

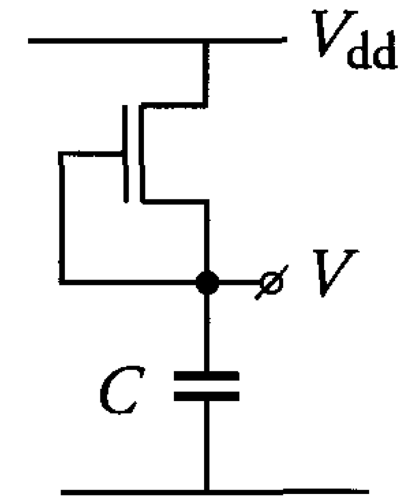


Figure 4.10: *Capacitor charging through a depletion transistor*

Charging a capacitance through a depletion transistor comprises two phases:

1. The transistor operates in the saturation region: $V_{ds_L} > V_{gs_L} - V_T$, i.e. $V_{dd} - V > -V_T$.
   In this first phase, the voltage V across the capacitance rises to a voltage V_1: $V_1 = V_{dd} + V_T$ $(V_T < 0)$.
2. For voltages above V_1, the capacitance C will be further charged through the non-saturated depletion transistor.

For these two regions of operation, two different equations have to be solved. With $V_L = 0.1 \cdot V_{dd}$, $V_T = -0.4 \cdot V_{dd}$ and $V = 0.9 \cdot V_{dd}$, the solution of these equations is:

$$\beta_L = \frac{11 \cdot C}{V_{dd} \cdot t} \tag{4.7}$$

Charging a capacitance through a resistance

Figure 4.11 shows a capacitance C, which is charged through a resistor R_L.

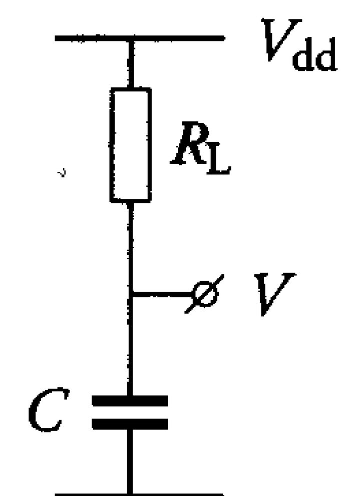


Figure 4.11: *Capacitor charging via a resistance*

As in the previous cases, the charge current and the current through the resistor are equated as follows:

$$C \cdot \frac{dV}{dt} = \frac{V_{dd} - V}{R_L}$$

This equation has the following solution, when $V_L = 0.1 \cdot V_{dd}$ and $V = 0.9 \cdot V_{dd}$:

$$R_L = \frac{0.5 \cdot t}{C} \tag{4.8}$$

Discharging a capacitance

The *driver transistor* which discharges the capacitive output load of an nMOS inverter is always of the enhancement type. Formulae that describe this discharge are derived with the aid of figure 4.12. This simplified schematic shows a load capacitance C, which is discharged through a driver transistor D.

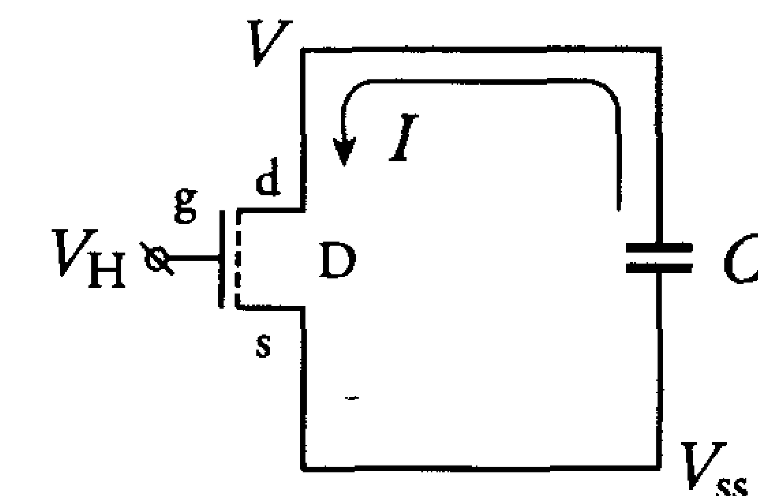


Figure 4.12: *Capacitor discharging through an enhancement transistor*

At time $t = 0$, the voltage V_H is applied to the gate of the driver transistor while the voltage on C is still V_H. Here, $V_g = V = V_H$ and thus $V_{ds_D} > V_{gs_D} - V_{T_D}$. The driver transistor therefore operates in the saturation region. It continues to do so until time t_1, when $V = V_1 = V_H - V_{T_D}$. At this time, the transistor enters the linear region and a different differential equation has to be solved from this time on. The solution for the differential equations, when $V_H = V_{dd}$, $V_T = 0.2 \cdot V_{dd}$ and $V = 0.1 \cdot V_{dd}$, is:

$$\beta_D = \frac{4 \cdot C}{V_{dd} \cdot t} \tag{4.9}$$

Comparison of the different types of nMOS inverters

NMOS inverters with different load elements are now compared. In this comparison, the current I_0 which flows in the load elements at $t = 0$ is equal for the different inverter types:

1. Saturated enhancement load transistor:

$$I_0 = \frac{\beta_L}{2}(V_{dd} - V_L - V_{T_L})^2$$

2. Non-saturated enhancement load transistor:

$$V_{gg} > V_{dd} + V_{T_L}$$
$$I_0 = \beta_L \left(V_{gg} - V_L - V_{T_L} - \frac{V_{dd}-V_L}{2}\right) \cdot (V_{dd} - V_L)$$

3. Depletion load transistor:

$$I_0 = \frac{\beta_L}{2} \cdot {V_{T_L}}^2$$

4. Resistive load:

$$I_0 = \frac{V_{dd}-V_L}{R}$$

5. Enhancement driver transistor for discharging:

(saturated when $t = 0$)
$$I_0 = \frac{\beta_D}{2} \cdot (V_{gs} - V_{T_D})^2$$

The following typical values are used to provide a quantitative comparison of the various types of nMOS inverters:
V_{dd}=2.5 V, V_L=0.25 V, I_0=100 μA and $\beta_{\square}$=240 $\mu A/V^2$.
Furthermore, we have the following typical values for the enhancement load transistor: V_{gg}=3.5 V, V_{T_D}=0.5 V, V_{T_L}=0.6 V, plus V_{T_L}=−1.5 V for the depletion load transistor.

Adopting a 1 pF load capacitance, a *circuit analysis* program was used to simulate the charging and discharging characteristics that correspond to these load and driver transistors, respectively. The charging

characteristic associated with the load resistance was also simulated. The results are shown in figure 4.13.

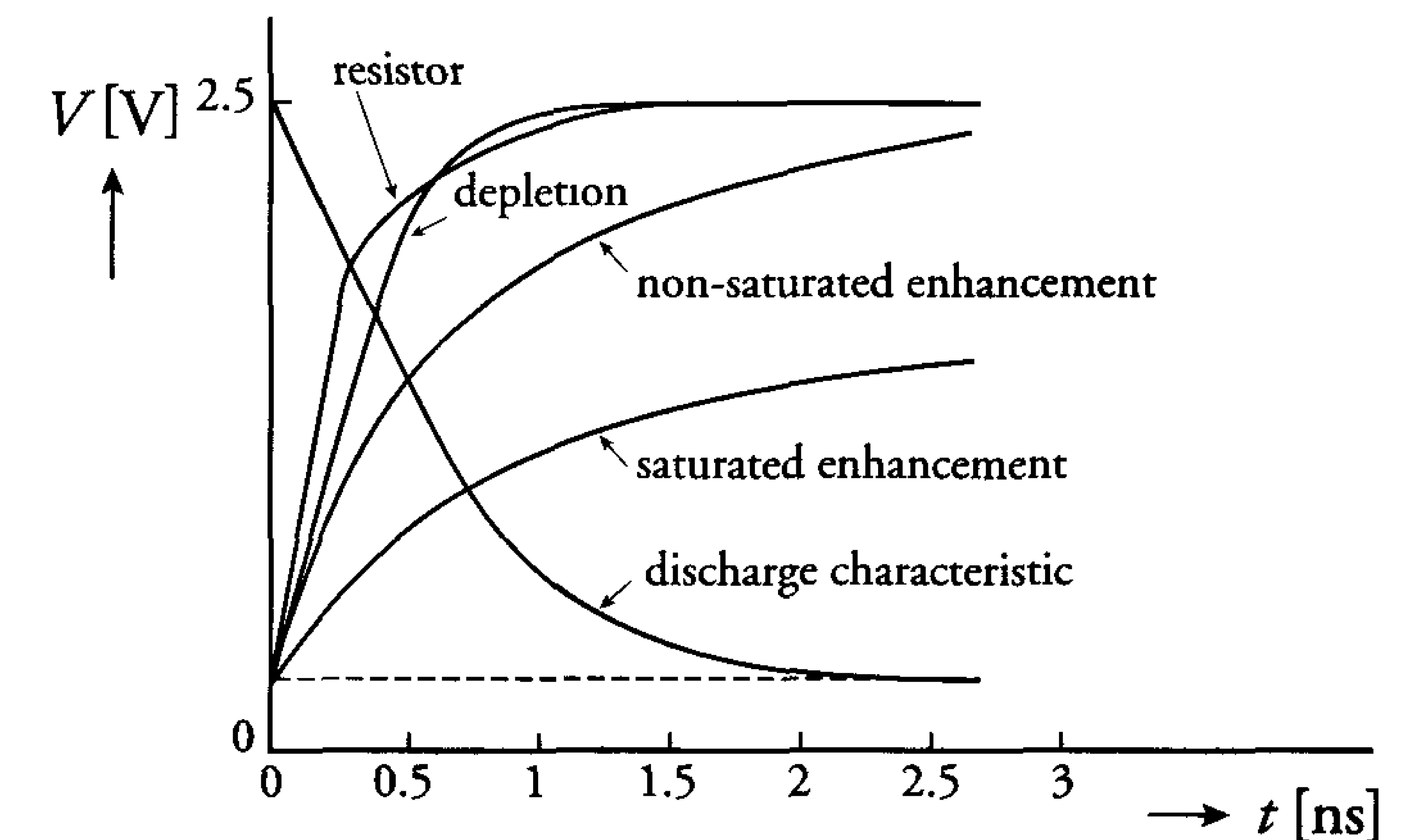


Figure 4.13: *Charging characteristics of nMOS inverters with different types of load, identical load capacitances and the same initial current* I_0

Figure 4.13 shows that the resistive and the depletion loads perform similarly for equal initial currents. Charging through a saturated enhancement load transistor is clearly very slow. In addition, the associated high level is merely $V_{dd} - V_{T_L}$. The non-saturated enhancement load transistor can be used to eliminate threshold loss and yield a high level V_H= V_{dd}. However, this type of load element is slower than a depletion load transistor. An additional disadvantage is the need for an extra supply voltage (V_{gg}) or bootstrapping.

4.2.4 Transforming a logic function into an nMOS transistor circuit

An inverter is transformed into a logic gate by replacing the driver transistor by a combination of MOS transistors. The combination may comprise series and/or parallel transistors. Each transistor gate is controlled by a logic signal. A complex logic function can therefore be implemented in a single logic gate with an associated propagation delay. The following transformation rules apply:

1. An *AND* function is realised by a *series connection* of transistors.
2. An *OR* function is realised by a *parallel connection* of transistors.

Because logic gates are an adaptation of the basic inverter, the output signal is always the inverse of the function that is derived when the transistors in the driver section are interpreted according to the above rules. In fact, implementations always comprise NAND, NOR or AND-OR-NOT functions.

Example: A 'full adder' is described by the following logic functions (see also section 7.3.5):

$$S = x\bar{y}\,\bar{z} + \bar{x}\,\bar{y}\,z + \bar{x}\,y\,\bar{z} + x\,y\,z$$
$$C_o = x\,y + x\,z + y\,z$$

Symbols x and y represent two bits which must be added. Symbol z represents the 'carry-in'. S represents the binary sum of x, y and z while C_o represents the 'carry-out'.

The logic function S can also be written as:

$$S = x\,(y\,z + \bar{y}\,\bar{z}) + \bar{x}\,(y\,\bar{z} + \bar{y}\,z)$$

This function corresponds to the implementation in figure 4.14, which realises the inverse ($\bar{S}$) of the sum function.

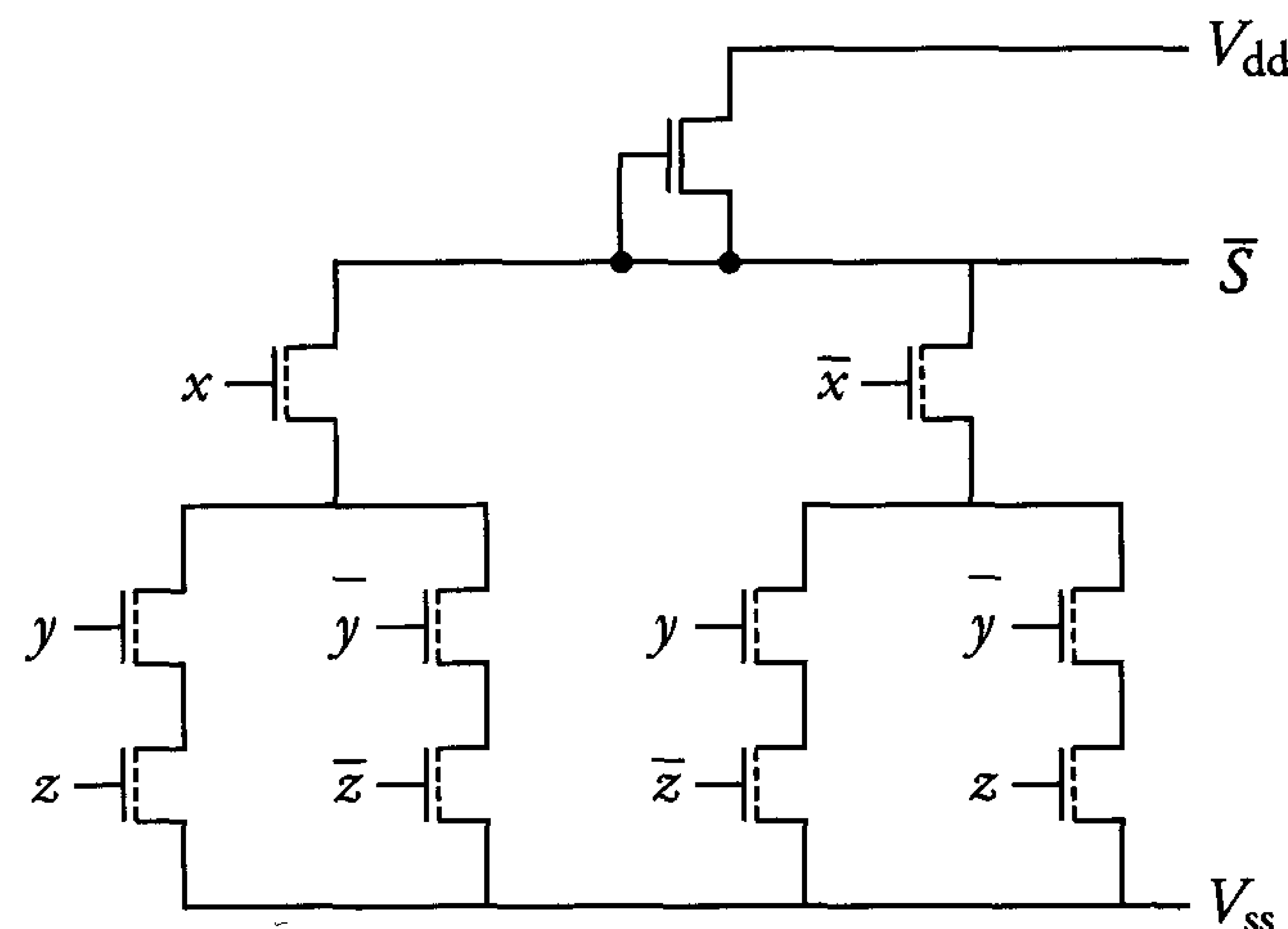


Figure 4.14: *An implementation of the function $\bar{S}$*

Figure 4.15 shows a realisation of the carry function.

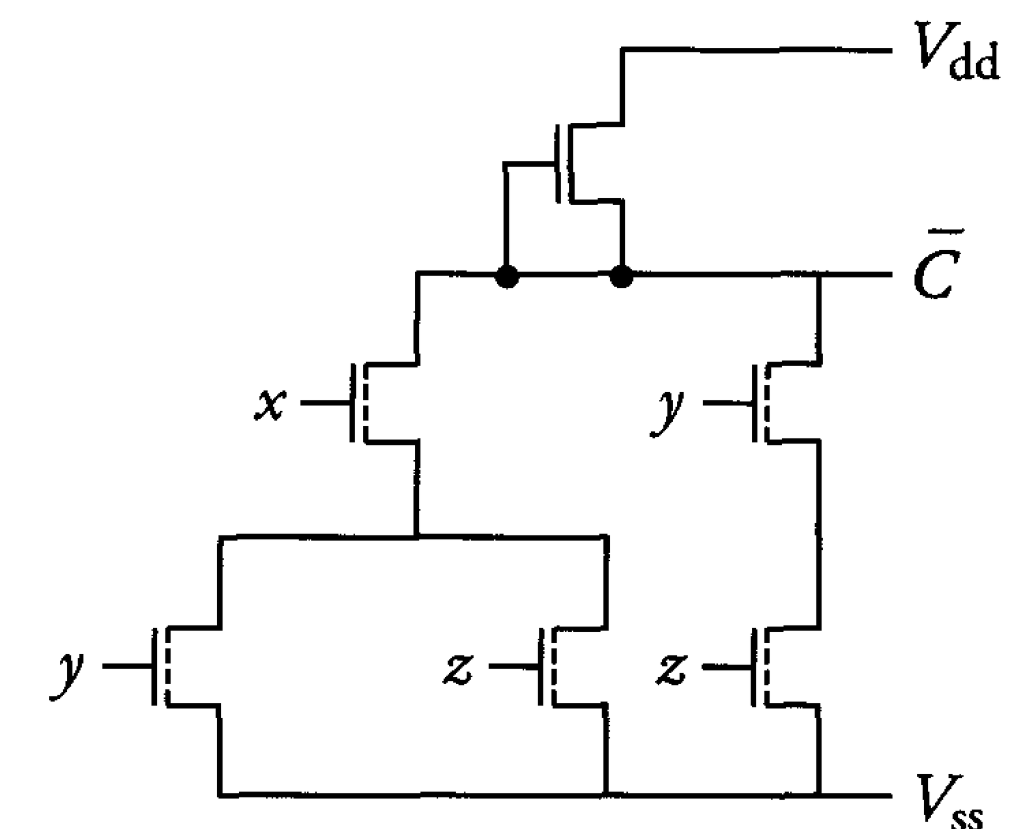


Figure 4.15: *Implementation of the full adder inverse 'carry-out' function*

An nMOS transistor's gain factor β equals $\beta_\square \cdot \frac{W}{L}$. The gain factor β_{total} of n transistors connected in series is expressed as follows:

$$\beta_{\text{total}} = (\frac{1}{\beta_1} + \frac{1}{\beta_2} + \ldots \frac{1}{\beta_n})^{-1}$$

If all the transistors have equal dimensions, then:

$$\beta_{\text{total}} = \beta / n$$

The discharge time constant associated with these n transistors is then directly proportional to n. In fact, the speed of a logic gate is largely determined by the number of transistors that are connected in series in the driver section. It is thus generally advisable to keep this number to a minimum. Figure 4.16, for example, shows a NAND gate with n driver transistors in series. The effective $(\frac{W}{L})$ ratio of these n transistors is expressed as follows:

$$(\frac{W}{L})_{\text{total}} = \frac{1}{(\frac{W}{L})_1^{-1} + (\frac{W}{L})_2^{-1} + \ldots + (\frac{W}{L})_n^{-1}} \quad (4.10)$$

The $(\frac{W}{L})$ aspect ratio of the driver transistor in an inverter can be calculated using the formulae in sections 4.2.2 and 4.2.3. For a NAND gate with n inputs, the inverter's driver transistor (D) must be replaced by n transistors in series. The NAND gate will be as fast as the inverter if its

transistors each have an aspect ratio $n \cdot (\frac{W_i}{L_i})$, where W_i and L_i are the width and length, respectively, of the inverter's driver transistor.

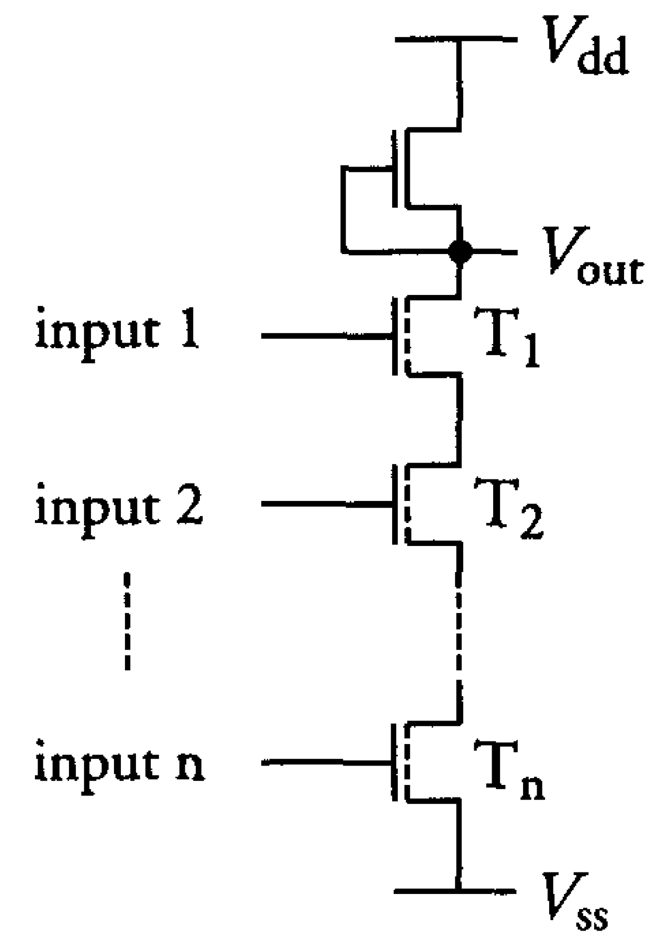


Figure 4.16: *NAND gate with n inputs and thus n transistors in series*

The number of parallel sections in a logic gate is also a critical implementation factor. The surface area and hence the parasitic capacitances associated with the logic gate increase with the number of parallel sections. This causes an increase in the gate's propagation delay.

This section presents an overview on the electrical design of nMOS circuits and the creation of basic nMOS logic gates. A major disadvantage of nMOS logic is the associated power consumption. Each logic gate with a low level at its output consumes DC power. Therefore, even when a large logic nMOS chip has no signal transitions, there is a large DC power consumption. CMOS circuits, which require more complex technologies than nMOS circuits, do not consume DC power when there is no activity. This is the most important reason for the domination of CMOS circuits in the integrated circuit market.

4.3 Electrical design of CMOS circuits

4.3.1 Introduction

The acronym CMOS stands for Complementary Metal Oxide Semiconductor'. The word 'complementary' indicates that transistors of different types can be manufactured in CMOS processes. The types are

n-channel and p-channel, or 'nMOS and 'pMOS'. The nMOS transistor and its operation have been extensively treated before. The pMOS transistor has been briefly mentioned. Where necessary, additional details about its operation are provided in this chapter. The nMOS and pMOS transistors used in CMOS processes are both of the enhancement type. Section 1.7 reveals that the threshold voltage of the nMOS transistor is therefore positive while that of the pMOS transistor is negative. This is shown in figure 4.17.

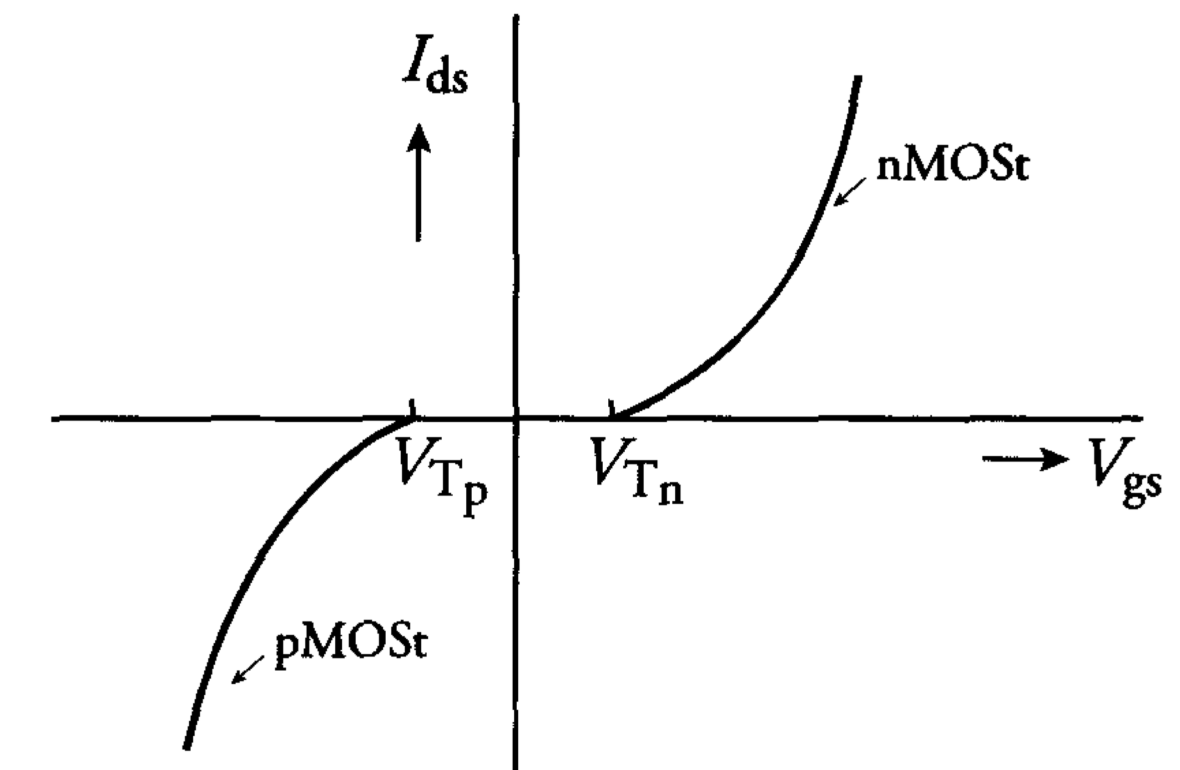


Figure 4.17: *The $I_{ds}=f(V_{gs})$ characteristics of nMOS ($V_{T_n}>0$) and pMOS ($V_{T_p}<0$) enhancement transistors*

The formulae discussed in section 1.5, which describe the back-bias effect on the threshold voltages of nMOS and pMOS transistors, are as follows:

$$V_{T_n} = V_{X_n} + K_n\sqrt{V_{sb} + 2\phi_f} \quad (\text{enhancement type} : V_{X_n} > 0,\ K_n > 0)$$

$$V_{T_p} = V_{X_p} + K_p\sqrt{V_{ws} + 2\,|\phi_f|} \quad (\text{enhancement type} : V_{X_p} < 0,\ K_p < 0)$$

In the CMOS process that is considered in this section, the pMOS transistor is integrated in an n-well. Voltage V_{ws} in the above expression for the threshold voltage V_{T_p} of a pMOS transistor represents the voltage between the source of the transistor and the n-well.

The above expressions and figure show that the operation of the pMOS transistor is the exact complement of the nMOS transistor's operation. The electrical operation of the nMOS and pMOS transistors

can be summarised as follows: the pMOS transistor's behaviour with respect to the supply voltage is identical to the nMOS transistor's behaviour with respect to ground and vice versa.

4.3.2 The CMOS inverter

A basic *CMOS inverter* consists of an nMOS transistor and a pMOS transistor connected as shown in figure 4.18. The n-well serves as a substrate for the pMOS transistor. It is formed by the diffusion or ion implantation techniques discussed in chapter 3.

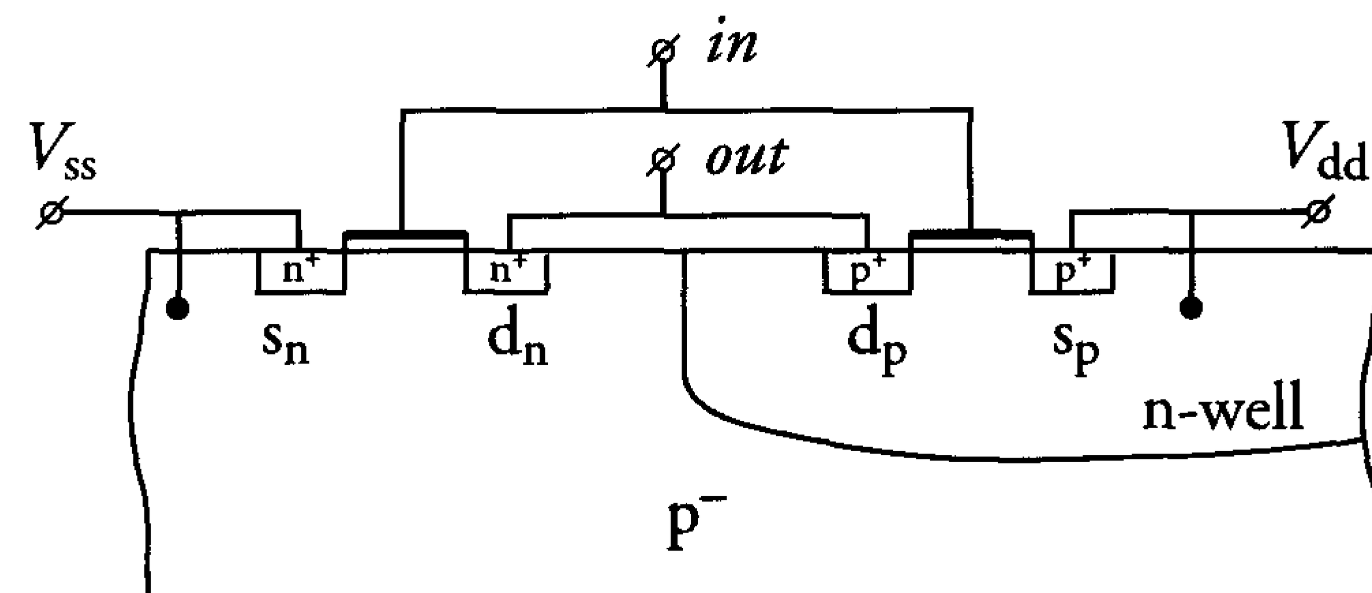


Figure 4.18: *Transistor connections for a CMOS inverter*

Figure 4.19 shows the circuit diagram of a CMOS inverter.

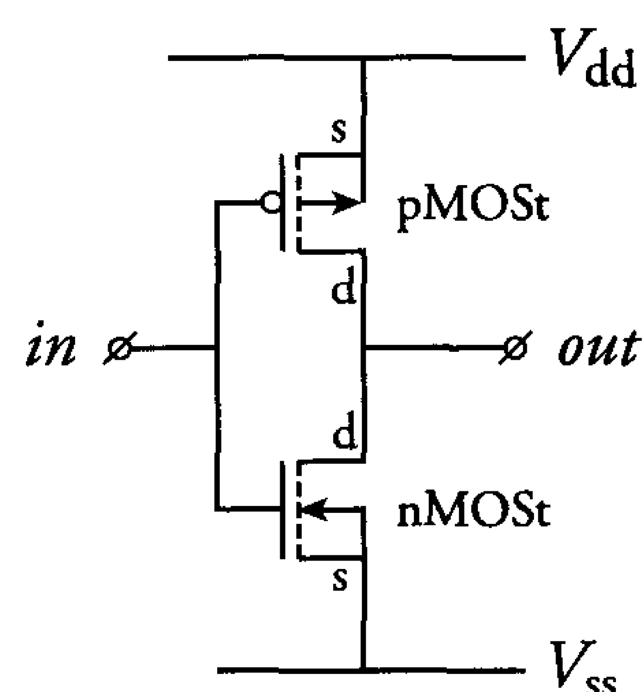


Figure 4.19: *Circuit diagram of a CMOS inverter*

The influence of substrate voltage on the threshold voltage of a transistor is discussed in section 1.5. This undesirable back-gate effect is proportional to the square root of the channel dope of the transistor

and is represented by the K-factor. The K-factor of the pMOS transistors in a retrograde twin well process can be of the same order as that of the nMOS transistors.

The performance of the pMOS transistor is hampered by the mobility of holes, which is approximately two to four times lower than the mobility of electrons. This leads to the following relationship between the effective $\beta_\square$ factors of nMOS and pMOS transistors (including second order effects):

$$\beta_{\square_n} \approx 3 \cdot \beta_{\square_p}$$

For equal absolute threshold voltage values, the pMOS transistor in the layout of an inverter with symmetrical behaviour will therefore be about 3 times the size of the nMOS transistor. This size ratio is expressed in the 'aspect ratio' A of the CMOS inverter as follows:

$$A = \frac{\left(\frac{W}{L}\right)_p}{\left(\frac{W}{L}\right)_n} \tag{4.11}$$

In many processes, all polysilicon areas and the sources and drains of nMOS transistors in an n-well CMOS process are n^+ areas. The sources and drains of the pMOS transistors are p^+ areas. It should be clear from figure 4.18 that p^+ and n^+ areas may *never* be directly connected in the ACTIVE mask, not even in a stick diagram. Such an interconnection would produce a pn diode which only conducts in one direction. Connections between n^+ and p^+ areas must therefore always be made in metal. Many CMOS processes currently include *double-flavoured polysilicon*, or *dual-dope polysilicon*: n^+ polysilicon gate for the nMOS transistor and p^+ polysilicon for the pMOS transistor.

The electrical behaviour of the CMOS inverter

An nMOS inverter comprises a driver and a load transistor. However, the pMOS and nMOS transistors in a CMOS inverter are both driver transistors. Figure 4.20 shows a CMOS inverter and its transfer characteristic $V_{out} = f(V_{in})$. The gates of the pMOS (T_p) and nMOS (T_n) transistors are connected to form the inverter input. It is important to remember that $V_{T_p} < 0$ and $V_{T_n} > 0$.

The transfer characteristic is explained as follows:

T_n is 'off' and T_p is 'on' for $V_{in} < V_{T_n}$.
The output voltage V_{out} then equals V_{dd}.

T_p is 'off' and T_n is 'on' for $V_{in} > V_{dd} + V_{T_p}$.
V_{out} then equals V_{ss}.

In both of the above stable situations, one transistor is always 'off' and no DC current can flow from supply to ground. The current characteristic $I = f(V_{in})$ in figure 4.20b reflects this fact. The absence of DC current in the two stable situations is the most important advantage of CMOS when compared with nMOS. A current only flows from supply to ground during an input voltage transition, for which the following conditions apply:

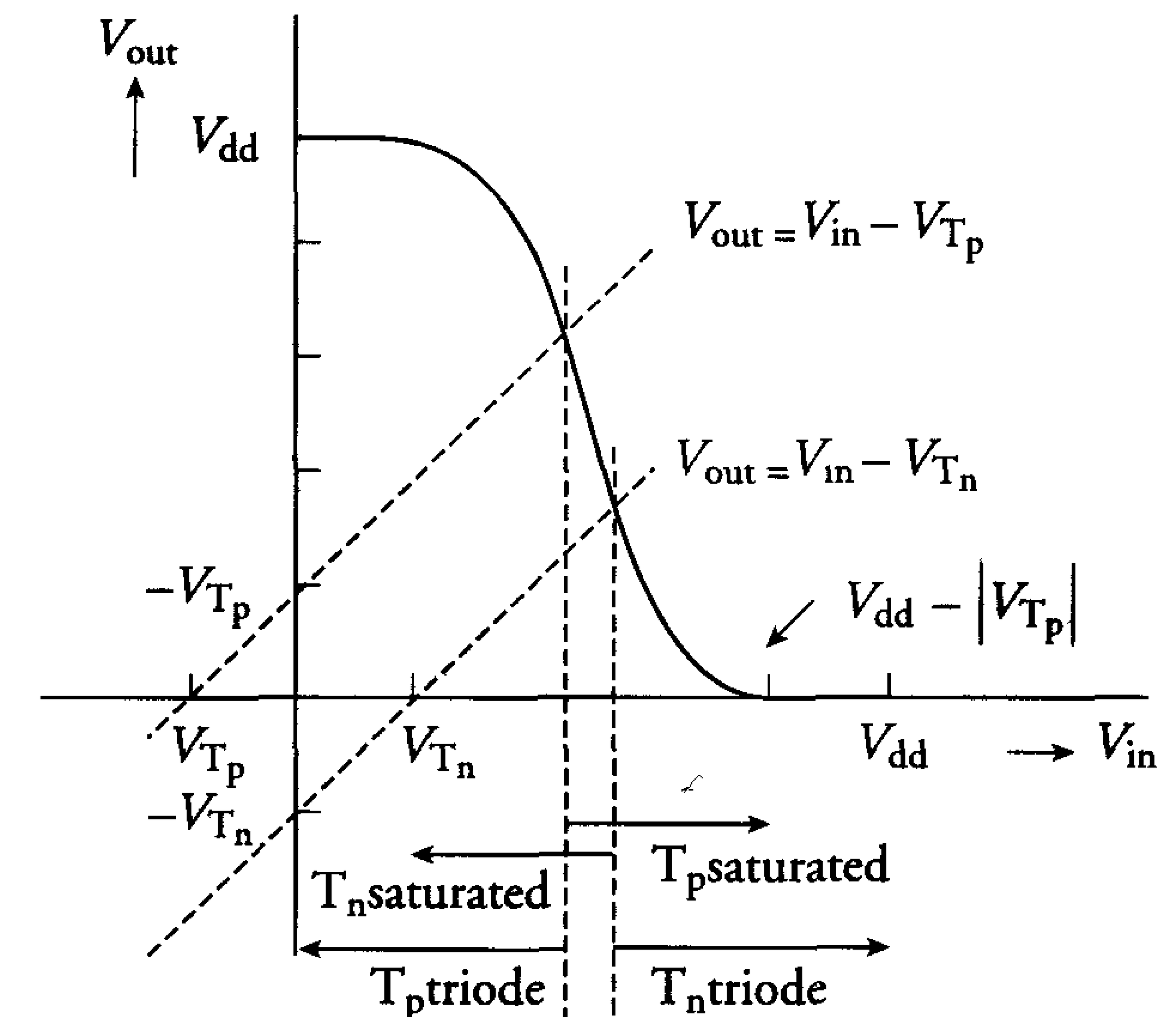
$$V_{T_n} < V_{in} < V_{dd} + V_{T_p}$$

Figure 4.20b shows the trajectory of the transient current associated with the input voltage transition from V_{ss} to V_{dd}. The areas where T_n and T_p operate in their respective saturation and triode regions are indicated in figure 4.20a.

The saturation areas are described for the two transistors as follows:

T_n: $V_{ds}|_{T_n} > V_{gs} - V_{T_n}$ i.e. $V_{out} > V_{in} - V_{T_n}$. This is the area above the dotted line $V_{out} = V_{in} - V_{T_n}$ in the transfer characteristic.

T_p: $V_{ds}|_{T_p} < V_{gs} - V_{T_p}$ i.e. $V_{out} - V_{dd} < V_{in} - V_{dd} - V_{T_p}$. This is the area below the dotted line $V_{out} = V_{in} - V_{T_p}$ in the transfer characteristic.



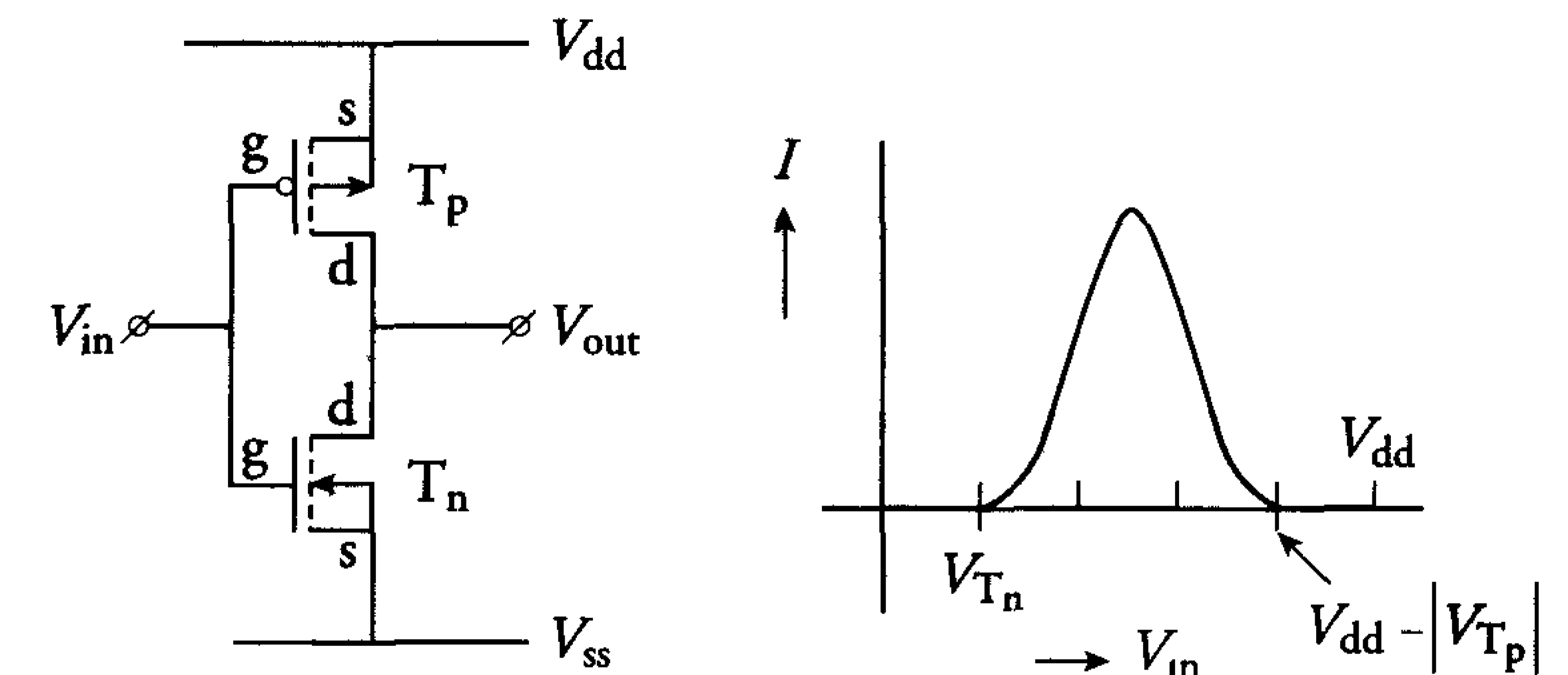


Figure 4.20: *Transfer characteristic (a) and current characteristic (b) of a MOS inverter*

Figure 4.20 shows that the transistors in an inverter are both saturated during transitions between logic levels. Theoretically, their output impedances are then infinite. Application of Ohm's Law reveals that a finite current should then cause an infinitely large change in the output

voltage. In practice, the output impedances are always finite and the maximum voltage change is limited. However, the transfer characteristic of the CMOS inverter is still very steep.

It must be noted that figure 4.20 is drawn on the basis of the assumptions that $V_{T_n}=-V_{T_p}$ and $V_{dd}>V_{T_n}+|V_{T_p}|$. The reader should verify that the transfer characteristic of the inverter displays hysteresis when $V_{T_n}+|V_{T_p}|>V_{dd}$.

The charging and discharging behaviour of a CMOS inverter can also be described by means of the static characteristic $I = f(V_{out})$ shown in figure 4.21. This characteristic is obtained when a pulse V_{in} with rise and fall times of 0 ns is applied at the inverter input. Capacitance C is the load capacitance present at the transistor's output. The currents through the pMOS and nMOS transistors are I_p and I_n, respectively.

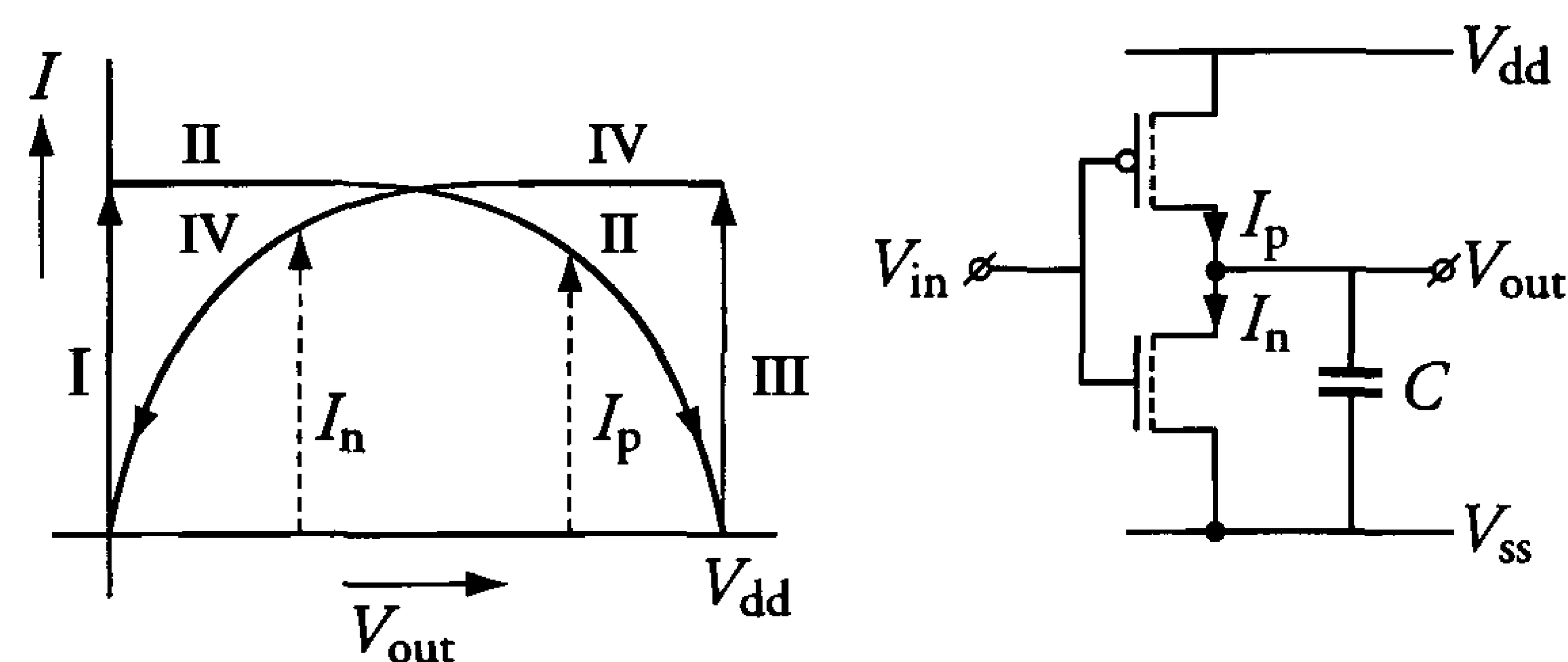


Figure 4.21: *Static CMOS-inverter characteristic*

The curves in figure 4.21 are explained as follows:

Trajectory I : I_p rises from 0 to $I_{p_{max}}$ when V_{in} falls from V_{dd} to V_{ss}.
Trajectory II : C charges to V_{dd} and I_p decreases to 0.
Trajectory III : I_n rises from 0 to $I_{n_{max}}$ when V_{in} rises from V_{ss} to V_{dd}.
Trajectory IV : C discharges to V_{ss} and I_n decreases to 0.

In figure 4.21, it is assumed that the βs and the V_Ts of the nMOS and pMOS transistors are equal. The current characteristics are therefore symmetrical with respect to $V_{out} = \frac{1}{2}V_{dd}$.

Designing a CMOS inverter

A true CMOS logic gate contains a pMOS transistor for every nMOS transistor. A *pseudo-nMOS* version, however, uses just one active pull-up pMOS transistor with its gate connected to ground. Here, a DC current flows from supply to ground when the output is 'low'. The complementary behaviour of the transistors in true CMOS circuits ensures the absence of DC current at both the low and high stable operating points. This type of CMOS logic is therefore 'ratioless' and the voltages V_H and V_L associated with the respective 'high' and 'low' output levels are independent of the transistor sizes. In fact, V_H equals the supply voltage V_{dd} while V_L equals 0 V. The dynamic discharge characteristic of a CMOS inverter is obtained when a step voltage (which rises from 0 V to V_{dd} in 0 ns) is applied to its input. This is illustrated in figure 4.22. As shown in figure 4.23, the dynamic charge characteristic is obtained when the input step voltage falls from V_{dd} to 0 V in 0 ns.

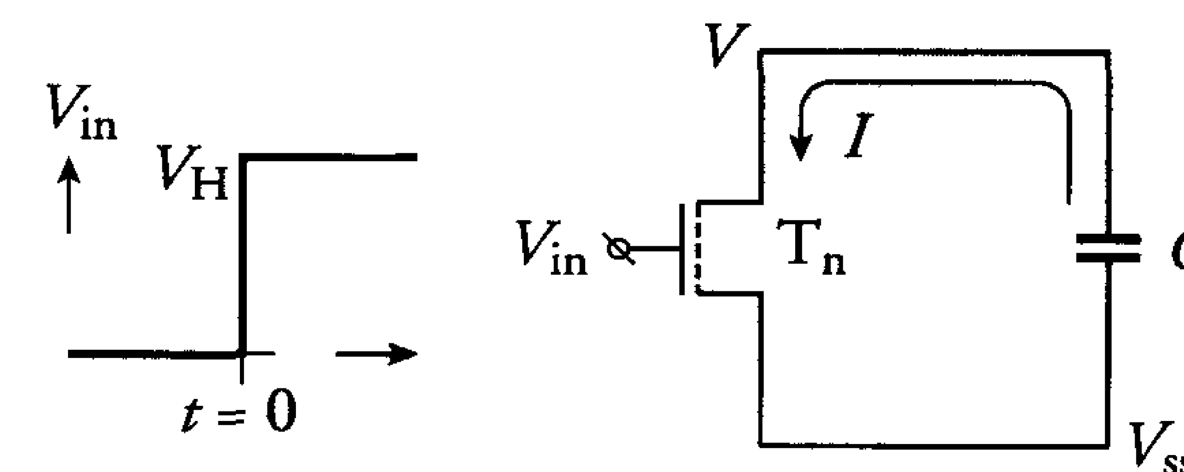


Figure 4.22: *Discharging a load capacitance through an nMOS transistor*

The complementary behaviour of the pMOS and nMOS transistors produces charging and discharging characteristics which are similar to those described for nMOS in section 4.2.3. The section also outlines the derivation of an expression for the gain factor β_n of an nMOS transistor which will discharge a capacitance C from V_{dd} to V in time t when a step voltage with amplitude V_{dd} is applied to its gate at $t = 0$. The formula is simply repeated here:

$$\beta_n = \frac{4 \cdot C}{V_{dd} \cdot t} \qquad (4.12)$$

The required dimensions of the nMOS transistor are obtained by equating the gain factor β_n to $\beta_{\square n} \cdot A_{T_n}$, where A_{T_n} is the aspect ratio of the transistor and equals $(W/L)_{T_n}$. Because of second-order effects described in chapter 2, μ_n and hence $\beta_{\square n}$ decrease for sub-micron channel

lengths. In this case, the calculated value of W must be increased in proportion to the decrease in $\beta_{\square n}$.

Example:
Given: A 0.25 μm CMOS process with $\beta_{\square n}$=240 μA/V^2.
Problem: Determine the aspect ratio A_{T_n} of an nMOS transistor T_n which will discharge a load capacitance C=100 fF from V_{dd} to $0.1 \cdot V_{dd}$ in 1 ns when a voltage V_{dd} is applied to its gate.
Solution: Substituting in (4.12) yields:

$$\beta_n = \frac{4 \cdot 100 \cdot 10^{-15}}{2.5 \cdot 10^{-9}}$$

Equating β_n to $\beta_{\square n} \cdot A_{T_n}$ and substituting $\beta_{\square n}$=240 μA/V^2 yields:

$$A_{T_n} = \left(\frac{W}{L}\right)_{T_n} = \frac{2}{3}$$

Because of second-order effects (chapter 2), which are not included in the simple basic current equations, the current capability of transistors in a 0.25 μm CMOS process is reduced by about a factor of 2. This means that the required $\left(\frac{W}{L}\right)$ for the nMOS transistor must be:

$$\left(\frac{W}{L}\right)_{T_n} = 1.3$$

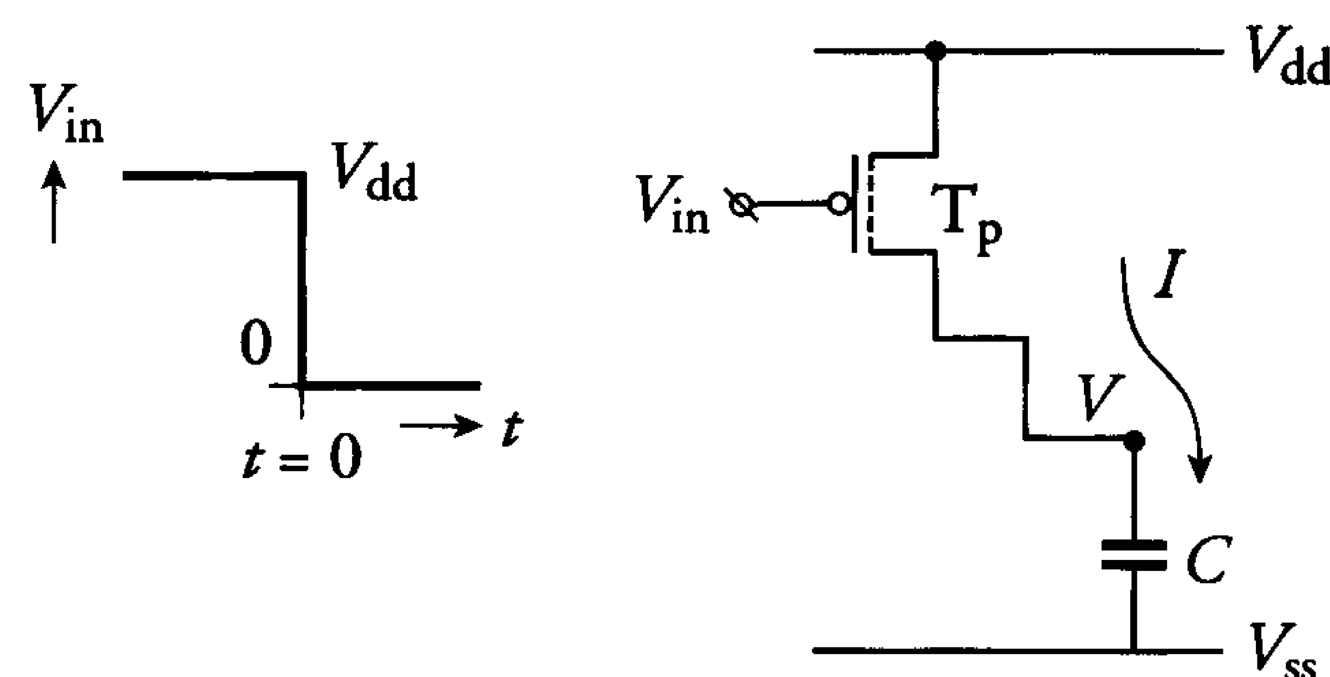


Figure 4.23: *Charging a load capacitance through a pMOS transistor*

The charging of a load capacitance through a pMOS transistor illustrated in figure 4.23 is analogous to discharging through an nMOS transistor. The expression for the gain factor β_p of a pMOS transistor, which

will charge a capacitance C from 0 V to a voltage $V = 0.9 \cdot V_{dd}$ in time t when its gate voltage falls from V_{dd} to 0 V in 0 ns is, therefore simply obtained by the same equation (4.12).

Example:
Given: The information in the previous example plus $\beta_{\square p}$=60 μA/V^2.
Problem: Determine the aspect ratio A_{T_p} of a pMOS transistor T_p which will charge the load capacitance C from 0 V to $0.9 \cdot V_{dd}$ in 1 ns when 0 V is applied to its gate.
Solution: This problem is the complement of the previous example.

Therefore, the following expression applies (see equation 4.11):

$$A_{T_p} = A_{T_n} \cdot A = 4$$

The rise and fall times of *buffer circuits* must be equal. These circuits must therefore use the previously-mentioned value of about 3 for the aspect ratio A expressed in formula (4.11). For CMOS logic, however, values for A of 1.5 to 2 are currently used. Larger values yield larger pMOS transistors and thus increase the load capacitance presented to previous logic gates. For CMOS circuits other than inverters, factors $(\frac{W}{L})_p$ and $(\frac{W}{L})_n$ in formula (4.11) are the effective values which apply to the transistors in the p and n sections, respectively. The dimensions of these transistors must be selected so that the value for A is optimal.

Dissipation of a CMOS inverter

During the last two decades, CMOS technology has become the most dominant technology for VLSI circuits. The most important reason for this is its low static power consumption. This is because of the absence of DC currents during periods when no signal transients occur in static CMOS circuits. However, a *short-circuit current* flows from supply to ground when a change in a logic circuit's input voltage causes the output voltage V_{out} to change. This short-circuit current leads to power dissipation [12].

The *power dissipation* of a basic CMOS inverter is explained with the aid of figure 4.24.

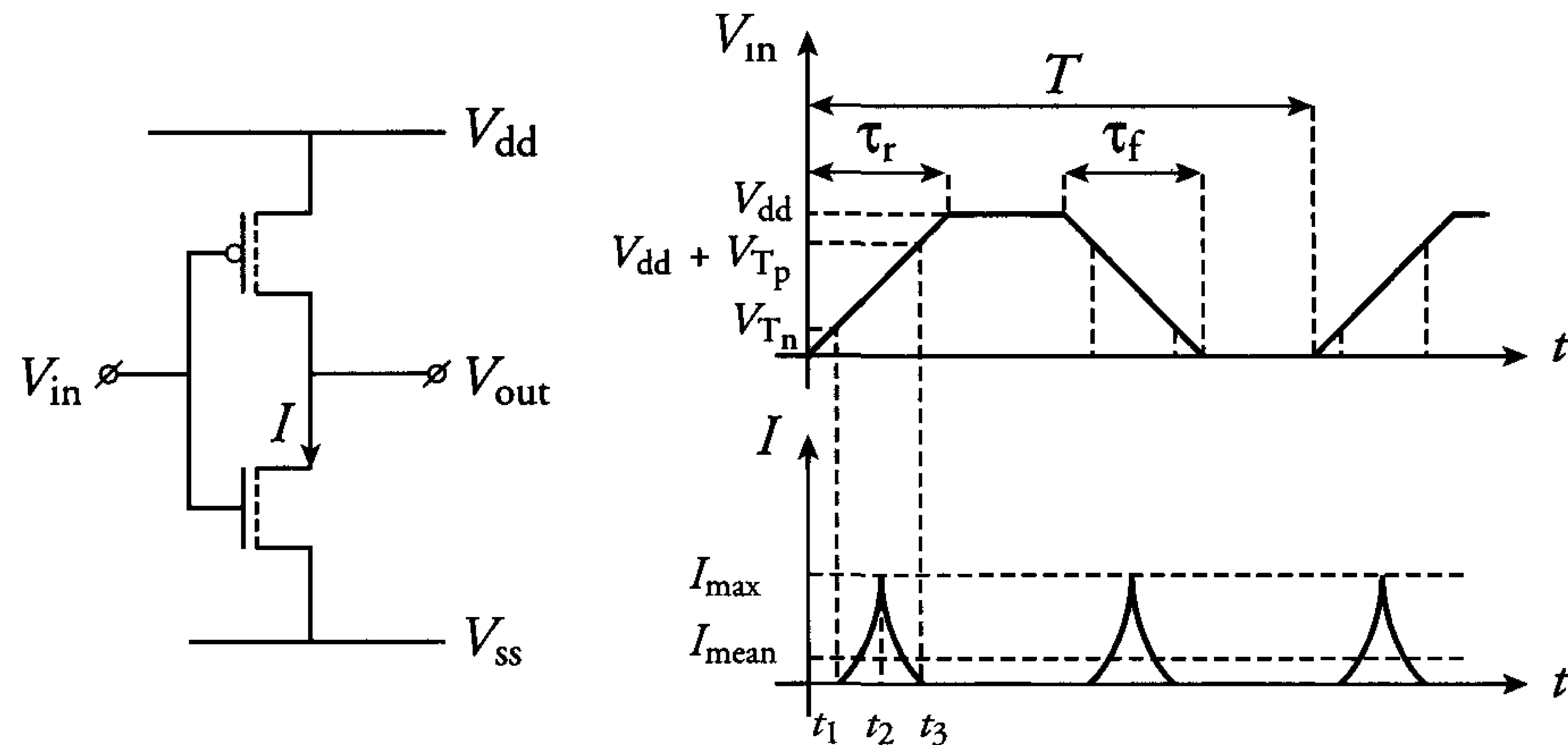


Figure 4.24: *Current through an unloaded inverter*

Only the nMOS transistor conducts when the input voltage V_{in} of this static CMOS inverter is 'high' (V_{dd}). Similarly, only the pMOS transistor conducts when the input voltage V_{in} is 'low' (V_{ss}). Therefore, the inverter does not dissipate power when the input is in either of the above stable states. However, during a transient at the input, there is a period when both the nMOS and pMOS transistors conduct. A short-circuit current then flows from supply to ground while the input voltage is between V_{T_n} and $V_{dd} - |V_{T_p}|$. This current I is shown in figure 4.24. If a load capacitance C_L is connected to the inverter output, then the dissipation consists of two components:

1. *Dynamic power dissipation*:

$$P_1 = C_L \cdot V^2 \cdot f \tag{4.13}$$

2. *Short-circuit power dissipation*:

$$P_2 = I_{mean} \cdot V \tag{4.14}$$

In the above equations, f $(= 1/T)$ is the frequency at which the voltage change V occurs on C_L and I_{mean} is the average short-circuit current. Clearly, the dynamic component P_1 is independent of transistor dimensions when parasitic capacitances at the output, such as pn-junction capacitances, are neglected. It is expressed in equation 4.13 and is explained with the aid of figure 4.25.

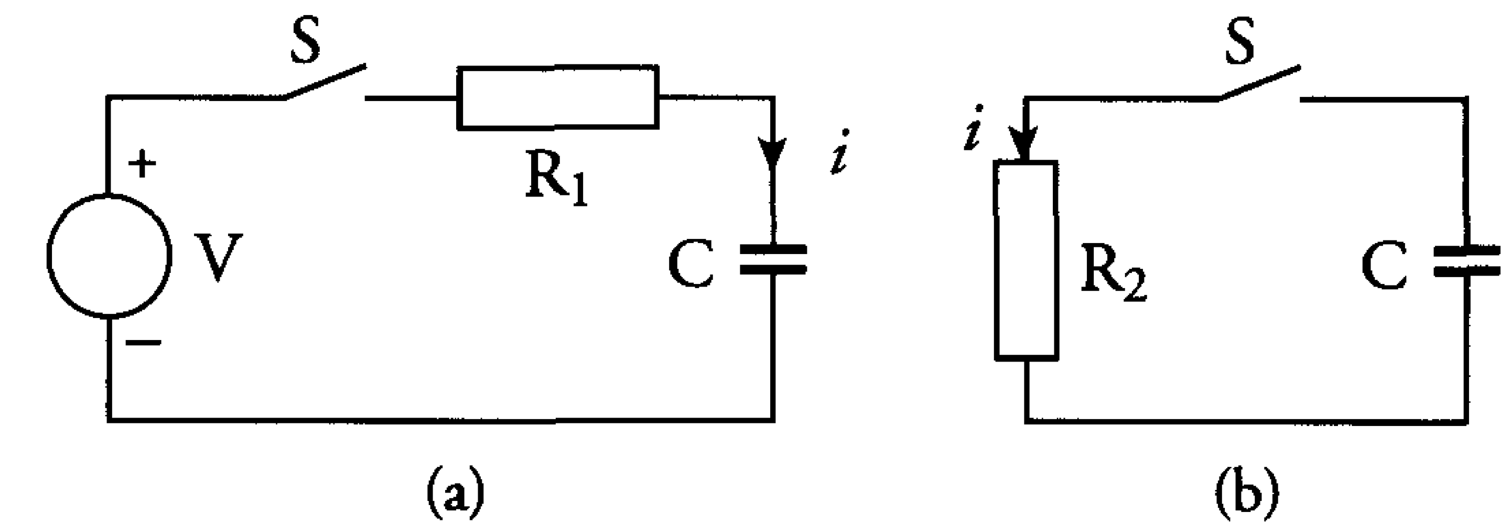


Figure 4.25: *Charging and discharging a capacitance*

Capacitance C is charged and discharged via resistors with values R_1 and R_2, respectively. During charging, the power dissipation in R_1 equals:

$$P_{R_1} = \int_0^\infty i^2(t) \cdot R_1 \cdot dt \qquad \text{with } i(t) = \frac{V}{R_1} \cdot e^{-t/(R_1 C)}$$

The solution to this integral is as follows:

$$P_{R_1} = \frac{1}{2} \cdot C \cdot V^2$$

P_{R_1} is thus independent of R_1. Similarly, the power dissipation P_{R_2} during discharging is independent of the value of R_2 and also equals $C \cdot V^2/2$.

The total power P supplied by the voltage source V during a complete charge-discharge cycle is the sum of P_{R_1} and P_{R_2}, i.e. $P = C \cdot V^2$. For f cycles per second the total power dissipation is:

$$P = C \cdot V^2 \cdot f$$

This dynamic power dissipation appears in all types of logic, including static MOS circuits, bipolar circuits, TTL circuits, etc.

The short-circuit component P_2, however, is proportional to transistor dimensions; it also depends on the size of the load capacitance. An expression for I_{mean} in formula (4.14) is derived on the assumption that the inverter's load capacitance is zero [12]. Although an asymmetric inverter is not fundamentally different, the inverter is also assumed to be symmetric. In this case, the following equations apply:

$$\beta_n = \beta_p = \beta \text{ and } V_{T_n} = -V_{T_p} = V_T$$

During the period t_1 to t_2 in figure 4.24, the short-circuit current I increases from 0 to I_{max}. Throughout this period, the output voltage

V_{out} is more than a threshold voltage V_{T_n} larger than the input voltage V_{in}. The nMOS transistor is therefore saturated and application of the simple MOS formulae (1.13) yields the following expression for I during this period of time:

$$I = \frac{\beta}{2}(V_{in} - V_{T_n})^2 \quad \text{for } 0 \leq I \leq I_{max}$$

The symmetry of the inverter produces a maximum value for I when V_{in} equals $V_{dd}/2$. In addition, the current transient during the period t_1 to t_3 is symmetrical with respect to the time t_2. The mean current I_{mean} (i.e. the effective current which flows during one cycle period T of the input signal) can therefore be expressed as follows:

$$I_{mean} = 2 \cdot \frac{2}{T} \int_{t_1}^{t_2} I(t)\mathrm{d}t = \frac{4}{T} \int_{t_1}^{t_2} \frac{\beta}{2}(V_{in}(t) - V_T)^2 \mathrm{d}t \qquad (4.15)$$

The input voltage V_{in} is assumed to have a symmetrical shape and linear edges, with rise and fall times equal to τ. The value of V_{in} as a function of time t during an edge is therefore expressed as follows:

$$V_{in}(t) = \frac{V_{dd}}{\tau} \cdot t$$

The following expressions for t_1 and t_2 can be derived from figure 4.24:

$$t_1 = \frac{V_T}{V_{dd}} \cdot \tau \quad \text{and} \quad t_2 = \frac{\tau}{2}$$

Substituting these expressions for $V_{in}(t)$, t_1 and t_2 in equation (4.15) yields:

$$I_{mean} = \frac{2\beta}{T} \cdot \int_{\tau/2}^{\frac{V_T}{V_{dd}} \cdot \tau} \left(\frac{V_{dd}}{\tau} \cdot t - V_T\right)^2 \mathrm{d}\left(\frac{V_{dd}}{\tau} \cdot t - V_T\right)$$

The solution to this equation is:

$$I_{mean} = \frac{1}{12} \cdot \frac{\beta}{V_{dd}} \cdot (V_{dd} - 2V_T)^3 \cdot \frac{\tau}{T}$$

Substituting this expression for I_{mean} into formula (4.14) yields the following expression for the short-circuit dissipation of a CMOS inverter with no load capacitance:

$$P_2 = \frac{\beta}{12} \cdot (V_{dd} - 2V_T)^3 \cdot \frac{\tau}{T} \qquad (4.16)$$

Formula (4.16) clearly illustrates that the short-circuit dissipation is proportional to the frequency $f = 1/T$ at which the input changes. Voltages V_{dd} and V_T are determined by the application and the process. Therefore, the only design parameters that affect P_2 are β and the rise and fall times (τ) of the inverter's input signal. For an inverter with a capacitive load, the transistor β values are determined by the required output rise and fall times. In this case, the short-circuit dissipation only depends on the input signal's rise and fall times, i.e. τ_r and τ_f, respectively. This is particularly true for buffer circuits which have transistors with large β values.

In the chapter on low-power design (chapter 8), the CMOS power contributions are discussed extensively. However, the design of large buffer circuits is discussed in this section on basic CMOS circuit design.

CMOS buffer design

Large capacitances associated with integrated circuits include those presented by bus lines and 'off-chip' circuits. These capacitances must often be driven at high frequencies. The required 'buffer' driving circuits dissipate a relatively large part of the total power consumed by a chip. Optimising these buffers therefore requires considerably more effort than the approach adopted for CMOS logic. Formula (4.16) shows that short-circuit dissipation is directly proportional to the rise and fall times (τ) of an input signal. The input signals of buffers which drive bus lines connected to large numbers of different sub-circuits on a chip must therefore have particularly short rise and fall times.

Suppose the signal on a bus line with capacitance C_L must follow a signal at the output node A of a logic gate which is capable of charging and discharging a capacitance C_0 in τ ns. An inverter chain such as illustrated in figure 4.26 can be used as a buffer circuit between node A and the bus line.

Figure 4.26: *A buffer circuit comprising an inverter chain*

From formula (4.16), it is clear that the rise and fall times on each input of the inverters in the above chain should be short. Moreover, it has been shown in literature that minimum dissipation can be achieved when the rise and fall times on each of these inputs are equal to the rise and fall times at the buffer output. The inverter chain must therefore be designed so that the rise and fall times on the inputs of each of its component inverters are τ ns. According to literature [4], a minimum propagation delay time for the buffer is obtained when the 'tapering factor' between the βs of successive inverters is e, the base of natural logarithm. In terms of dissipation and silicon area, however, this will not lead to an optimum design. Design optimisation for minimum dissipation and silicon area requires a different approach [9].

Example:
A signal is produced by a logic gate and must be buffered to drive a capacitive load $C_L = 10\text{pF}$. The required rise and fall time τ at the buffer output is 1 ns. The channel length of both the pMOS and the nMOS transistor is $0.25\,\mu$m. $\beta_{\Box n} = 240\,\mu\text{A/V}^2$ and $\beta_{\Box p} = 60\,\mu\text{A/V}^2$.

The design procedure is as follows:

- Define β_{nN} of the last driver according to formula (4.12)

  $\beta_{nN} = \frac{2 \cdot 10 \cdot 10^{-12}}{2.5 \cdot 10^{-9}} = 8 \cdot 10^{-3}\,\text{A/V}^2$

  $\frac{W}{L} \cdot \beta_{\Box n} = \beta_{nN}$ and $\beta_{\Box n} = 240\,\mu\text{A/V}^2$ lead to: $\left(\frac{W}{L}\right)_{nN} = 33$

- Take second-order effects (chapter 2) into account. With respect to the transistor current capability, the simple equations (1.17) from which (4.12) is derived are much too optimistic. Chapter 2 discusses second-order effects such as narrow and short channel effects. For a $0.25\,\mu$m CMOS process, these second-order effects may reduce the current capability by a factor as high as about 2. Therefore, the required transistor needs to have a $\left(\frac{W}{L}\right)_{nN} = 33 \times 2 = 66$ to do the job.

- Determine the right tapering factor in the inverter chain. Figure (4.27) shows three different inverter chains with different tapering factors and with the required $\left(\frac{W}{L}\right)_{nN}$ ratio as determined above.

Figure 4.27: *Inverter chains with different tapering factors, all driving the same load with equal rise times*

- The performance of these inverter chains can be expressed with the variables: power dissipation, propagation delay, maximum current change $\frac{dI}{dt}$ and area. Figure (4.28) shows the simulation results of these inverter chains.

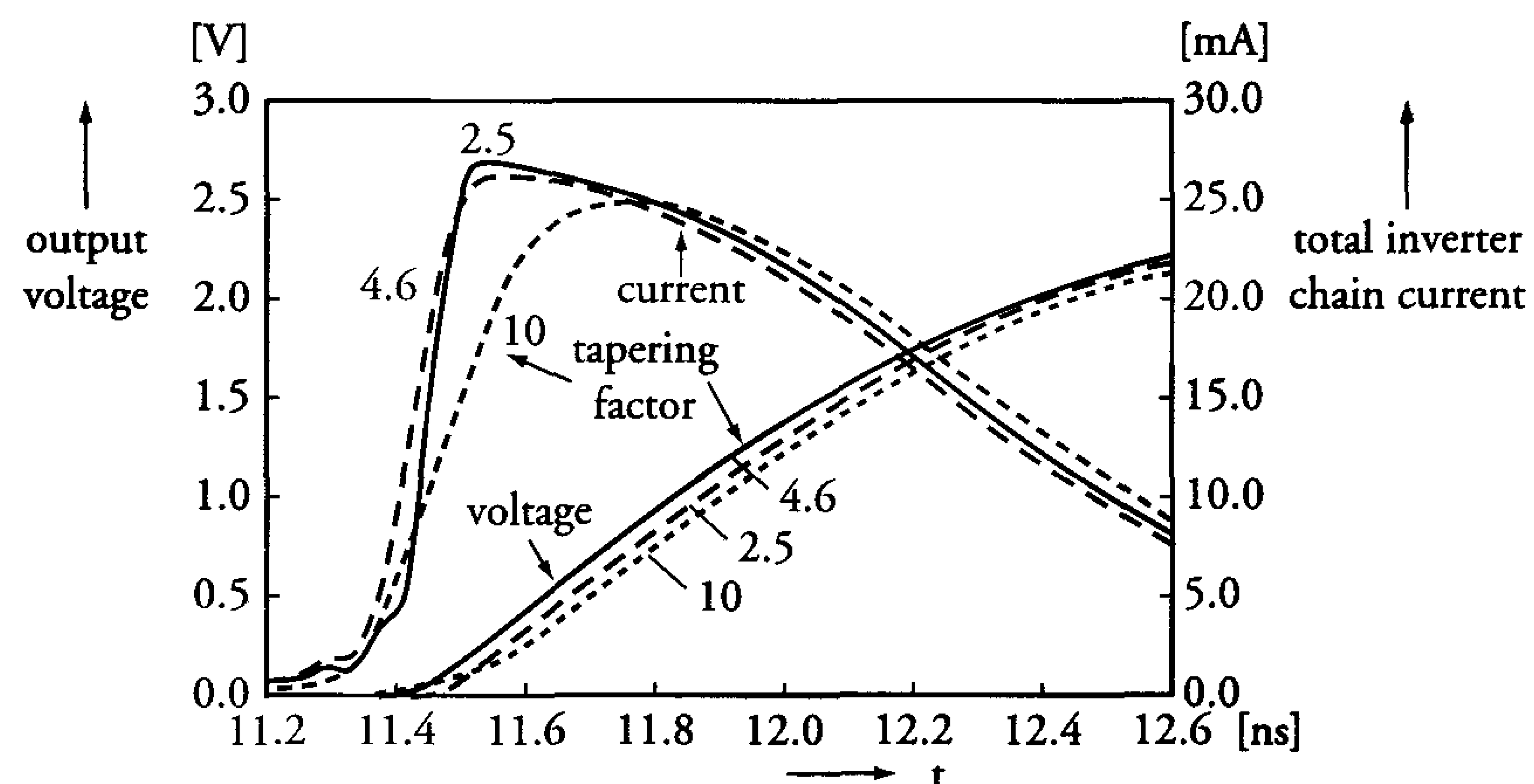


Figure 4.28: *Simulation results for inverter chains, with different tapering factors*

All inputs, V_{in}, are connected to an equivalent logic gate with the same effective $\frac{W}{L}$ ratio as the first inverter of circuit 1 (and 2). The diagram shows the input signals (V_{in}), the supply currents (I_{supply}) and the output signals (V_{out}). Detailed overall results for these circuits are given in table 4.1.

Table 4.1: Comparison of inverter chain buffers with different tapering factors

Circuit number	**1**	**2**	**3**	
tapering factor	2.5	4.6	10	
number of inverters	6	4	3	
total power (relative)	1.14	1.11	1	
total area (relative)	1.55	1.21	1	
max $\frac{dI}{dt}$ (relative)	4.6	3	1	
(absolute)	$2.8{\cdot}10^8$	$1.8{\cdot}10^8$	$0.6{\cdot}10^8$	[A/s]
propagation delay (relative)	0.98	0.94	1	
(absolute)	0.92	0.88	0.94	[ns]

The tapering factor e (close to 2.5), which is derived in literature [4] to achieve minimum propagation delay, scores very badly with respect to the maximum $\frac{dI}{dt}$ and to the area. As the maximum level of $\frac{dI}{dt}$ is a very important figure that determines noise (see chapter 9), it should be as low as possible, but without deteriorating the performance too much. From the table, we can conclude that circuit 3, with a tapering factor of 10, yields optimum general performance. In summary, overall circuit performance (power, delay, area, noise), is better when a CMOS buffer is optimised for minimum power dissipation rather than for minimum propagation delay. Generally speaking, tapering factors between 8 and 16 should be applied in practical 0.25 μm CMOS buffer designs.

Noise margins

The maximum amplitude of a noise signal that can be superimposed on all nodes of a long inverter chain without causing the output logic level to change is called *noise margin*. Figure 4.29 shows the transfer characteristic of a CMOS inverter for three different gain factor ratios. The noise margins for both high and low levels are very large because of the almost rectangular shape of these transfer characteristics. For the symmetrical inverter, with $\beta_{\text{n}}=\beta_{\text{p}}$ and $V_{\text{T}_\text{n}}=-V_{\text{T}_\text{p}}$, the noise margins are equal for both levels. Of course, not every inverter is symmetrical. In such cases, the noise margin is different for the two levels. However, the difference is only significant for highly asymmetrical inverters.

Generally, the following equation applies:

$$\text{Noise margin} \geq 0.42 \cdot V_{dd} \tag{4.17}$$

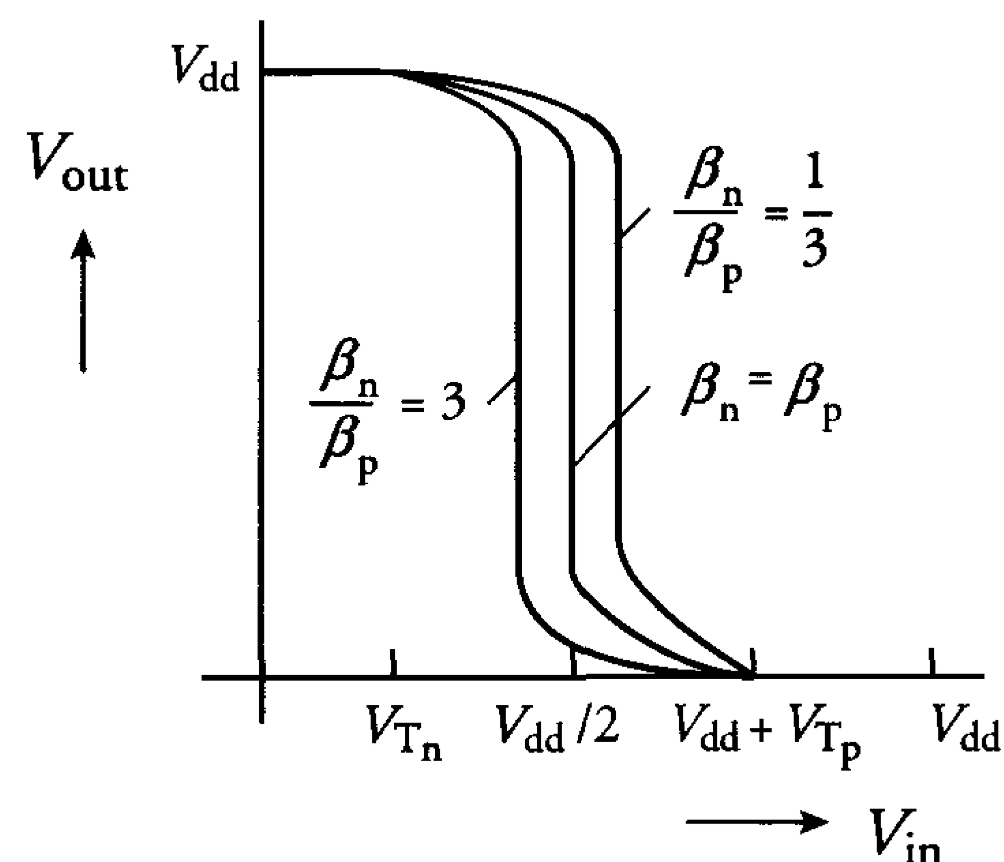


Figure 4.29: *CMOS inverter transfer characteristics for different aspect ratios*

4.4 Digital CMOS circuits

4.4.1 Introduction

CMOS circuits can be implemented in static or dynamic versions. The choice is mainly determined by the type of circuit and its application. Two important factors which influence the choice are chip area and power dissipation. The differences between these factors for the two types of implementation are treated in this section.

4.4.2 Static CMOS circuits

A logic function in static CMOS must be implemented in both nMOS and pMOS transistors. An nMOS version only requires implementation in nMOS transistors. A single load transistor is then used to charge the output. This load transistor also conducts when the output is 'low'. A current therefore flows from supply to ground and causes DC dissipation while the output of an nMOS logic gate is 'low'. In a CMOS logic gate, a current only flows between supply and ground during output transitions.

Figure 4.30 shows some static CMOS implementations of logic gates. Back-bias connections for both the nMOS and the pMOS transistors are indicated in the inverter in figure 4.30(a). The respective back-bias voltages, V_{sb} and V_{ws}, are both 0 V. The back-bias connections are no longer shown in figures 4.30(b), 4.30c and all subsequent figures. Unless otherwise stated, the substrate voltages are assumed to be V_{ss} for the nMOS transistors and V_{dd} for the pMOS transistors. Figures 4.30(b) and 4.30(c) show nMOS and pMOS transistors, respectively, connected in series. The sources of some of these transistors are not connected to V_{ss} or V_{dd}. The back-bias effect has a considerable influence on nMOS and pMOS transistors whose sources are not connected to V_{ss} and V_{dd}, respectively. This is particularly true when the source is loaded.

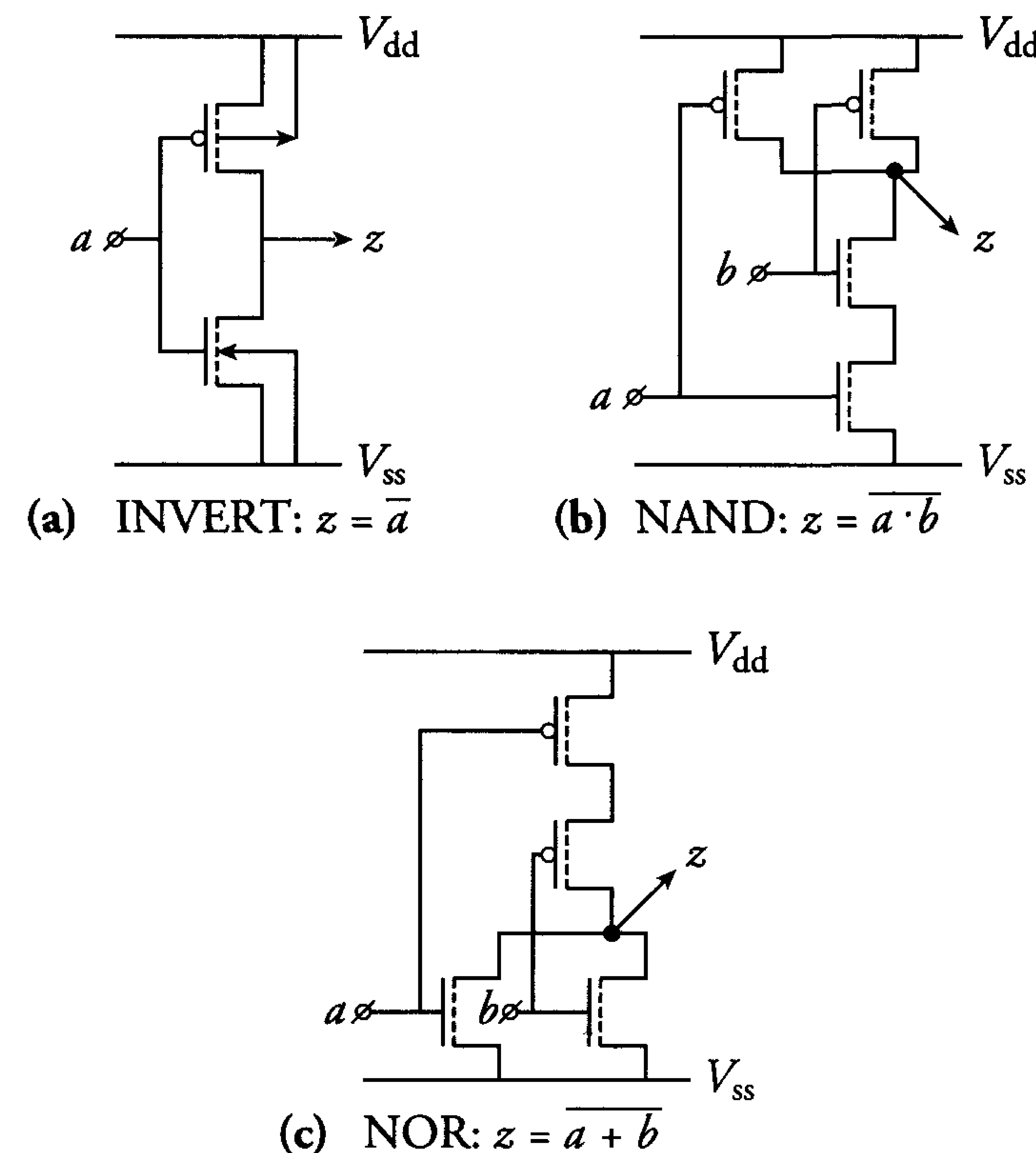


Figure 4.30: *Examples of static CMOS logic gates*

In general, a series connection of transistors in the nMOS section of a CMOS logic gate will reflect a parallel connection of transistors in the

pMOS section and vice versa. This is illustrated in figure 4.31, which shows an example of a static CMOS implementation of a complex logic function and its equivalent logic gate diagram.

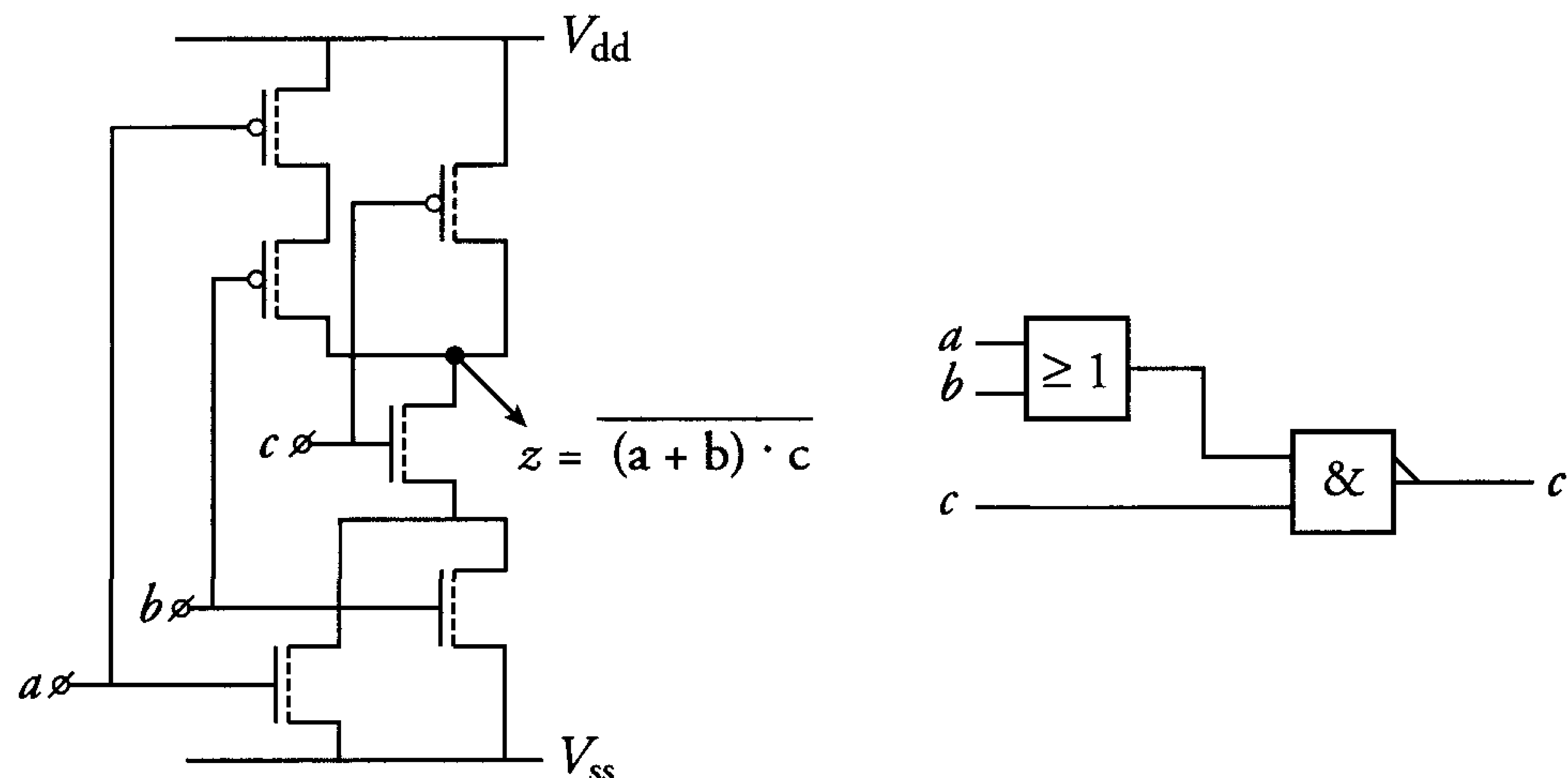


Figure 4.31: *An example of a more complex static CMOS logic gate*

The performance of a pMOS transistor is considerably poorer than that of an nMOS transistor. The number of pMOS transistors in series in a CMOS logic gate should therefore be minimised. If this number becomes very large then, only in exceptional cases, can a pseudo-nMOS implementation be used.

Figure 4.32 is an example of a pseudo-nMOS implementation of the CMOS equivalent in figure 4.31. The pseudo-nMOS version is identical to its nMOS counterpart except that the nMOS load element is replaced by a pMOS transistor with its gate connected to V_{ss}. Both nMOS and pseudo-nMOS logic gates have the advantage of the same low input capacitance. The output rise time of a pseudo-nMOS logic gate is determined by only one pMOS transistor and should therefore be short. A disadvantage of such a gate is the static power dissipation when the output is 'low'. This type of gate must therefore be used sparingly. The output low level and noise margins are determined by the ratio of the widths of the nMOS and pMOS transistors. Pseudo-nMOS logic is therefore also a form of ratioed logic, as discussed in section 4.2.2.

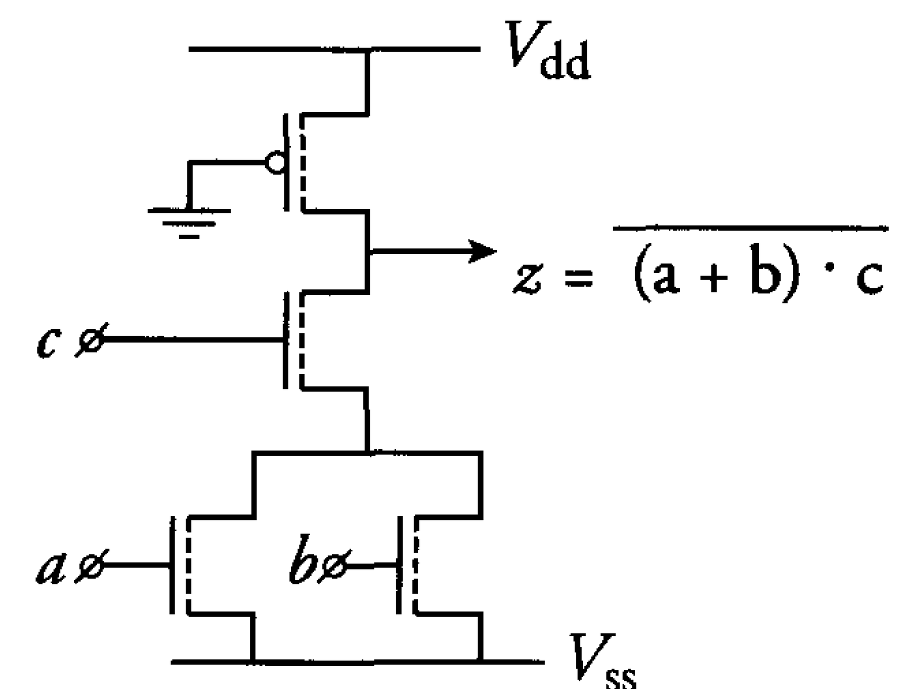


Figure 4.32: *A pseudo-nMOS logic gate*

The CMOS transmission gate (pass transistor)

Figure 4.33 shows a *transmission gate* comprising a complementary pair of transistors. This is an important component in both static and dynamic circuits. It is used to control the transfer of logic levels from one node to another when its control signals are activated. A single nMOS enhancement transistor can also be used to implement a transmission gate. Such an implementation has only one control signal but is disadvantaged by threshold loss. The threshold voltage of the transistor may be relatively high because of the body effect and the maximum high output level equals a threshold voltage below the control voltage. For this reason, the CMOS implementation is preferred.

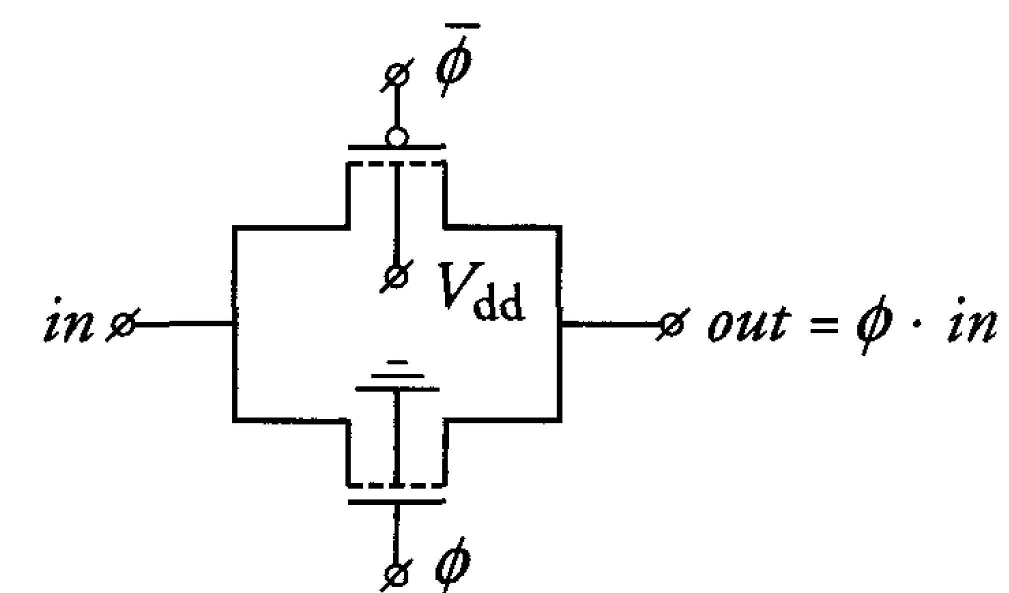


Figure 4.33: *CMOS transmission gate*

If the gate of the nMOS transistor in the CMOS transmission gate is controlled by a signal ϕ, the gate of the pMOS transistor must be controlled by the complementary signal $\overline{\phi}$. When the input voltage is 0 V

and ϕ is 'high', the output will be discharged to 0 V. The complementary behaviour of the pMOS transistor ensures that the output voltage is 2.5 V when the input voltage is 2.5 V and $\overline{\phi}$ is 'low'.

Figure 4.34 shows the contributions of both MOS transistors to the charge and discharge characteristics of a CMOS transmission gate. The pMOS and nMOS transistors prevent threshold loss on the output 'low' and 'high' levels, respectively.

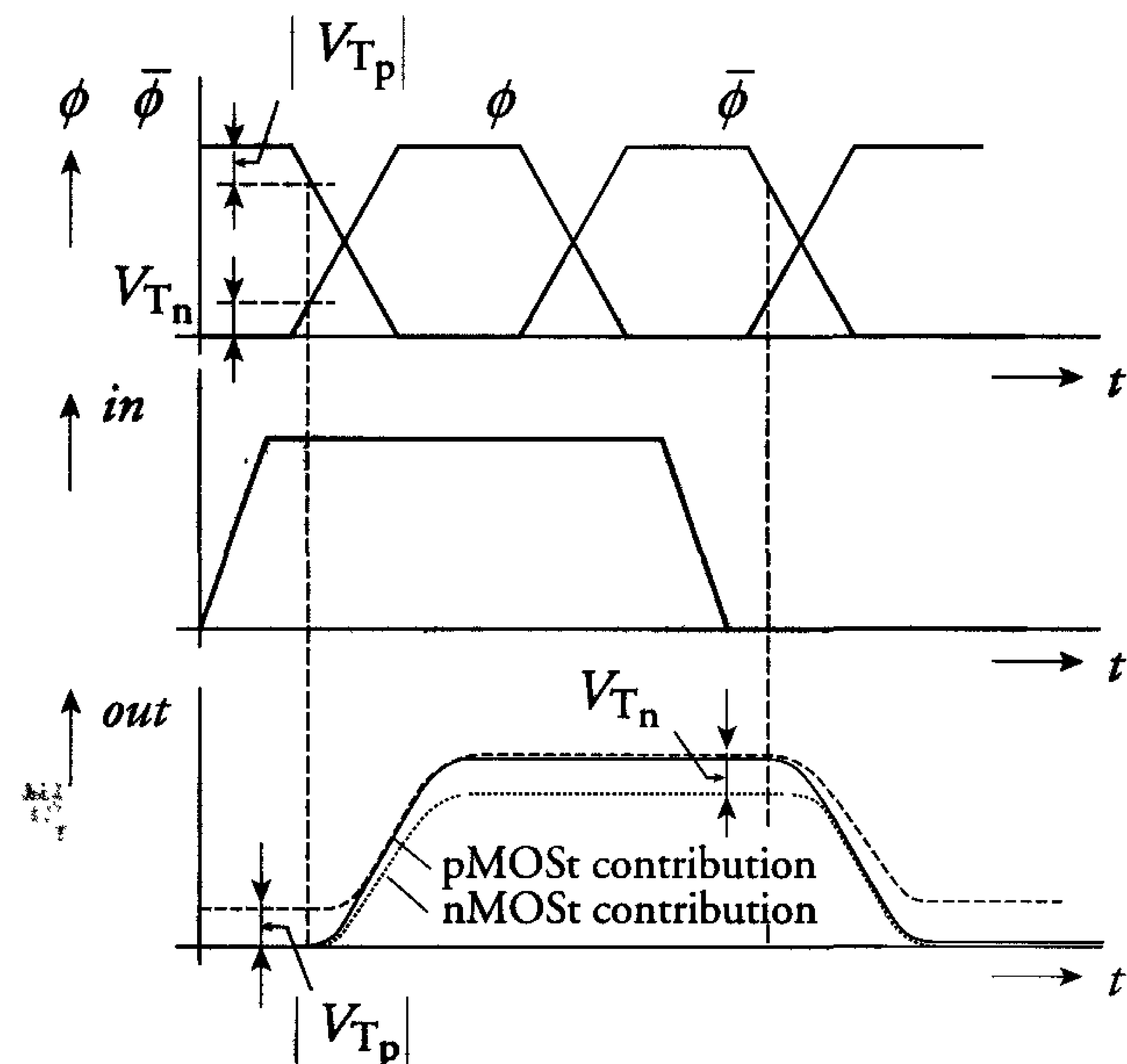


Figure 4.34: *CMOS transmission gate behaviour and the individual contributions of the nMOS and pMOS transistors to the charge and discharge characteristics*

Pass-transistor logic

In static CMOS circuits, transmission gates are used in latches, flip-flops and in 'pass-transistor logic'. Examples of pass-transistor logic are exclusive OR (EXOR) logic gates and multiplexers. Figure 4.35 shows pass-transistor logic implementations of an EXOR gate. The nMOS transmission gate implementation in figure 4.35(a) is disadvantaged by high threshold loss resulting from body effect. The complementary implementation in figure 4.35(b) yields shorter gate delays at the expense of larger chip area. When connecting the outputs of these gates to a latch circuit (e.g. two cross-coupled pMOS loads), a static CMOS logic family is created (figure 8.15). The threshold voltage loss over the nMOS pass gates is compensated by the level restoring capability of the latch.

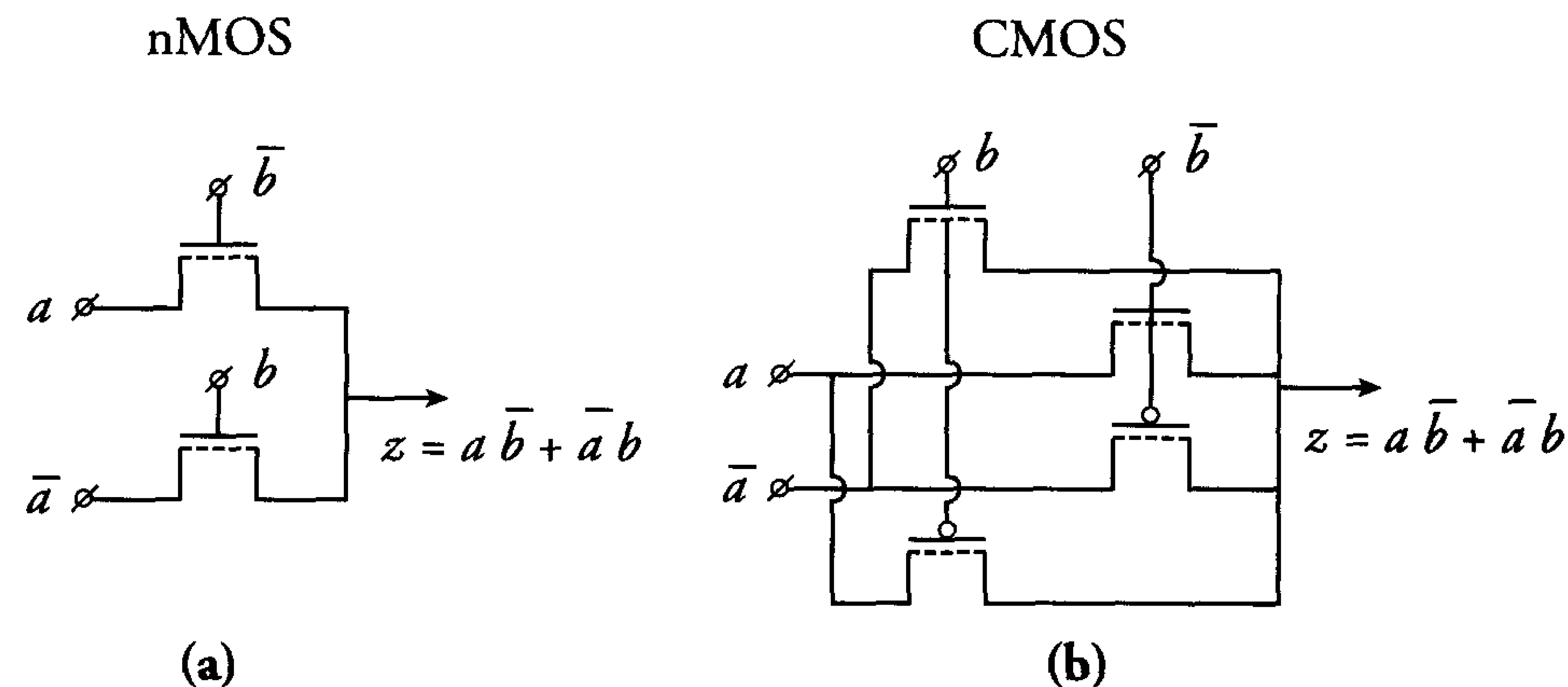


Figure 4.35: *Pass-transistor logic implementations of an EXOR logic gate with (a) nMOS pass transistors (b) CMOS pass-transistor gates*

A general disadvantage of pass-transistor logic as presented in figure 4.35 is the series resistance between the inputs a and $\overline{a}$ and the output z. The charging and discharging of a load at the output through the pass transistor causes additional delay. Other disadvantages include the need for complementary control signals. The potentials of pass-transistor logic challenge the creativity of the designers. Several alternatives have been published. These are discussed in detail in the low-power chapter 8, together with their advantages and disadvantages.

Finally, circuit designs implemented with pass-transistor logic must be simulated to prevent unexpected performance degradation or even erroneous behaviour caused by effects such as charge sharing (section 4.4). With decreasing voltages in current and future processes, the performance of pass-transistor logic tends to drop with respect to standard static CMOS logic. Therefore, the importance and existence of pass-transistor logic is expected to decrease in the coming years. The forms of CMOS logic discussed above can be used in both asynchronous circuits and synchronous, or 'clocked', circuits. The latter type of circuits are the subject of the next section.

4.4.3 Clocked static CMOS circuits

Signals which flow through different paths in a complex logic circuit will ripple through the circuit asynchronously if no measures are taken. It is then impossible to know which signal can be expected at a given node and time.

Controlling the data flow inside a circuit therefore requires synchronisation of the signals. Usually, this is done by splitting all the different paths into sub-paths with a uniform delay. The chosen delay is the worst case delay of the longest data ripple. In synchronous static MOS, the sub-paths are separated by means of 'latches' and/or 'flip-flops' which are controlled by means of periodic clock signals. Dynamic circuits may also use latches and flip-flops. Alternatively, data flow in dynamic circuits may be controlled by including the *clock signals* in every logic gate.

Static latches and flip-flops

Latches and flip-flops are used for temporary storage of signals. Figure 4.36 shows an example of a static *CMOS latch* and an extra transmission gate. The transmission gate on the left-hand side is an integral part of the latch; it also comprises two cross-coupled inverters. Complementary logic values can be written into this latch via the transmission gates when the clock signal is high, i.e. when $\phi = 1$ and $\overline{\phi} = 0$. Feedback in the latch ensures that these values are held when $\phi = 0$ and $\overline{\phi} = 1$. This basic principle is used in static full-CMOS memory cells and flip-flops.

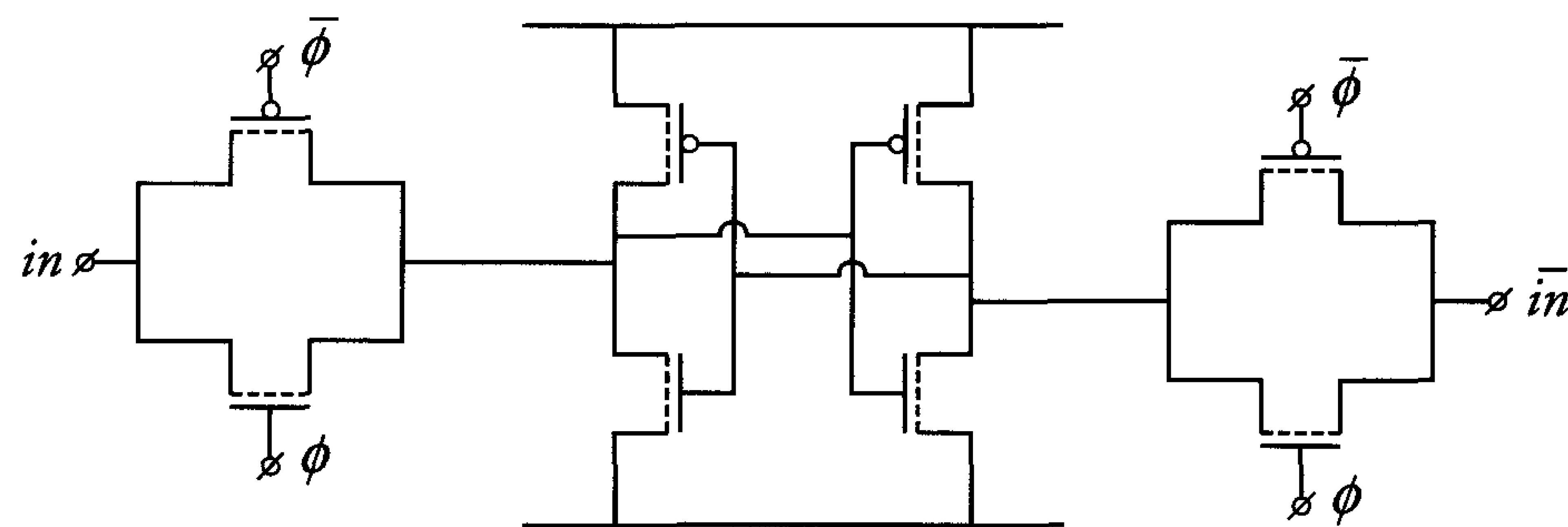


Figure 4.36: *CMOS static latch*

A flip-flop can temporarily store data and is controlled by one or more clock signals. The maximum clock frequency of a clocked static CMOS circuit is determined by the worst case delay path between two flip-flops. This path has the longest propagation delay as a result of a combination of logic gates and/or long signal tracks with large capacitances. There are several implementations of static CMOS flip-flops. The discussions below are limited to different forms of D-type flip-flops.

A *D-type flip-flop* can be built by connecting two latches in series, as shown in figure 4.37. The latches in this example use nMOS transmission gates. When the clock ϕ_1 goes 'high', data at the D input is latched into the 'master' latch of the flip-flop while the 'slave' latch stores the previous input data. The D-input has to compete with the latch's feedback inverter via the nMOS transmission gate. The W/L aspect ratios of the transistors in the feedback inverter are therefore very small. The threshold loss of the nMOS transmission gate produces a 'poor' high level at the input of the large inverter.

The aspect ratio, as expressed in equation (4.11), used for the large inverter must ensure that its output is 'low' when the poor high level is present at its input. The high level is then regenerated by the small feedback inverter. Static dissipation therefore does not occur. In practice, the aspect ratio of the large inverter must be larger than 0.5. This ensures that the inverter's switching point is lower than half the supply voltage.

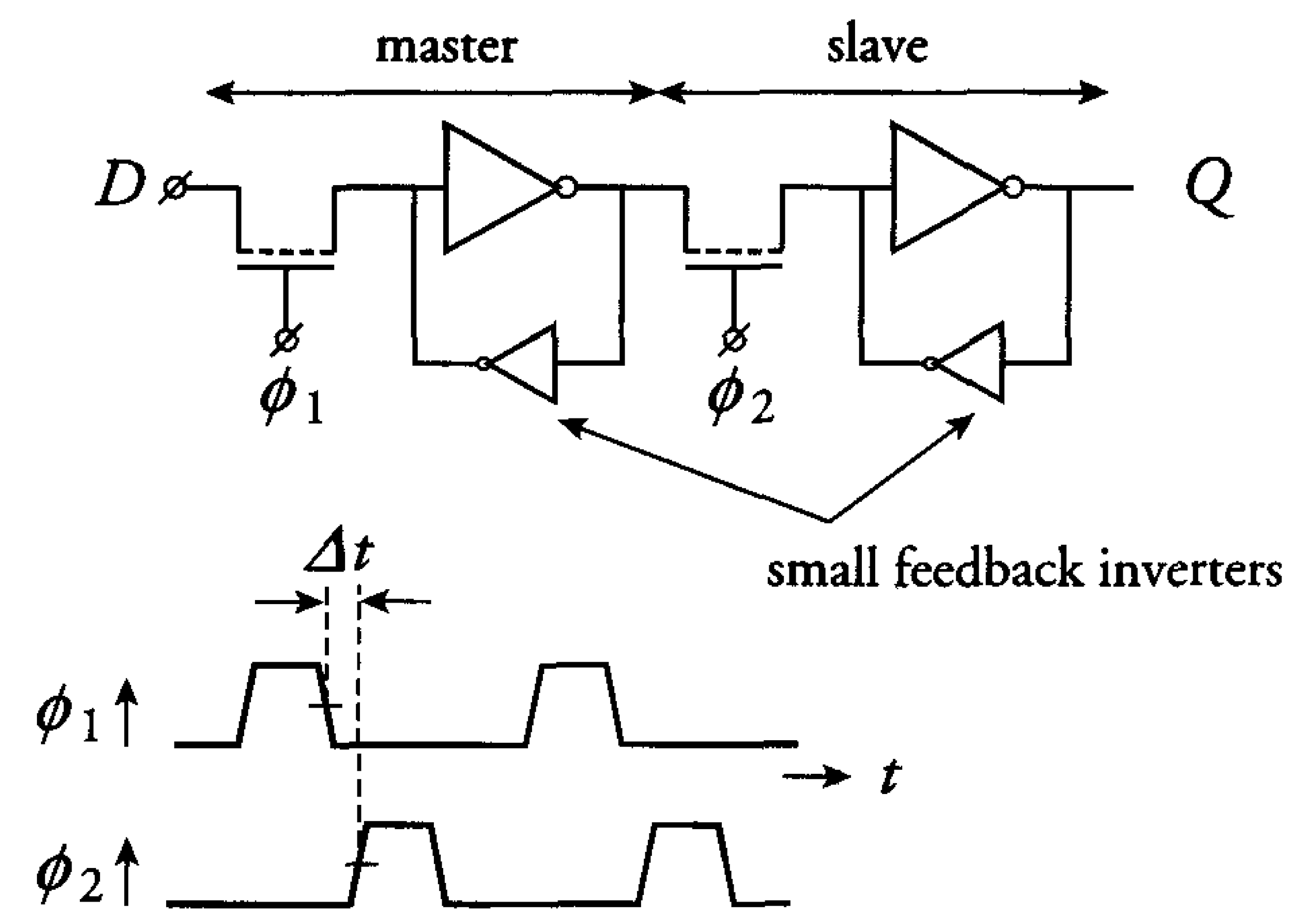


Figure 4.37: *(a) D-type flip-flop with nMOS transmission gates and (b) its 2-phase non-overlapping clock signals*

The flip-flop in figure 4.37 can also be implemented with complementary transmission gates. In this case, however, the nMOS and pMOS transistors in the first transmission gate are controlled by ϕ and $\overline{\phi}$, respectively. The nMOS and pMOS transistors in the second transmission gate are controlled by $\overline{\phi}$ and ϕ, respectively.

Another implementation of the D-type flip-flop is shown in figure 4.38. The additional transmission gates in the feedback loops of each latch interrupt these loops when data is being written into the latch. This reduces the driving requirements of the input circuit and the master, which makes it easier to change the state of the flip-flop.

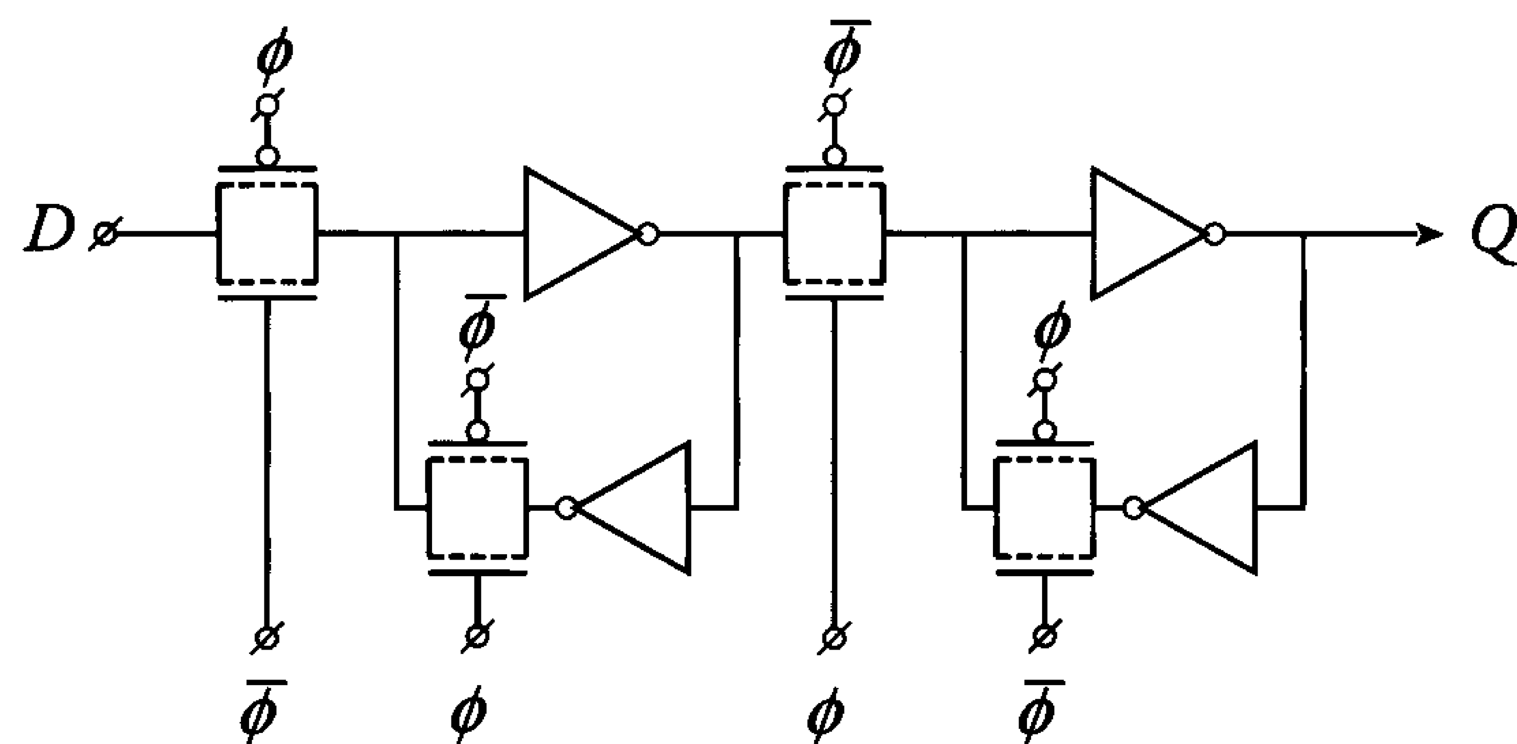


Figure 4.38: *Another implementation of a D-type flip-flop with complementary transmission gates*

Two clocks must be routed in chips with flip-flops which require complementary clocks, such as $\phi1$ and $\phi2$ in figure 4.37 or ϕ and $\overline{\phi}$ in figure 4.38. If the routing area is critical, a single clock flip-flop must be used. Such a flip-flop must then include an inverter to locally generate the inverse of the routed clock. However, there is then an increased risk of 'transparency'. This occurs when the *'clock skew'* causes a flip-flop's transmission gate to simultaneously conduct for a short period of time. This causes the flip-flop to be briefly transparent and data can *'race'* directly from the input to the output. This effect occurs when the flip-flop's complementary clocks arrive via different delay paths. If the clock ϕ_1 in figure 4.37, for instance, is delayed by more than a time period Δt with respect to clock ϕ_2, the flip-flop would be briefly transparent.

Clocks ϕ_1 and ϕ_2 in figure 4.37 are *non-overlapping*, i.e. ϕ_1 is 'low' before ϕ_2 goes 'high' and vice versa. The use of non-overlapping clocks is a good means of preventing transparency in flip-flops.

A discussion of the many more types and variants of static D-type flip-flops is beyond the scope of this book. However, the D-type flip-flop presented in figure 4.39 is particularly interesting. This flip-flop is primarily implemented with NAND logic gates. It requires only a single clock and is very robust. Unfortunately, it consists of 15 nMOS and 15 pMOS transistors and therefore requires considerably more chip area than the 10-transistor flip-flop in figure 4.37. A 'high-density gate array' layout of the flip-flop in figure 4.39 is shown in figure 7.32.

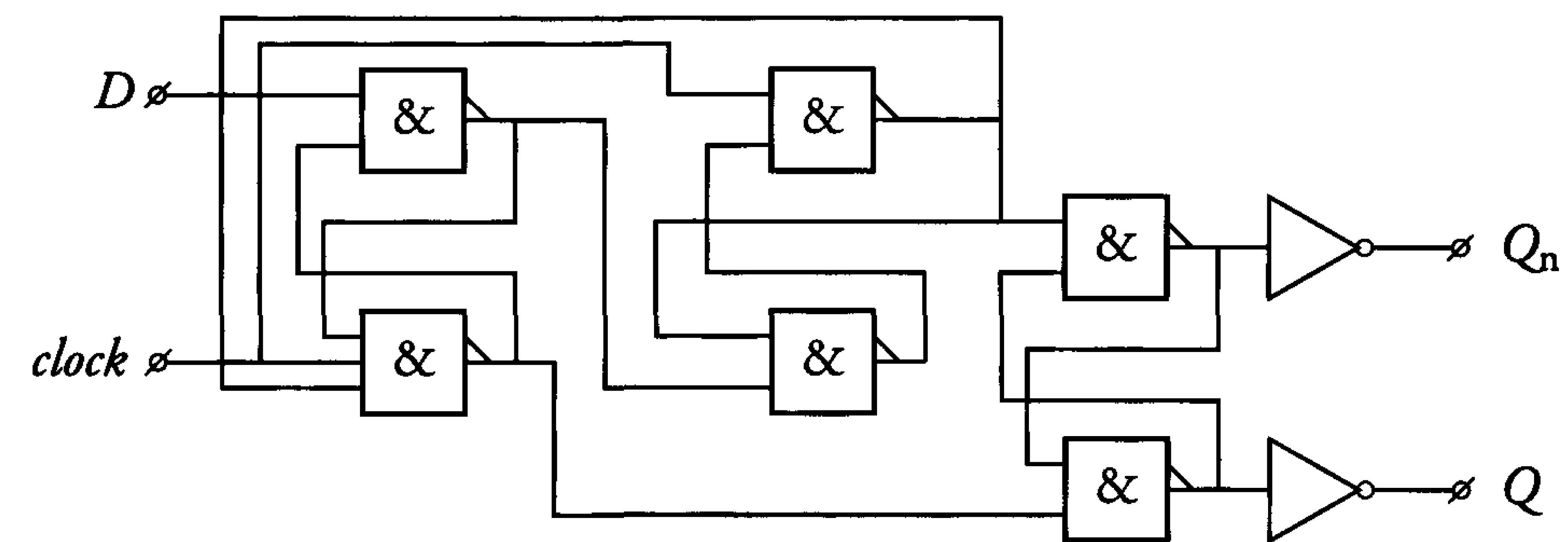


Figure 4.39: *A D-type flip-flop comprising NAND logic gates*

4.4.4 Dynamic CMOS circuits

The main advantage associated with dynamic CMOS circuits is the small chip area that they require. The explanation lies in the fact that logic functions are only implemented in nMOS transistors. Only one pMOS transistor is used per logic gate to charge its output node. Dynamic CMOS circuits are therefore *'nMOS-mostly'* and can occupy significantly less chip area than their static CMOS equivalents. This is particularly true for complex gates.

Figure 4.40 shows a *dynamic CMOS* implementation of a NOR gate. A dynamic CMOS gate of this type requires four different clocks for proper operation, i.e. ϕ_1, $\overline{\phi_1}$, ϕ_2 and $\overline{\phi_2}$. Inputs a and b must be generated by a gate in which ϕ_1 and ϕ_2 are interchanged. The output may also only serve as an input for a gate with ϕ_1 and ϕ_2 interchanged.
The operation of the NOR gate is described as follows:

- Node Z is precharged to V_{dd} when clock ϕ_1 is 'low'.
- When ϕ_1 goes 'high', Z will be discharged if either a or b is 'high'.

- Clock ϕ_2 is then 'low' and the transfer gate passes the value on Z to the input of another logic gate.

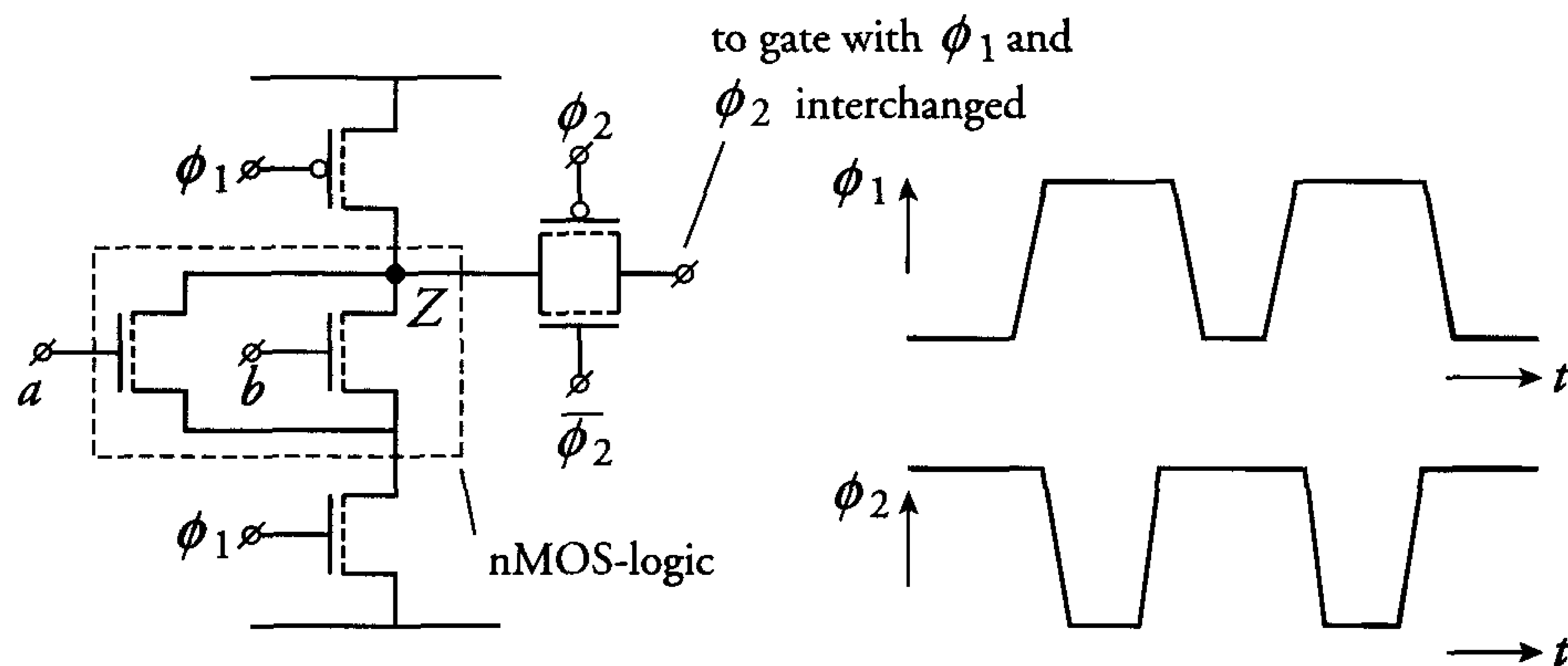


Figure 4.40: *A dynamic CMOS implementation of* $Z = \overline{a+b}$

There is a wide variety of dynamic CMOS logic implementation forms. These include the race-free, pipelined CMOS logic from the Catholic University of Leuven and Bell Labs' DOMINO-CMOS. In contrast to the form of dynamic CMOS shown in figure 4.40, all logic gates in a DOMINO-CMOS circuit are simultaneously precharged during the same part of the clock period. The logic gates sample their inputs when the precharge period ends. In keeping with the domino principle, however, each logic gate can only switch state after its preceding gate has switched. Figure 4.41 shows an example of a DOMINO-CMOS logic gate. The output Y of the dynamic gate is precharged when the clock ϕ is 'low'. The output Z of the static inverter is then 'low'. In fact, the inverter output nodes of all logic gates are 'low' during precharge. These outputs can therefore either stay 'low' or switch to 'high' when ϕ is 'high'. Clearly, each node can only make one transition during this sample period. A node stays in its new state until the next precharge period begins. The data must obviously be given enough time to ripple through the worst case delay path during a sample period. The sample period will therefore be much longer than the precharge period. An important disadvantage of DOMINO-CMOS logic is that all gates are non-inverting. Circuit adaptations are therefore required to implement logic functions with inverse inputs, e.g. an EXOR gate.

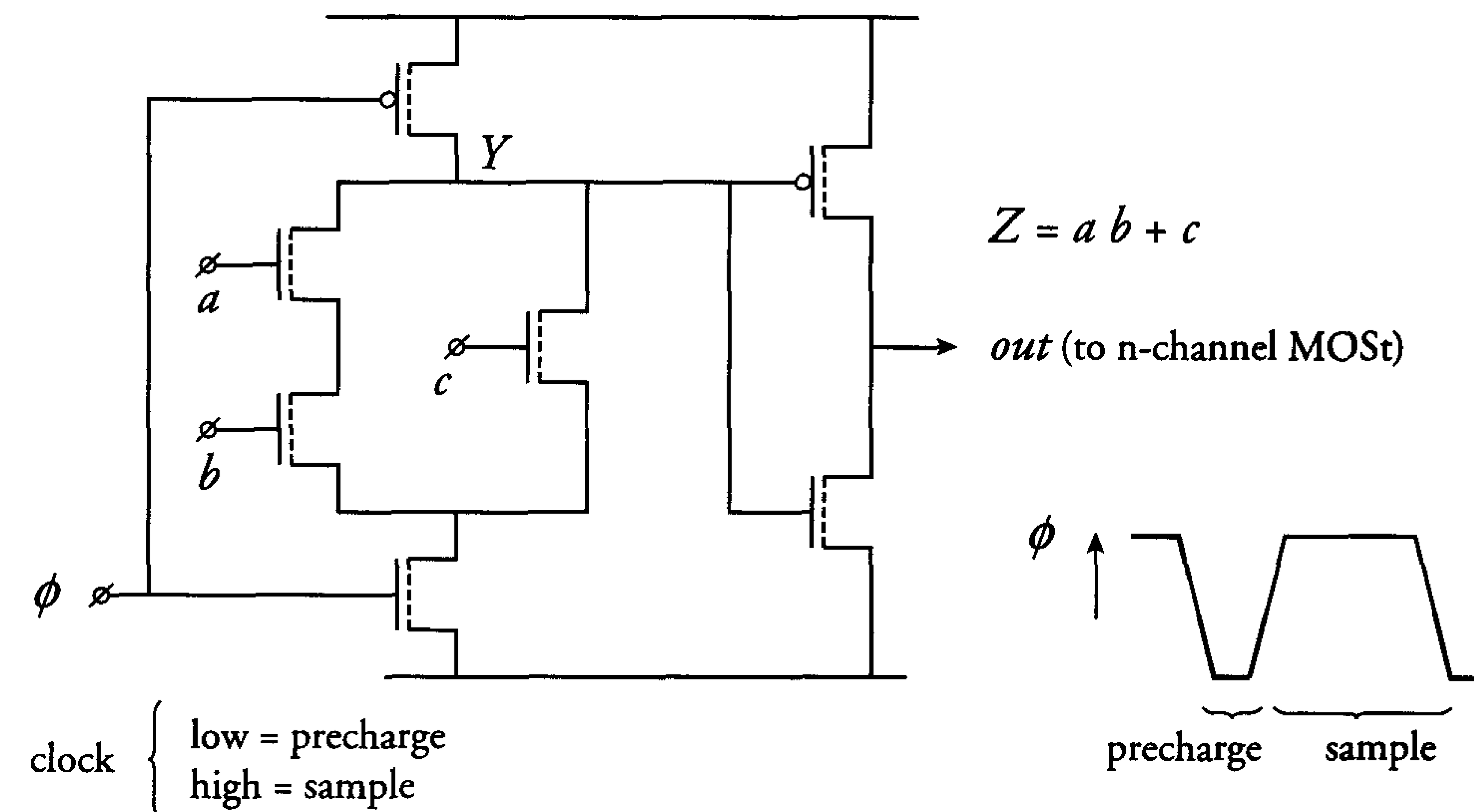


Figure 4.41: *An example of a DOMINO-CMOS logic gate*

Another disadvantage is the need to buffer each logic gate with an inverter; this requires extra silicon area. Today, DOMINO-CMOS logic is often used in high-performance processors. Particularly the most delay-critical circuits, like multipliers and adders are implemented in some style of DOMINO-CMOS [11]. With respect to power dissipation, several remarks on dynamic circuits are made in chapter 8.

Dynamic CMOS latches, shift registers and flip-flops

There are many variations of dynamic CMOS shift registers. However, most of them (like their static CMOS counterparts) basically consist of inverters and transfer gates. A *shift register* is in fact a series connection of flip-flops. Dynamic versions of latches and flip-flops therefore also exist. A dynamic *flip-flop* is also referred to as a dynamic shift register cell because it dynamically shifts data from its input to its output during a single clock cycle.

A minimum clock frequency is required to maintain information in circuits that use dynamic storage elements. This minimum frequency is usually several hundred Hertz, and is determined by the sub-threshold leakage current and the leakage current of the reverse-biased diffusion to substrate pn-junctions in both nMOS and pMOS transistors. There are many different types of dynamic CMOS storage elements.

By deleting the feedback inverters in figure 4.37, we get the dynamic D-type flip-flop shown in figure 4.42. Of course, this flip-flop comprises two dynamic latches.

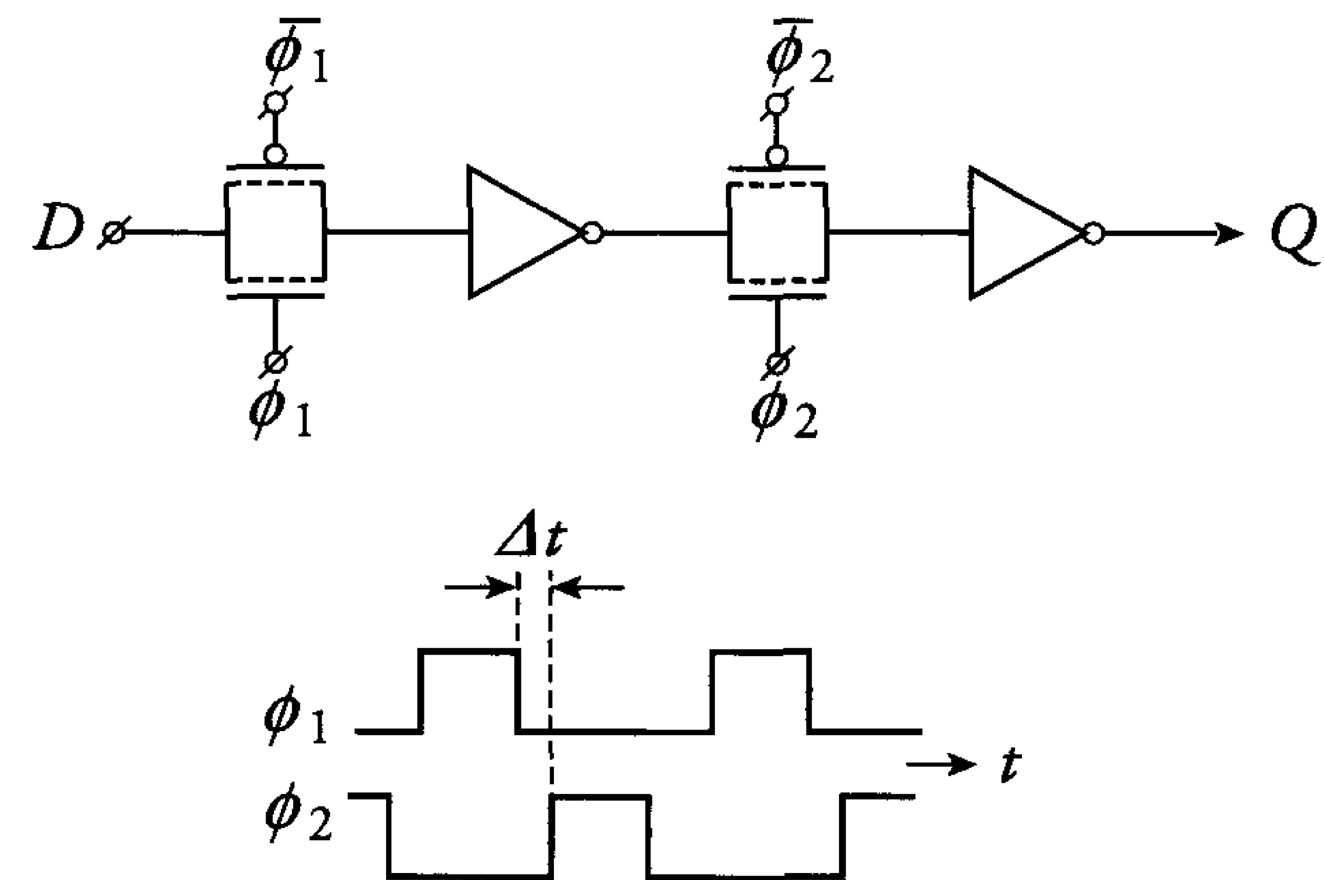


Figure 4.42: *Dynamic D-type flip-flop with non-overlapping clock signals*

The input data D in the above flip-flop is dynamically stored on the input capacitance of the first inverter when ϕ_1 is 'high'. When ϕ_2 is 'high', the output level of the first inverter is dynamically stored on the input capacitance of the second inverter. The *non-overlapping clocks* are intended to prevent the latch from becoming transparent and allowing data to race through the cell during a clock transition. Just as in the static flip-flop, however, this flip-flop will become transparent if the clock skew exceeds Δt. A shift register operates incorrectly when transparency occurs in its flip-flops.

Figure 4.43 presents another type of dynamic CMOS shift register cell. An advantage of this implementation is the reduced layout area resulting from the absence of complementary transfer gates. The clocks in the first section could also be switched and used in the second section. The resulting risk of transparency requires considerable attention.

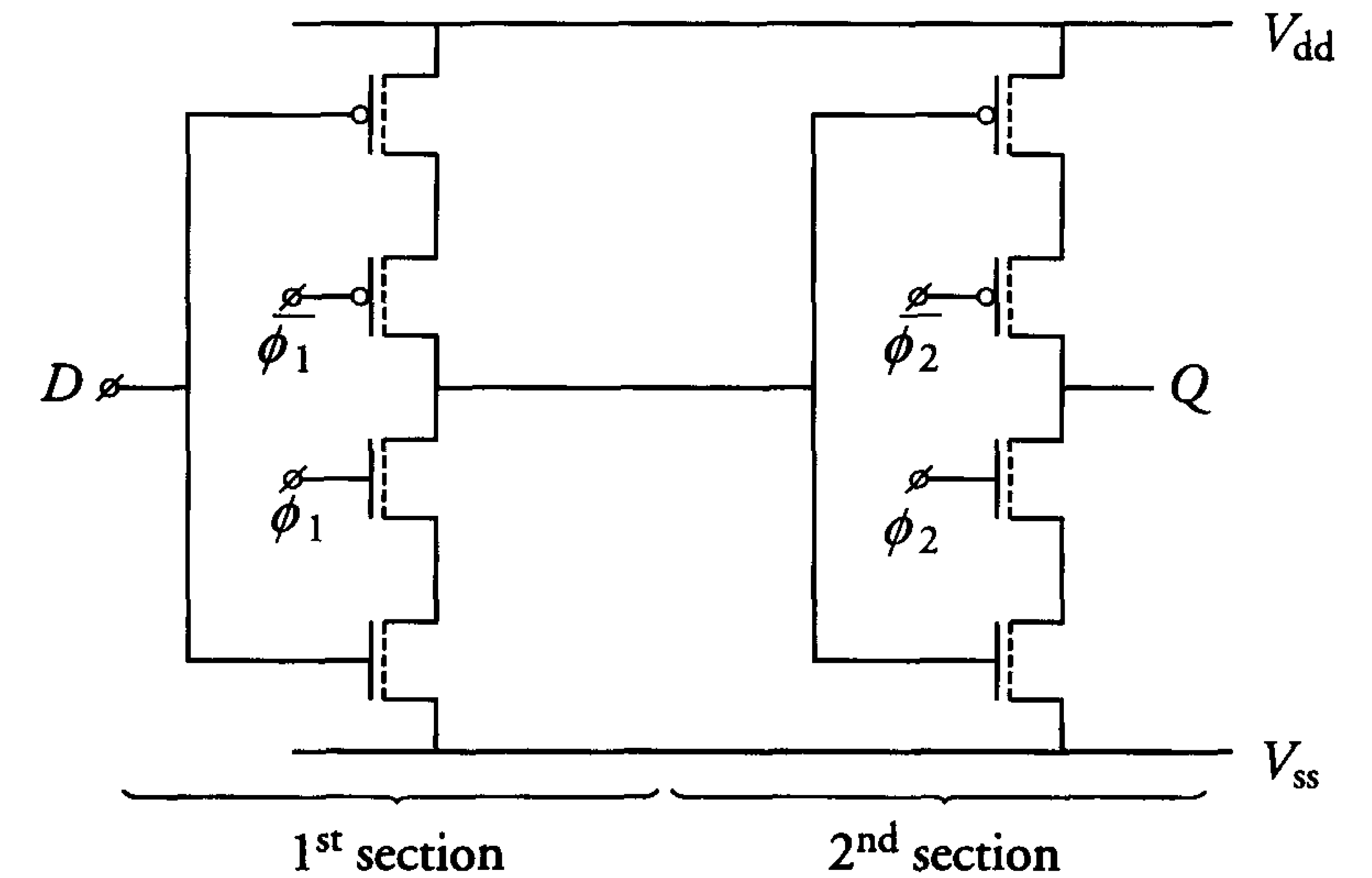


Figure 4.43: *Another dynamic CMOS shift register cell*

Critical phenomena in dynamic circuits

The operation of dynamic MOS circuits relies on the parasitic capacitances that store the logic levels. During a certain period of the clock cycle, several nodes in a dynamic circuit become floating, which makes them very susceptible to such effects as charge sharing and cross-talk.

- **Charge sharing**
  A typical example of charge sharing is shown in figure 4.44.

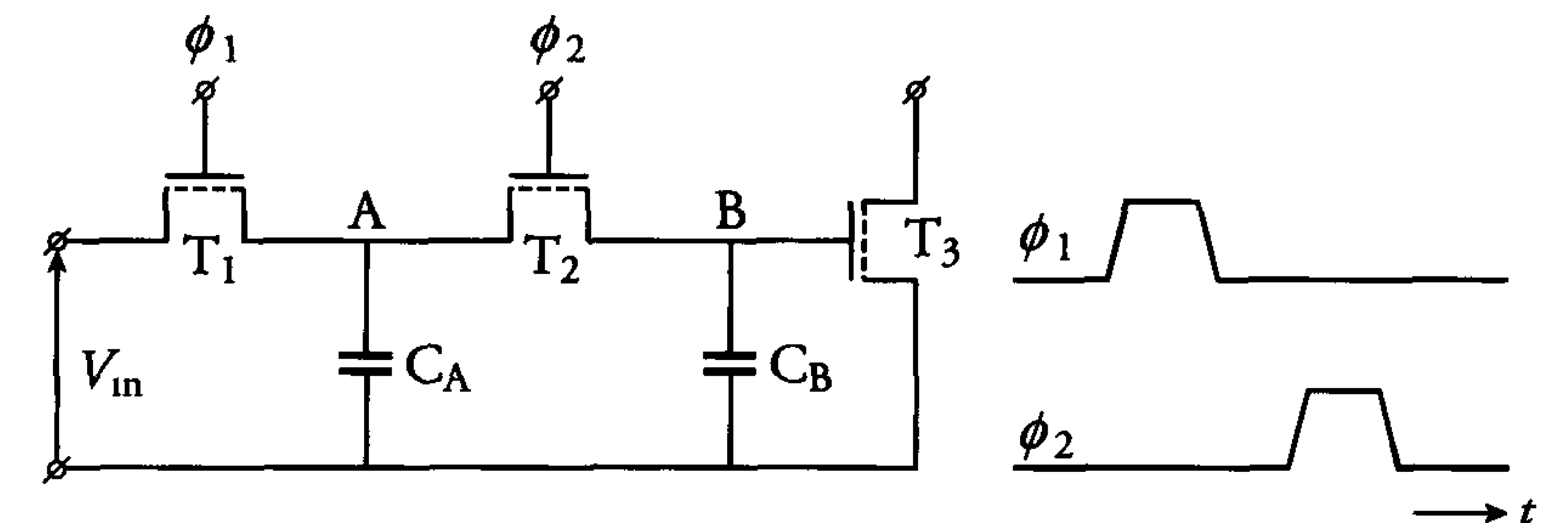


Figure 4.44: *An example of charge sharing*

The high levels of clocks ϕ_1 and ϕ_2 are assumed to cause no threshold loss in transistors T_1 and T_2. When ϕ_1 goes 'high', C_A is charged to the voltage V_{in} and remains at this level when ϕ_1 goes low again. During the period when ϕ_2 is 'high', the charge on C_A is shared between C_A and C_B. The voltages at nodes A and B are then described by:

$$V_A = V_B = \frac{C_A}{C_A + C_B} \cdot V_{in} \tag{4.18}$$

As long as $C_B \ll C_A$, then $V_A \approx V_{in}$. However, if C_B is relatively large, then a 'high' level will be significantly degraded when charge is shared between C_A and C_B. Charge sharing circuits must therefore be used with caution and, if possible, should be avoided.

- **Cross-talk**
  Figure 4.45 shows a schematic of a situation in which cross-talk can occur. A capacitance C exists between node A and a signal track B which crosses it. When ϕ_1 goes from '1' to '0', capacitance C_A is supposed to act as temporary storage for the logic signal that was at A when ϕ_1 was '1'. However, node A has a very high impedance when ϕ_1 is '0', and a voltage change ΔV_B on the signal track B results in the following voltage change at node A:

$$\Delta V_A = \frac{C}{C_A + C} \cdot \Delta V_B$$

The value of the 'cross-over' capacitance C is proportional to the area of the overlap between node A and track B. A large value for C can lead to a disturbance of the logic levels at node A. The area and the number of potentially dangerous crossings must therefore be kept to a minimum during the layout phase of dynamic circuits. Each dynamic node in the finished layout must be checked to ensure that cross-talk noise remains within acceptable margins.

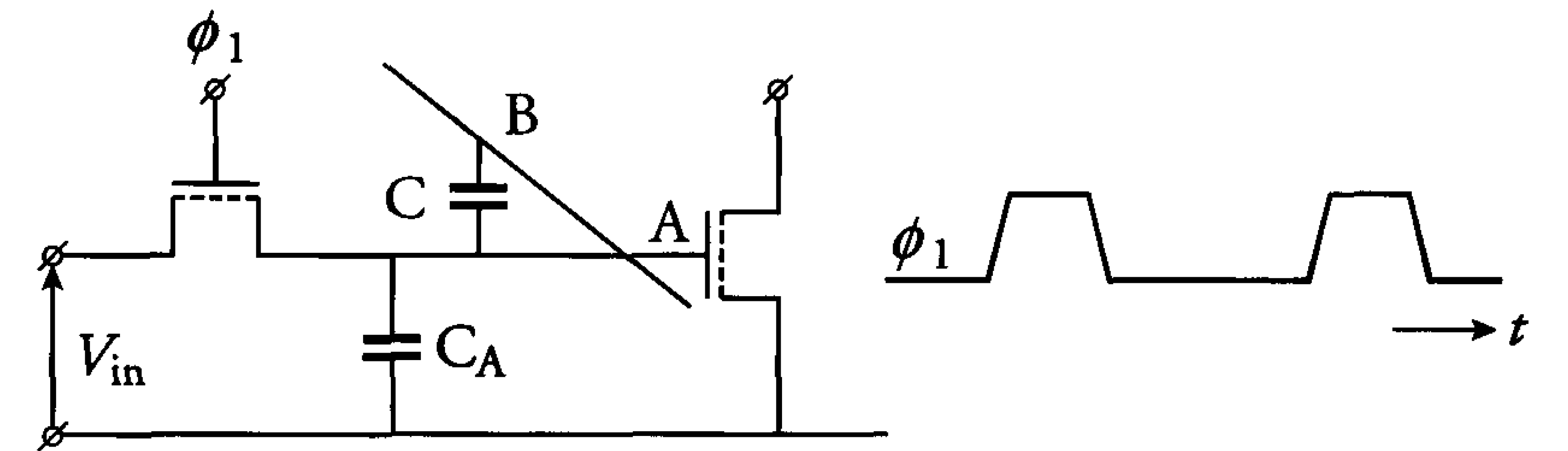


Figure 4.45: *A potential cross-talk situation*

The properties of dynamic MOS circuits can be summarised as follows:

- dynamic MOS circuits have less fan-in capacitance and consume less chip area than static equivalents.
- phenomena such as charge sharing and cross-talk make the electrical design and layout of dynamic nMOS circuits considerably more difficult than for static circuits.

Full CMOS (static CMOS) circuits are currently clearly ahead of dynamic CMOS circuits in the VLSI race. Significant numbers of CMOS ICs, however, still use dynamic CMOS circuits for the implementation of special functions.

4.4.5 Other types of CMOS circuit

The most important characteristics of different CMOS circuits have been presented. These include the small chip area associated with dynamic implementations of logic gates, the low power dissipation associated with static implementations, large logic swings and large noise margins, etc. The advantages and disadvantages associated with an implementation choice can therefore be weighed up. Power dissipation, for instance, can be sacrificed for speed, or speed can be achieved when lower noise margins are accepted.

In the past, several articles have appeared on specialised forms of CMOS, including Cascode Voltage Swing Logic (CVSL) [8]. A CVSL logic gate is obtained by replacing the pMOS transistors in a conventional static CMOS logic circuit by nMOS transistors, which require inverse input signals. The reduction in chip area (at the expense of speed) is particularly noticeable when complex logic gates are implemented in

static or dynamic CVSL. A modified form of CVSL called Differential Split Level (DSL) Logic uses a reduced logic swing. It therefore operates about 2 to 3 times faster but dissipates more power than CVSL.

Some advice which may simplify the task of selecting the right logic implementation is given in the next section.

4.4.6 Choosing a CMOS implementation

An important decision at the start of a new CMOS design is the choice of logic implementation. The choice of a static or dynamic form is determined by a number of factors. The most dominant are power dissipation, speed, chip area and noise immunity. These factors are examined below.

Power dissipation

As previously shown, static CMOS circuits do not dissipate power when the circuit is stable. Instead of the sub-threshold leakage, power is only dissipated in gates that change state. In clocked static CMOS circuits, most power dissipation occurs during and immediately after clock transitions. In clocked dynamic CMOS, however, each gate output is precharged every clock cycle.

Consider the dynamic inverter as an example. If the input remains 'high' during successive clock periods, then the output should be 'low'. However, the output is precharged during every clock period. This repeated charging and discharging of the output leads to high power consumption. A static CMOS inverter in the same situation would not change state and would therefore consume no power. Circuits for low-power or battery-operated applications and many memory circuits are therefore implemented in static CMOS. Chapter 8 presents extensive discussions on low-power issues.

Speed and area

Dynamic CMOS logic circuits are generally faster than their static CMOS counterparts. The nMOS-mostly nature of dynamic CMOS logic means that pMOS transistors are largely reserved for precharge and/or transfer functions while logic functions are only implemented in nMOS transistors. The input capacitance of a dynamic logic gate is therefore lower than a static equivalent. In addition, complex logic gates implemented in static CMOS may contain many pMOS transistors in series in the 'pull-up' path. A dynamic CMOS implementation offers increased speed and

a smaller area because it uses only one pMOS transistor as an active pull-up.

Noise immunity

In a static CMOS logic circuit, there is always a conduction path between a logic gate's output and ground or the supply. Therefore, no logic gate output nodes are floating. Noise-induced voltage deviations on their logic levels are automatically compensated by current flows which restore levels. Dynamic circuits suffer from charge sharing and cross-talk effects, as already mentioned. There is also always a minimum clock frequency required because of the leakage of charge from floating nodes. As a result, static circuits are more robust. For this reason, most semi-custom design libraries are implemented in static CMOS.

4.4.7 Clocking strategies

Advantages and disadvantages of several implementations of single-phase and multi-phase *clocking strategies* have been described in the previous discussions of static and dynamic CMOS circuits. Single-phase circuits are the most efficient in terms of routing area. However, they may require more transistors than multi-phase alternatives. These transistors form the inverter required in each DOMINO CMOS logic gate or the inverter per flip-flop when a second phase is locally generated. The many transistors required for a NAND gate implementation of a flip-flop should also be remembered. In addition, the timing behaviour of *single-phase* circuits is critical and requires many circuit simulations to ensure equivalent functionality for best and worst cases, i.e. when delays are shortest and longest, respectively. *2-phase* circuits that use non-overlapping clocks have less critical timing behaviour.

Clock skew is always present in clocked circuits. Chapter 9 describes clocking strategies and alternatives, and also extensively discusses potential timing problems involved in designs with relatively large clock skew(s).

4.5 CMOS input and output (I/O) circuits

The electrical 'interfaces' between a CMOS IC and its external environment must ensure that data is received and transmitted correctly. These input and output interfaces must be able to withstand dangers that they

may be reasonably expected to encounter. CMOS input and output circuits and the associated protection circuits are discussed below.

4.5.1 CMOS input circuits

MOS ICs often have to communicate with several other types of logic, such as ECL and TTL. A *TTL-compatible* input buffer must interpret an input voltage below 0.8 V as 'low' while voltages above 2 V must be interpreted as 'high'. The switching point of a TTL-compatible CMOS inverter must therefore be about 1.5 V. However, the switching point of a symmetric CMOS inverter (i.e. an inverter with equal transconductances for the nMOS an pMOS transistors) is half the supply voltage. The effects of asymmetry on the switching point of an inverter are shown in the transfer characteristic in figure 4.29. This figure clearly illustrates that a TTL-compatible CMOS inverter must be asymmetric.

Figure 4.46 shows a TTL-CMOS input buffer with the correct transconductance ratios between the transistors. The first inverter converts the TTL input signal to a CMOS level. An input buffer is usually located quite a distance from the logic gates that it drives. The required routing then forms a considerable load capacitance. A clock signal's input buffer is even more heavily loaded. The size of the load capacitance determines the required widths of the nMOS and pMOS transistors in an input buffer's second inverter. To achieve equal rise and fall times, the ratio of these widths must be approximately as shown.

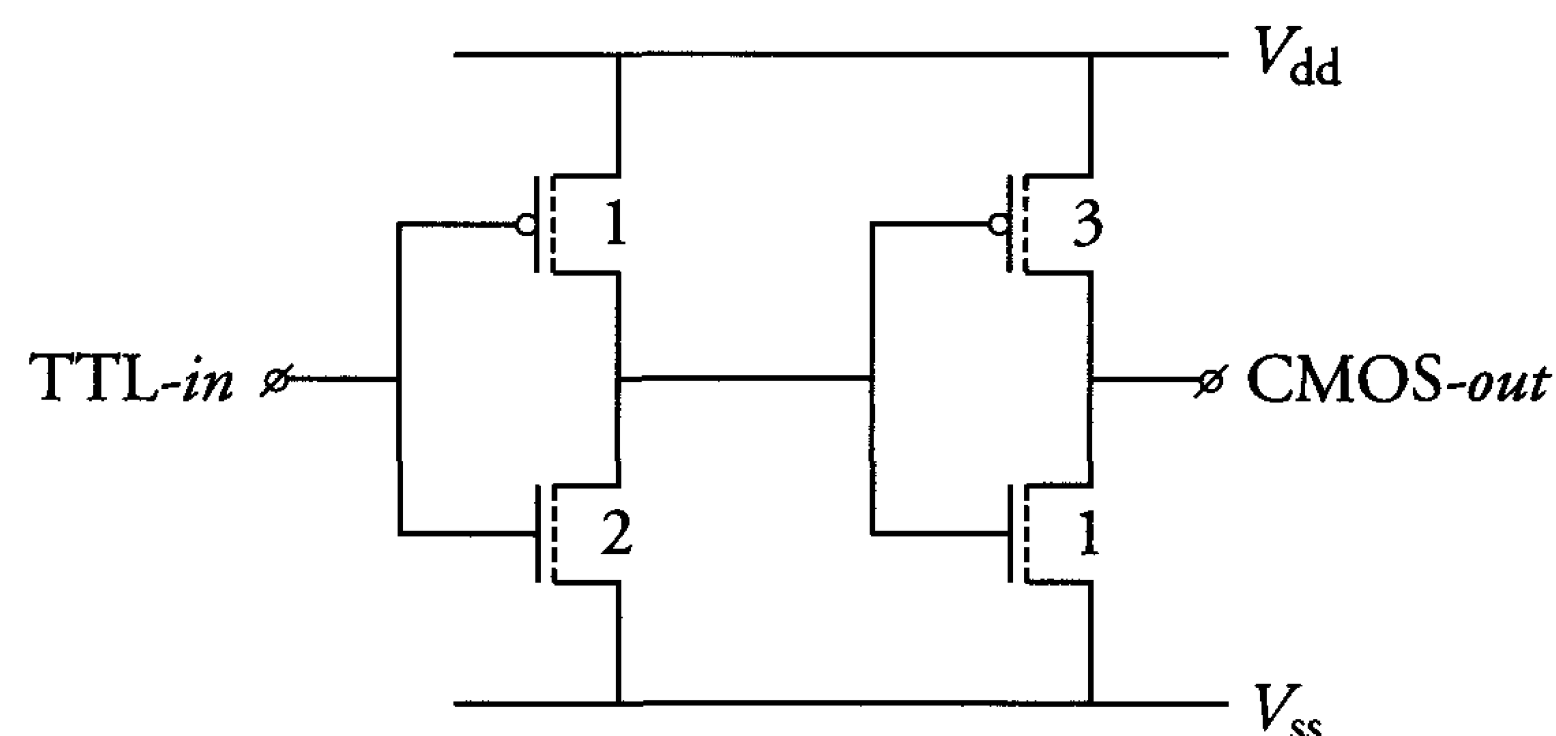


Figure 4.46: *TTL-CMOS input buffer*

The widths and lengths of manufactured transistors may vary independently as a result of processing variations. The effects of these variations

are particularly evident for smaller dimensions. Minimum allowed dimensions should therefore not be used to achieve the required accuracy for the switching point of about 1.5 V for the first inverter in figure 4.46. In a CMOS process, for instance, with a minimum channel length of 0.25 μm and minimum channel width of 0.3 μm, the first inverter could be dimensioned as follows:

$$\left(\frac{W}{L}\right)_{\mathrm{p}} = \frac{0.5}{0.5}\,\mu\mathrm{m} \text{ and } \left(\frac{W}{L}\right)_{\mathrm{n}} = \frac{1}{0.5}\,\mu\mathrm{m}$$

4.5.2 CMOS output buffers (drivers)

There are many different output buffer designs. They usually contain a tapered chain of inverters, as discussed in section 4.3.2. Transistor sizes in the output buffer are determined by the specifications of the output load and the clock frequency. Output load capacitances usually range from 20 to 50 pF, and clock frequencies vary between 50 and 500 MHz.

Several problems arise when many outputs switch simultaneously at a high frequency. The resulting peak currents through metal tracks may exceed the allowed maxima. These currents also cause large *voltage drops* across the intrinsic inductances in the bond wires between a chip's package and its bond pads. The accumulation of peak currents in power and ground lines leads to relatively large noise signals on the chip. These problems (which are also discussed in chapter 9) must be taken into account when designing output buffers.

The very large transistors required in output drivers would result in unacceptably large *short-circuit currents* between supply and ground if the charge and discharge transistors were allowed to conduct simultaneously. Figure 4.47 shows an example of a short-circuit free output buffer. This tri-state buffer is combined with an output flip-flop and can drive a 30 pF load at 100 MHz. Signals 1, 2 and 3 represent the input data, the clock and the tri-state control, respectively. The logic circuits II and III control the gates of the nMOS and pMOS output driver transistors, respectively. These circuits ensure that the driver transistors never conduct simultaneously.

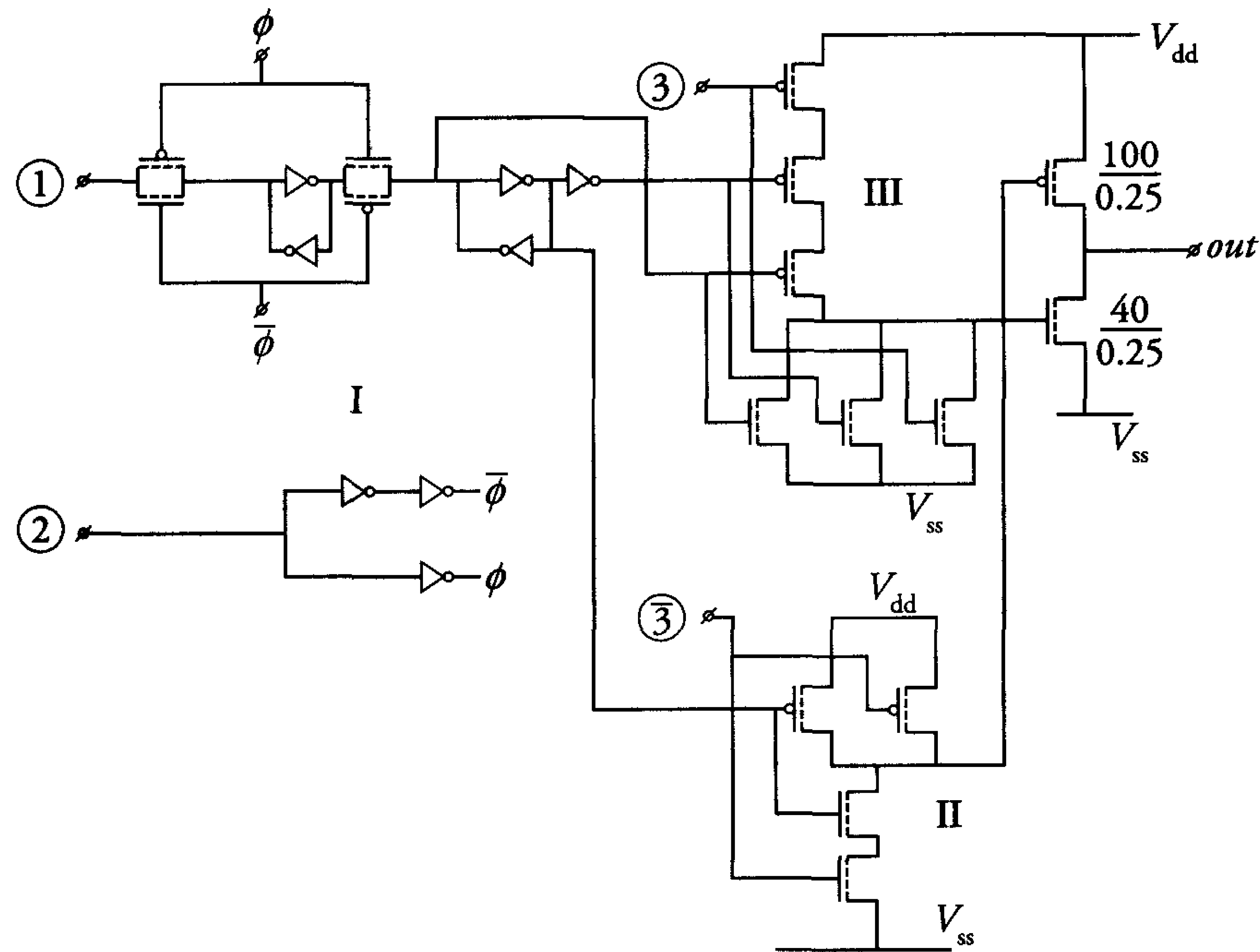


Figure 4.47: *Short-circuit free tri-state CMOS output buffer*

4.6 The layout process

4.6.1 Introduction

In this section, we present a simple set of basic design rules for a *CMOS process* containing a single polysilicon and a single metal layer. These layout *design rules* reflect a 0.25 μm CMOS process. Although such a process usually incorporates about six metal layers, only one metal layer will be used in this layout design process. This is because many of the libraries only use the first metal layer for the local interconnections inside each library cell. After a description of each individual mask, the creation of a stick diagram and the layout process are demonstrated with an example. Finally, a process cross-section shows the real silicon implementation.

4.6.2 Layout design rules

The process *masks* of the chosen technology are listed below in the order of the process sequence. Many of these masks are described in section 3.8.

ACTIVE (layout colour: green)
This mask defines the *active areas* inside which the transistors will be created. Outside the active areas, thick oxide will be formed with STI or LOCOS. The width of an ACTIVE pattern determines the transistor channel width.

NWELL (layout colour: yellow)
This mask defines the areas where the pMOS transistors will be located. The n-well actually serves as a substrate for the pMOS transistors. As the CMOS process offers complementary transistors, the creation of a p-type substrate (p-well) for nMOS transistors is also required. This is usually automatically generated from the NWELL mask: a p-well will be created everywhere where no n-well pattern is defined.

POLY (layout colour: red)
This mask defines the polysilicon pattern. A transistor channel is formed where POLY crosses an ACTIVE region. On top of thin gate oxide, polysilicon acts as a MOS transistor gate. Outside the active areas, polysilicon is used as a local interconnection only over small distances inside the library cells. The minimum width of the polysilicon determines the transistor channel length.

NPLUS (layout colour: orange)
The sources and drains of nMOS transistors need n^+ implants. The NPLUS mask defines the areas in which n^+ is implanted. During the n^+ implantation, the STI (thick oxide regions) and the polysilicon gate act as barriers, e.g. we get self-aligned n^+ regions (sources and drains) everywhere within ACTIVE which is not covered by POLY and surrounded by NPLUS.

PPLUS (layout colour: purple)
Analogous to the NPLUS mask, sources and drains of the pMOS transistor are p-type doped by means of the PPLUS mask.

CONTACT (layout colour: black)

This mask defines contact holes in the dielectric layer below the first metal layer (METAL). Through these contact holes, the metal layer can contact polysilicon (POLY) and source or drain regions (ACTIVE). Some vendors offer technologies that allow *direct contact* from polysilicon to a source or drain region. Such a contact is also called *buried contact*.

METAL (layout colour: blue)

This defines the pattern in the first metal layer, which can be aluminium, tungsten or copper. Currently, tungsten is often used as the first level metal. A track in this layer can be used for both short and long interconnections because its sheet resistance is relatively low. Precautions should still be taken with the use of tungsten, because its sheet resistance is about five times higher than that of commonly-used aluminium-copper alloys.

The following set of design rules will be used in an example of a layout and in several exercises at the end of this chapter. Figure 4.48 serves as an illustration of each of the design rules.

Set of design rules

ACTIVE		
a.	Track width	0.3
b.	Track spacing	0.5
NWELL		
c.	Track width	1.2
d.	Track spacing	1.2
e.	Extension NWELL over ACTIVE	0.6
POLY		
f.	Track width	0.25
g.	Track spacing	0.4
h.	Extension POLY over ACTIVE (gate extension)	0.3
i.	Extension ACTIVE over POLY (source/drain width)	0.4
j.	Spacing between ACTIVE and POLY	0.2
NPLUS		
k.	Track width	0.5
l.	Track spacing	0.5
m.	Extension NPLUS over ACTIVE (n^+ ACTIVE)	0.25
n.	Spacing between n^+ ACTIVE and NWELL	0.6
PPLUS		
o.	Track width	0.5
p.	Track spacing	0.5
q.	Extension PPLUS over ACTIVE (p^+ ACTIVE)	0.25
CONTACT		
r.	Minimum and maximum dimensions	0.3×0.3
s.	Spacing between contacts	0.5
t.	Extension ACTIVE over CONTACT	0.1
u.	Extension POLY over CONTACT	0.1
v.	Extension METAL over CONTACT	0.1
w.	Spacing CONTACT and POLY gate	0.25
x.	CONTACT on gate regions not allowed	!
METAL		
y.	Track width	0.4
z.	Spacing between tracks	0.4

Figure 4.48: *Illustration of each of the design rules of the previous page*

The minimum width and spacing in a certain mask pattern is defined by the different processing steps involved. For instance, the ACTIVE is defined by the STI formation process, while a METAL pattern is the result of deposition and etching techniques. Minimum overlaps or separations between patterns in different masks are defined by alignment tolerances with respect to a common reference location and by the different processing steps involved. The minimum width of the POLY mask pattern determines the channel length of the transistors and is usually referred to in the process notation, e.g. a $0.25\,\mu$m CMOS process means that the minimum POLY width is $0.25\,\mu$m. Usually, when a complex layout has to be developed, a stick diagram is first drawn to explore the different possibilities of layout interconnections. The use of a stick diagram is discussed first.

4.6.3 Stick diagram

A *stick diagram* is used as an intermediate representation between circuit diagram and layout. This topological representation of the circuit is drawn in colours which correspond to those used in the layout. Only the connections of the different mask patterns are depicted, without paying attention to the sizes. The EXNOR circuit of figure 4.49 serves as an example for the development of a stick diagram. This EXNOR circuit represents the Boolean function: $Z = \overline{(a+b)\overline{ab}} = a\,b + \bar{a}\,\bar{b}$

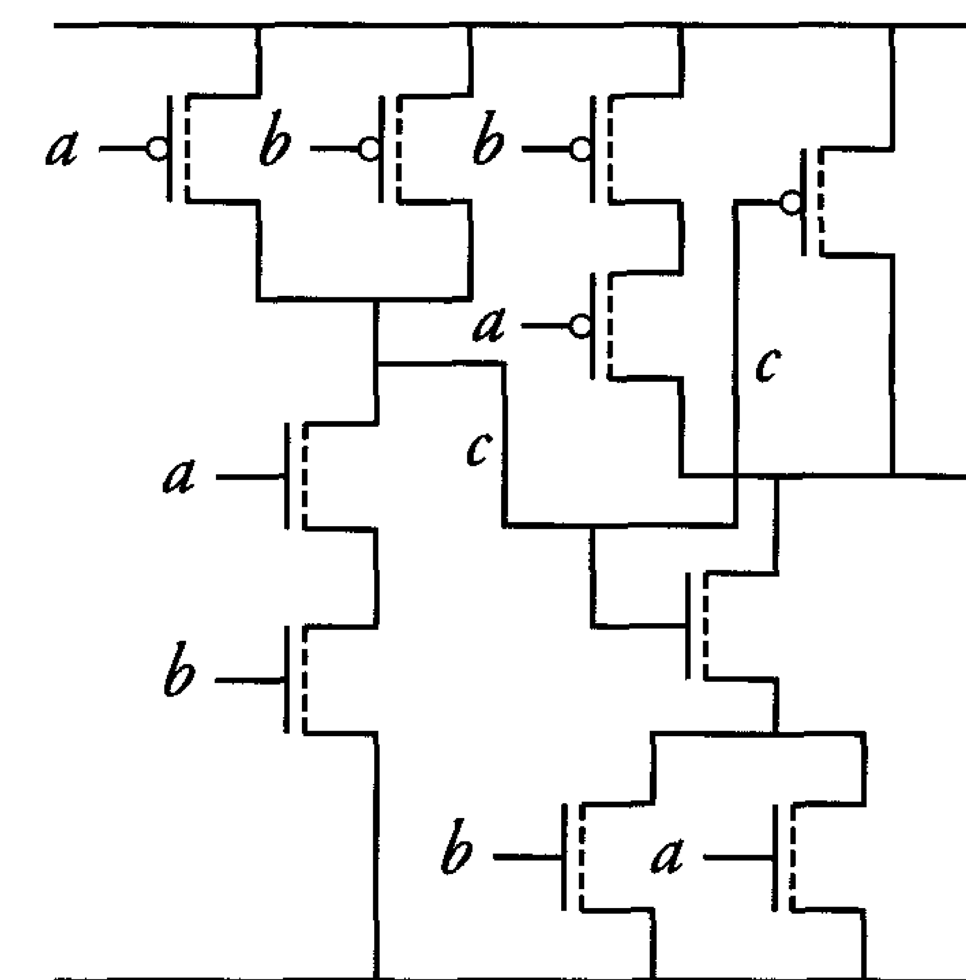


Figure 4.49: *Circuit diagram of a CMOS EXNOR logic gate*

Figure 4.50 illustrates the procedure for the generation of the stick diagram for the EXNOR logic gate.

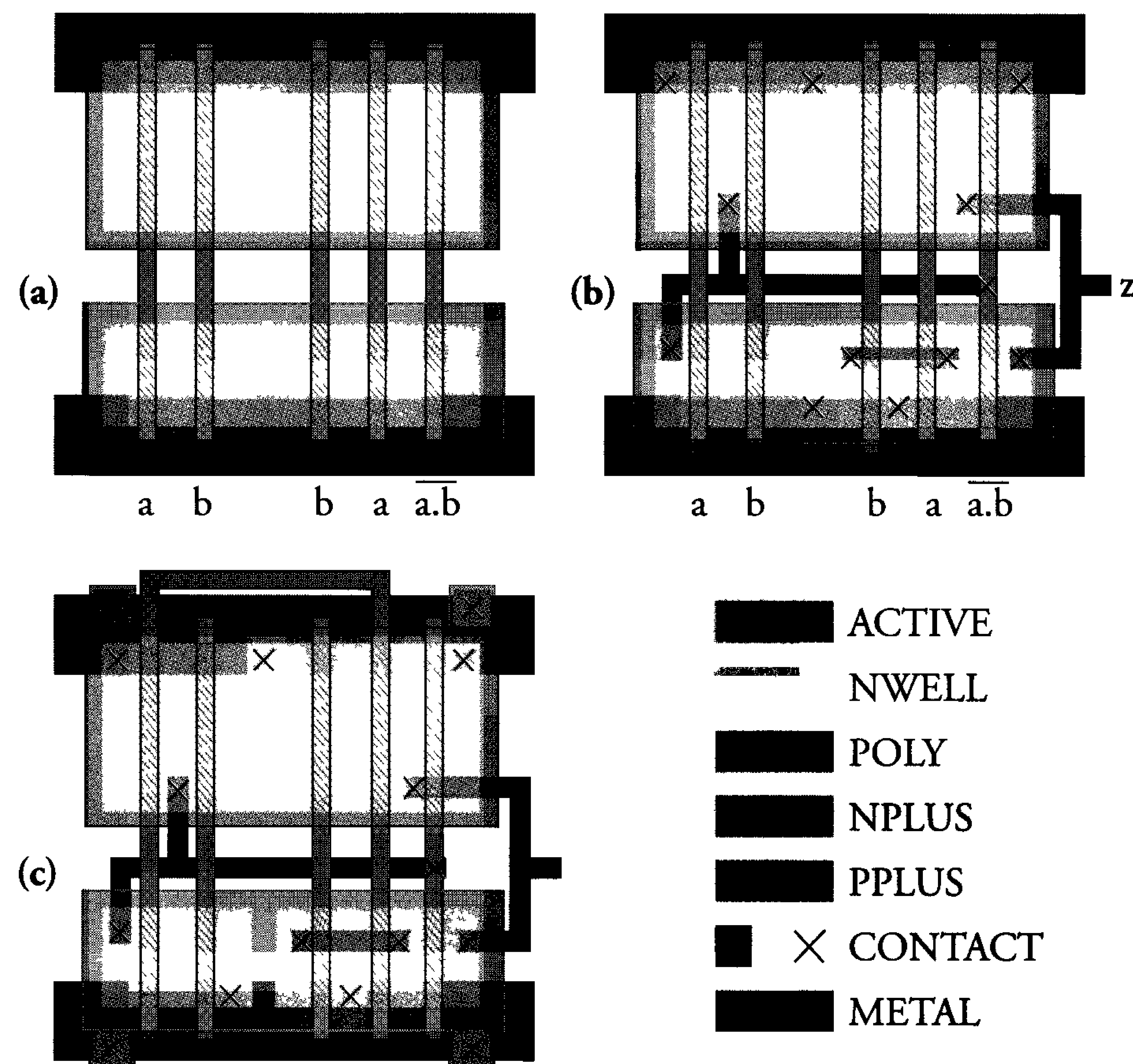


Figure 4.50: *Various steps in the design of a stick diagram*

The creation of this topological view is divided into three phases, represented by (a), (b) and (c) in the figure. These phases are explained as follows:

(a) Two horizontal parallel thin oxide (ACTIVE) regions are drawn. The lower ACTIVE region is usually reserved for nMOS transistors while the upper region is for the pMOS transistors. The envisaged CMOS process uses NPLUS and PPLUS masks to define the n^+ and p^+ diffusion regions of the source/drain areas of the nMOS and pMOS transistors, respectively. An NPLUS boundary is therefore drawn around the lower ACTIVE region in the stick diagram while the upper region is surrounded by a PPLUS boundary. The n-well is indicated by the NWELL area, which overlaps ACTIVE areas surrounded by PPLUS. It is not required to draw the PWELL mask, because it is the inverse of the NWELL mask; everything outside the NWELL area becomes PWELL. Parallel polysilicon (POLY) gates are drawn vertically across both ACTIVE regions. Metal (METAL) supply and ground lines are drawn horizontally over the PPLUS and NPLUS regions, respectively.

(b) Additional METAL and POLY lines indicate transistor connections according to the function to be implemented. The source/drain diffusion areas of neighbouring transistors are merged and black crosses represent contacts. These transistor connections are implemented from left to right. The two nMOS transistors on the left of the stick diagram, for example, correspond to the nMOS transistors of the NAND gate on the left of the circuit diagram in figure 4.49. The drains of two pMOS transistors and one nMOS transistor are connected with METAL to form the NAND gate output. This connection is represented by a metal interconnection of n^+ and p^+ diffusion areas. A direct diffusion connection between an n^+ and p^+ area is not possible as it would form a diode. Connections between n^+ and p^+ areas therefore always occur via metal. The NAND gate output is connected to the gate of the most right nMOS and pMOS transistors.

(c) The third nMOS source/drain area from the left in figure 4.50(b), is connected to ground and to another node. This is clearly not according to the required functionality and such diffusion areas are therefore split into separate diffusion areas in figure 4.50(c). No back-bias voltage is used in the chosen process. The p-type substrate is therefore connected to ground and the n-well is connected to the supply. These substrate and n-well connections are indicated at the edges of the active NPLUS and PPLUS areas, respectively, of the EXNOR's final stick diagram shown in figure 4.50(c).

There should be enough connections from PWELL to ground and from NWELL to V_{dd} to keep latch-up sensitivity to a low level. (latch-up is

discussed in section 9.2.2) These contacts reduce the values of R_1 and R_2, respectively, in figures 9.1 and 9.2. A conservative, but effective, measure is to place an n-well contact adjacent to every supply contact and a substrate contact near every ground contact. One contact may be omitted in situations where two well or two substrate contacts would be very close. This subject is further addressed in the layout discussion below.

4.6.4 Example of the layout procedure

The following example shows the complete layout process from a basic Boolean function, through Boolean optimisation, circuit diagram and stick diagram to a layout. Consider the following Boolean function:

$$Z = \bar{a}\bar{b}\bar{c} + \bar{a}\bar{c}\bar{d} + \bar{a}c\bar{d} + \bar{a}\bar{b}c\bar{d}$$

To optimise this function for implementation in CMOS, an inverse Boolean expression in the format $Z = \bar{f}$ must always be found, because every single CMOS logic gate implements an inverted expression:

$$\begin{aligned} Z &= \bar{a}\bar{b}\bar{c} + \bar{a}\bar{c}\bar{d} + \bar{a}c\bar{d} + \bar{a}\bar{b}c\bar{d} \\ &= \bar{a}(\bar{b}\bar{c} + \bar{c}\bar{d} + c\bar{d} + \bar{b}c\bar{d}) \\ &= \bar{a}(\bar{b}\bar{c} + (\bar{c} + c + \bar{b}c)\bar{d}) \\ &= \bar{a}(\bar{b}\bar{c} + \bar{d}) \\ &= \overline{\overline{\bar{a}(\bar{b}\bar{c} + \bar{d})}} = \overline{a + \overline{(\bar{b}\bar{c} + \bar{d})}} = \overline{a + (b+c)d} \end{aligned}$$

Therefore, the optimised function for implementation as a single CMOS logic gate is: $Z = \overline{a + (b+c)d}$. The circuit diagram for this logic function is shown in figure 4.51.

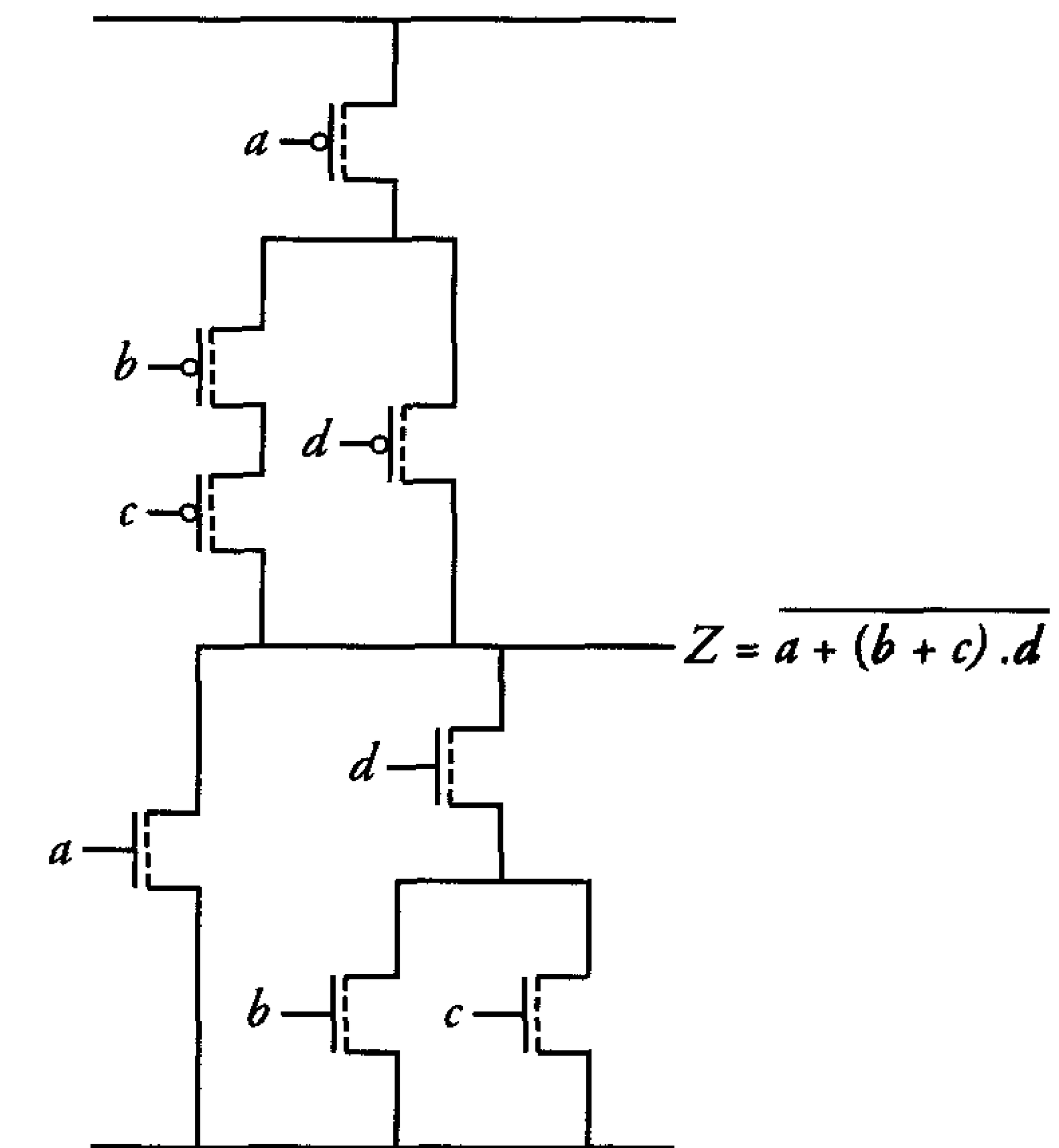


Figure 4.51: *Circuit diagram implementing* $Z = \overline{a + (b+c)d}$

The corresponding CMOS stick diagram and layout can be found in figure 4.52(a) and figure 4.52(b) respectively. Figure 4.52(c) shows a cross-section through the line D-D' in the layout.

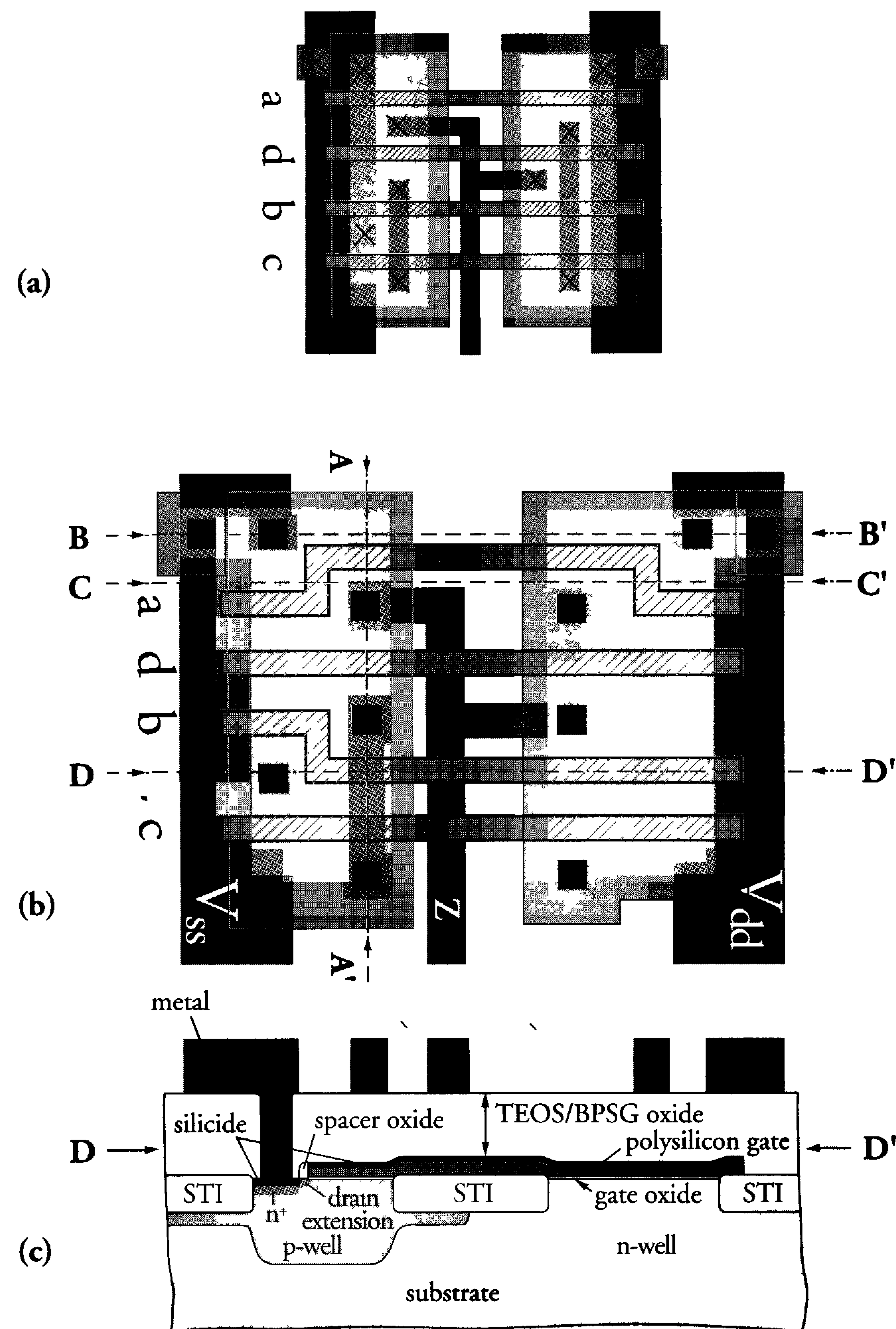


Figure 4.52: *(a) Stick diagram, (b) layout and (c) cross-section of the sample logic gate along the line D-D'*

The layout contains one substrate (p-well) and one n-well contact. The use of extra n-well and p-well contacts reduces latch-up sensitivity but may lead to an increased layout area. A practical compromise is to place at least one substrate and n-well contact per five nMOS and pMOS transistors, respectively. This rule of thumb applies to logic circuits. The large transistors in driver and I/O circuits require considerably more substrate and n-well contacts.

The n-wells in a CMOS circuit layout are usually connected to the supply voltage. Generally, different neighbouring n-wells (which are connected to the same voltage) should be extended to form one large well. Although shown in the figure, there are no additional p^- substrate (PWELL) and NWELL contact holes (CONTACT) needed, because adjacent n^+ source regions and PWELL areas are connected through the silicided top layer.

The output node of a static CMOS logic gate is formed by an interconnection of n^+ and p^+ diffusion areas. The p^+ diffusion area is usually the larger. The parasitic capacitance of such an output node is therefore larger than its nMOS counterpart. In addition, the width of a pMOS transistor is usually larger than an nMOS transistor width.

As a result of silicided p^+ diffusion regions, the series resistance of sources and drains are low and only one contact is sufficient per connected node. These resistances are only several ohms per square in CMOS technologies with silicided source and drain regions. Minimum source and drain areas can then be used to keep parasitic capacitances low.

The process cross-section in figure 4.52(c) is made along the line D-D'. The cross-section includes n^+ diffusion, transistor gate areas, STI oxide areas, n-well and p-well areas and a source contact. A detailed study of the relationship between the cross-section and the layout should enable the reader to draw a cross-section at a line anywhere in the layout.

Circuit density and performance are often improved by using several polysilicon layers (memories) and five to seven metal layers (VLSI). The area reduction must compensate for the costs associated with the additional masks and processing steps. However, with the ever-increasing current density, more and more metal layers are required to distribute the power properly across the chip.

The use of slanting lines is largely restricted to memory layouts only. In logic chips, the area reduction associated with the use of slanting rather than orthogonal lines is generally no more than 5 %. This does

not compensate for the potential physical problems (high electric fields in the corners) or for the problems that slanting lines cause for CAD tools. This latter category of problems arises especially when growing or scaling operations must be performed by software.

4.6.5 Guidelines for layout design

Designing a correct layout involves more than just a translation of the circuit diagram into a layout that meets the relevant design rules. Attention must be paid to several key issues:

- Minimise layout area.
  A minimum layout area will especially reduce the overall silicon costs with the development of a new library that is to be used for the design of numerous chips. Moreover, when ICs become smaller, they generally show a higher performance, consume less power and are cheaper.

- Pay attention to parasitic elements.
  Each design, whether a library cell or a large logic block, must be optimised with respect to parasitic capacitances (source and drain junctions, metal interconnects) and resistances (mainly of long interconnections). This is necessary to achieve better performance and again reduces the power consumption.

- Pay attention to parasitic effects.
  Effects such as cross-talk, charge sharing and voltage drop across supply lines particularly greatly reduce the performance as well as the signal integrity. Such effects are extensively discussed in chapter 9.

Table 4.2 shows some typical values of the capacitances and resistances of different components and materials used in a 0.25 μm CMOS technology with a gate oxide thickness $t_{ox} = 50$ Å (5 nm).

Table 4.2: Parasitic capacitances and resistance values in a 0.25 μm process ($t_{ox} = 5$ nm)

Material	Capacitances	Resistances
Polysilicon (POLY)	gate cap: 7 fF/μm^2 * edge cap: 0.17 fF/μm * track area cap: 0.1 fF/μm^2 track edge cap: 0.05 fF/μm	poly 200 Ω/□ polycide 5 Ω/□
Aluminium (Al) (METAL)	track cap: 0.1 fF/μm^2 edge cap: 0.05 fF/μm	60 mΩ/□
Tungsten (W) (METAL)	track cap: 0.1 fF/μm^2 edge cap: 0.05 fF/μm	250 mΩ/□
Copper (Cu) (METAL)	track cap: 0.1 fF/μm^2 edge cap: 0.035 fF/μm	60 mΩ/□ (at 30% thickness reduction with respect to Al and W)
Source/Drain implants (ACTIVE)	track cap: 0.08 fF/μm^2 thick oxide edge cap: 0.05 fF/μm cap to POLY edge: 0.03 fF/μm	$n^+ \approx$ 40 to 80 Ω/□ $p^+ \approx$ 50 to 150 Ω/□ silicided $n^+ \approx$ 10 Ω/□ silicided $p^+ \approx$ 8 Ω/□

Note: * on thin oxide

It is clear that polysilicon and n^+/p^+ junctions can only be used for very short connections inside library cells as a result of the relatively high sheet resistance values.

Especially deep sub-micron CMOS processes include five to seven layers of metal. In many cases, the upper metal layer has a greater thickness, a larger minimum feature size and a larger spacing. Therefore, this upper level must be used for a structured and proper overall chip power supply network.

The above discussions on CMOS layout implementation conclude this chapter. More information on the design of CMOS circuits and layouts can be found in the reference list.

4.7 Conclusions

CMOS has become the major technology for the manufacture of VLSI circuits, and now accounts for about 90 % of the total IC market. The main advantage of CMOS is its low power dissipation. This is an important requirement in current VLSI circuits, which contain millions of transistors.

Static CMOS circuits are characterised by high input and parasitic capacitances and large logic gate structures. The silicon area occupied by a static CMOS logic circuit is about twice that of an nMOS counterpart. Dynamic CMOS circuits are nMOS-mostly and are therefore generally smaller than their CMOS counterparts. The use of a static rather than a dynamic implementation must therefore be justified by a sufficient reduction in power dissipation. Generally, static CMOS shows the lowest τD product and is thus the most power efficient implementation for VLSI. Moreover, its robustness is very important in current deep sub-micron ICs as these show increasing noise, caused by cross-talk and supply voltage drops. Low-power issues and maintaining signal integrity at a sufficiently high level are the subjects of chapter 8 and 9, respectively.

Basic technologies for the manufacture of MOS devices are explained in chapter 3. Various nMOS circuit principles are introduced. This chapter emphasises the most important differences between CMOS and nMOS circuits. These differences are evident in the areas of technology, electrical design and layout design. A structured CMOS layout design style is presented in this chapter while using a limited set of representative design rules. The combination of the CMOS and nMOS circuit design and layout principles discussed in this chapter should afford the reader sufficient insight into the basic operation of different CMOS circuits.

4.8 References

CMOS physics and technology (see also chapter 3)

[1] Richard C. Jaeger,
'Introduction to Microelectronic Fabrication',
Modular Series on Solid-State Devices, Volume 5, 1988

[2] Y. Sakai, et al.,
'Advanced Hi-Cmos Device Technology',
IEEE IEDM, pp. 534-537, Washington DC, 1981

[3] S.M. Sze,
'VLSI Technology',
McGraw-Hill, New York, 1983

[3a] S. Wolf and R.N. Tauber,
'Silicon processing for the VLSI Era',
Volume 1, Process Technology, Lattice Press, 1986

[3b] S.M. Sze,
'Modern Semiconductor Device Physics',
John Wiley & Sons, 1997

CMOS design principles (general)

[4] C. Mead, L. Conway,
'Introduction to VLSI Systems',
Addison-Wesley, 1980

[5] N. Weste, K. Eshraghian,
'Principles of CMOS VLSI Design, a Systems Perspective',
Addison-Wesley, 1993

[6] L.A. Glasser, D.W. Dobberpuhl,
'The Design and Analysis of VLSI circuits',
Addison-Wesley, 1985

[7] M. Annaratone,
'Digital CMOS circuit Design',
Kluwer Academic Publishers, 1986

[8] L.G. Heller, et al.,
'Cascode Voltage Switch Logic',
IEEE Digest of technical papers of the ISSCC, 1984

[9] Jan M. Rabaey,
'Digital Integrated Circuits: A Design Perspective',
Prentice Hall, 1995

[10] Kerry Bernstein, et al.
'HIGH SPEED CMOS DESIGN STYLES',
Kluwer Academic Publishers, 1999

[11] International Solid-State Circuits Conference
Digest of Technical papers, February 2000,
pp. 90-11, pp. 176-177, pp. 412-413, pp. 422-423

Power dissipation in CMOS

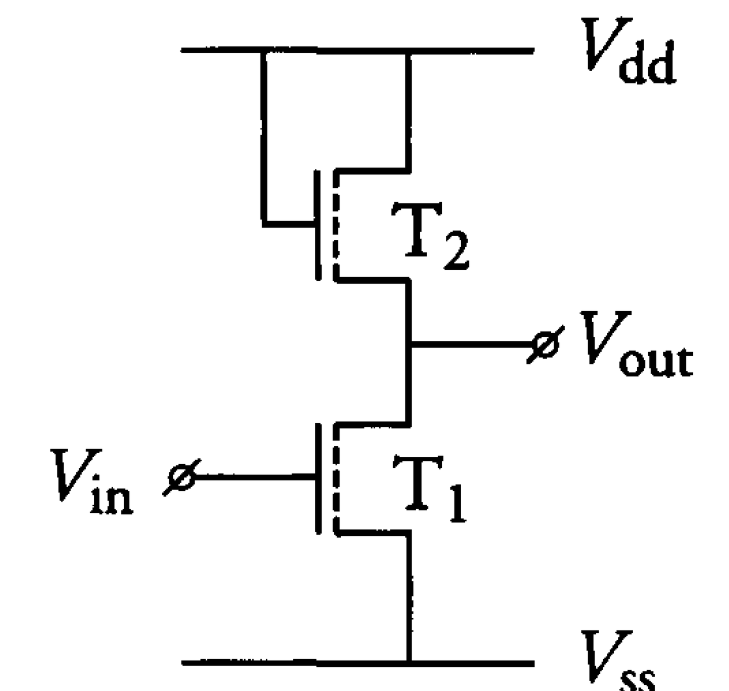
[12] H.J.M. Veendrick,
'Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits',
IEEE Journal of Solid State Circuits, Vol. SC-19,
No. 4, August 1984, pp. 468-473

For further reading

[13] 'IEEE Journal of Solid-State Circuits'

[14] 'ISSCC and ESSCIRC conferences, digests of technical papers'

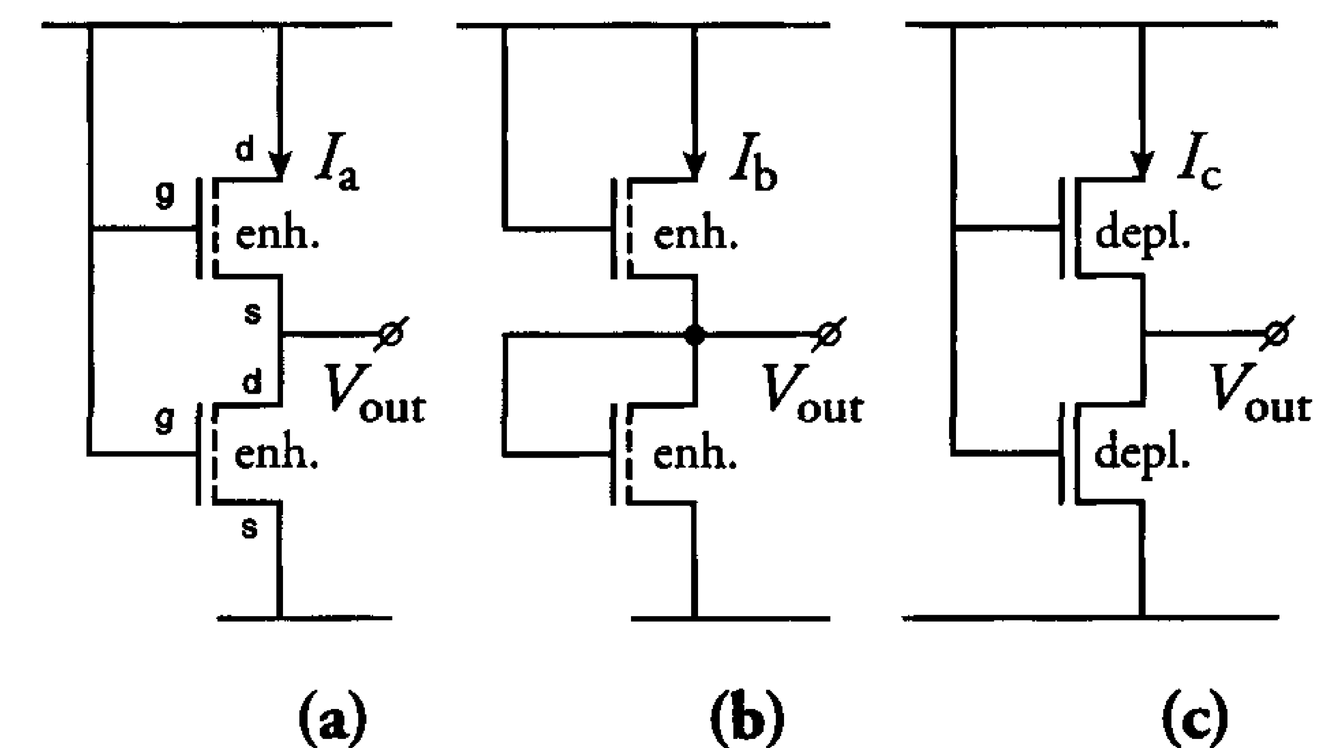
4.9 Exercises



1. The following values apply for the adjacent inverter:
$V_{dd} = 2.5\,V$, $V_{ss} = 0\,V$,
$V_{bb} = -2\,V$, $\beta_\square = 240\,\mu A/V^2$,
$V_x = 0.3\,V$, $\left(\frac{W}{L}\right)_{T_2} = \frac{0.3}{0.25}$,
$L_{T_1} = 0.25\,\mu m$.

   a) Determine $V_{out}=V_H$ when $V_{in}=V_L<0.25\,V$ for the cases when $K=0\,V^{1/2}$ and $K=0.3\,V^{1/2}$

   b) Calculate the width W_{T_1} of transistor T_1 so that $V_L<0.25\,V$ when the input is driven by a circuit with the same type of load transistor and $K=0.3\,V^{1/2}$.



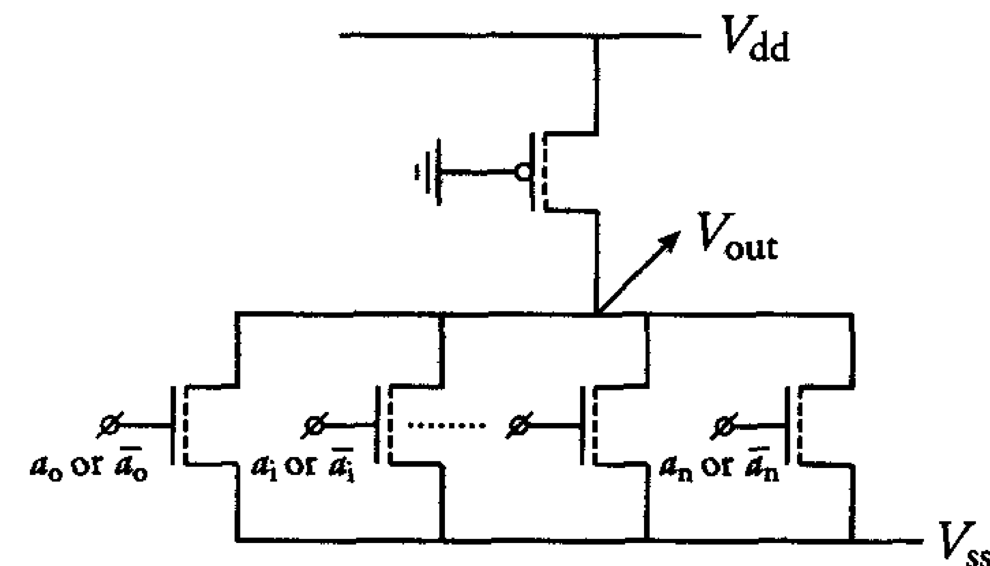
(a) (b) (c)

2. The following values apply for the above circuits:
$V_{dd} = 2.5\,V$
$K = 0\,V^{1/2}$
$|V_x| = 0.5\,V$
All transistors are of the same size.

   a) Which of the currents I_a, I_b and I_c is larger and why?

   b) What is the value of V_{out} in b and c? Explain.

3. The adjacent circuit can be used as a word line selection circuit in a random-access memory. The gate of the pMOS load transistor is connected to ground and the following values apply:

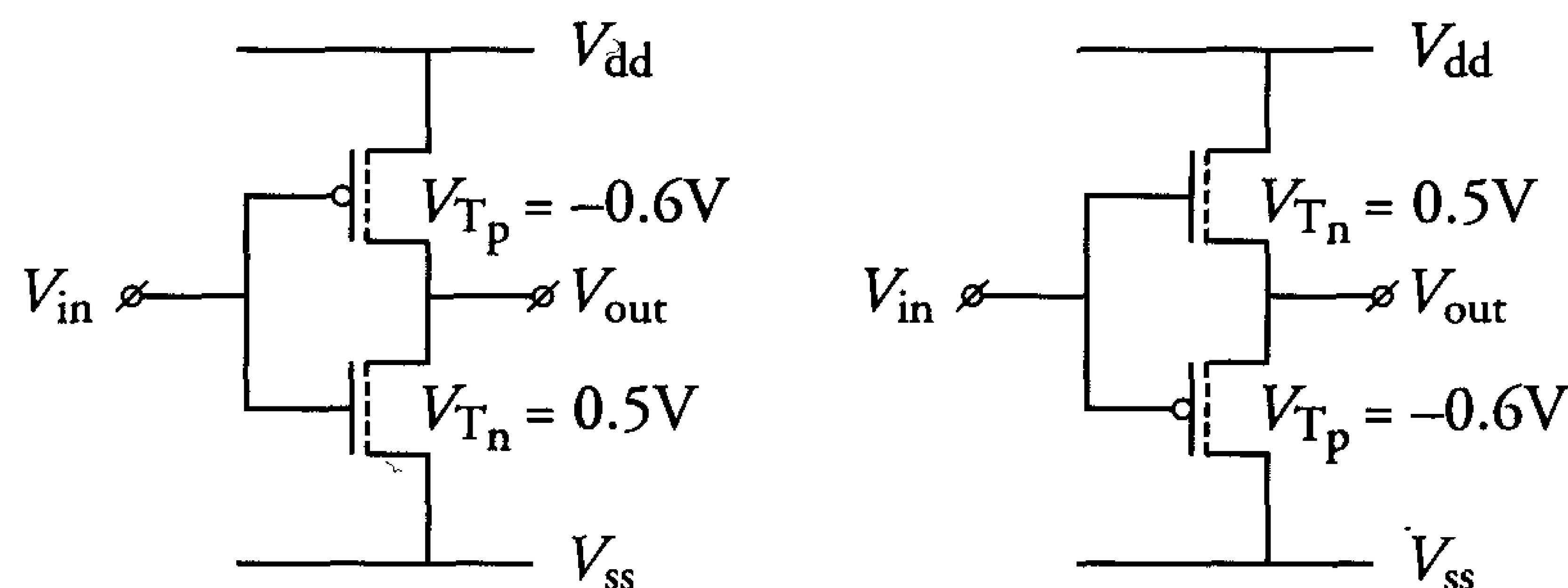


$V_{X_n} = 0.5\,\mathrm{V} = -V_{X_p}$
$\beta_{\square n} = 240\ \mu A/V^2 \qquad \beta_{\square p} = 60\ \mu A/V^2$
$K_n = 0.1\ \mathrm{V}^{1/2} \qquad K_p = -0.3\,\mathrm{V}^{1/2}$
$2\phi_f = 0.64\,\mathrm{V} \qquad V_{dd} = 2.5\,\mathrm{V}$

a) Explain the disadvantages that would arise if the logic was implemented in complementary static CMOS.

b) Calculate the aspect ratio $(\frac{W}{L})_n/(\frac{W}{L})_p$ that yields an output low level $V_L \leq 0.25\,\mathrm{V}$ when only one of the inputs a_i is 'high'.

c) Calculate the aspect ratio $(\frac{W}{L})_p$ when an output load capacitance $C = 500\,\mathrm{fF}$ must be charged to 90 % of the supply voltage in 10 ns.



4. a) If V_{dd}=0.9 V in the above figure, explain what would happen at the output of circuit (a) when V_{in} switches from 0 V to V_{dd} and back. Draw this in the inverter characteristic: $V_{out}=f(V_{in})$.

b) Repeat (a) for V_{dd}=2.5 V.

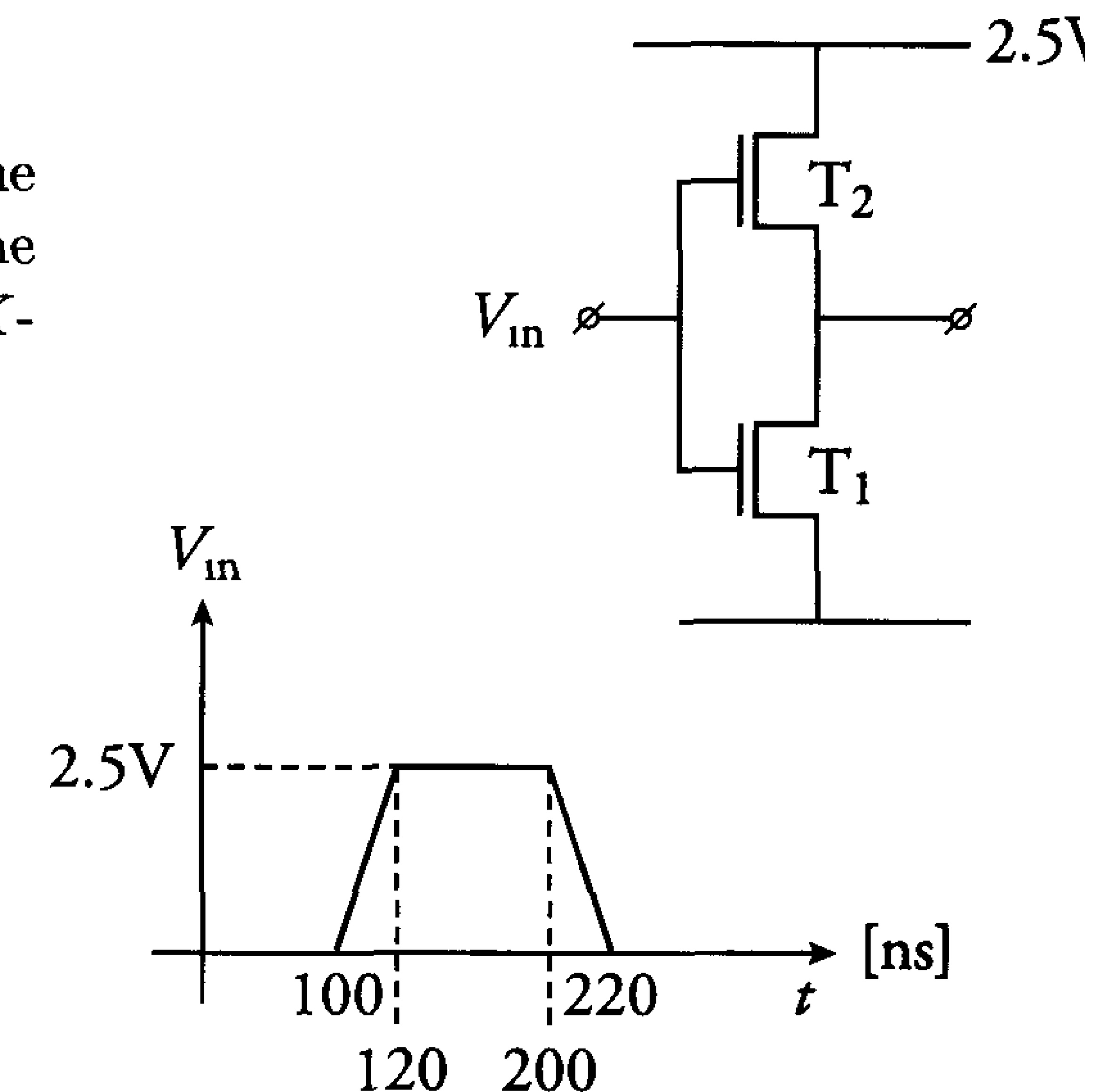
c) If V_{dd}=2.5 V in circuit (b) and V_{in} switches from 0 V to V_{dd} and back, draw $V_{in}=f(t)$ and $V_{out}=f(t)$ in the same diagram.

5. a) Explain in no more than *ten* lines the cause of short-circuit dissipation and design measures to reduce it.

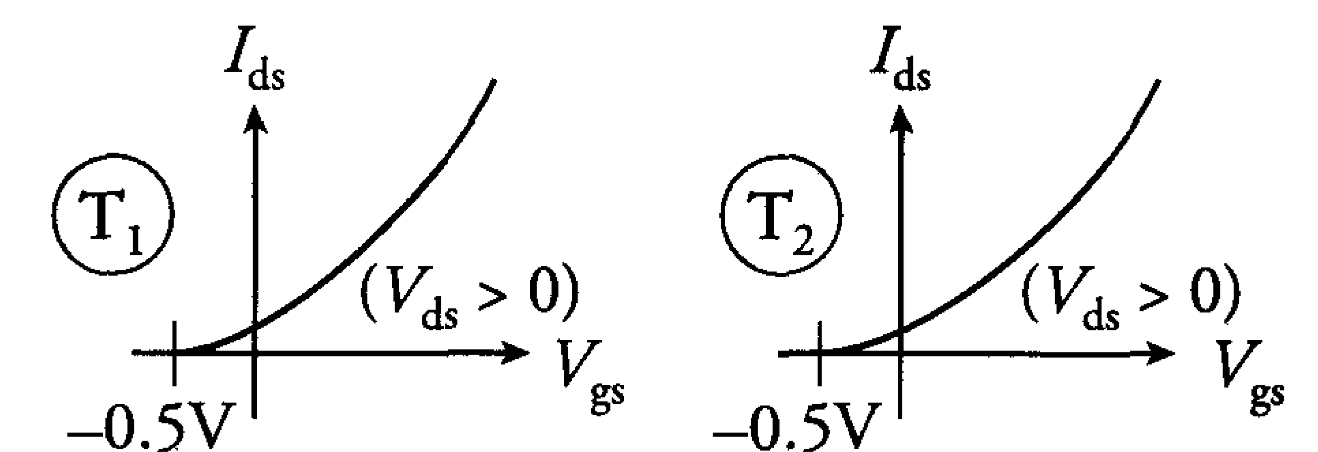
b) Present and describe one or two solutions that reduce the short-circuit dissipation in a driver circuit to zero.

6. Draw a process cross-section along the line indicated by B-B' in the layout in figure 4.52.

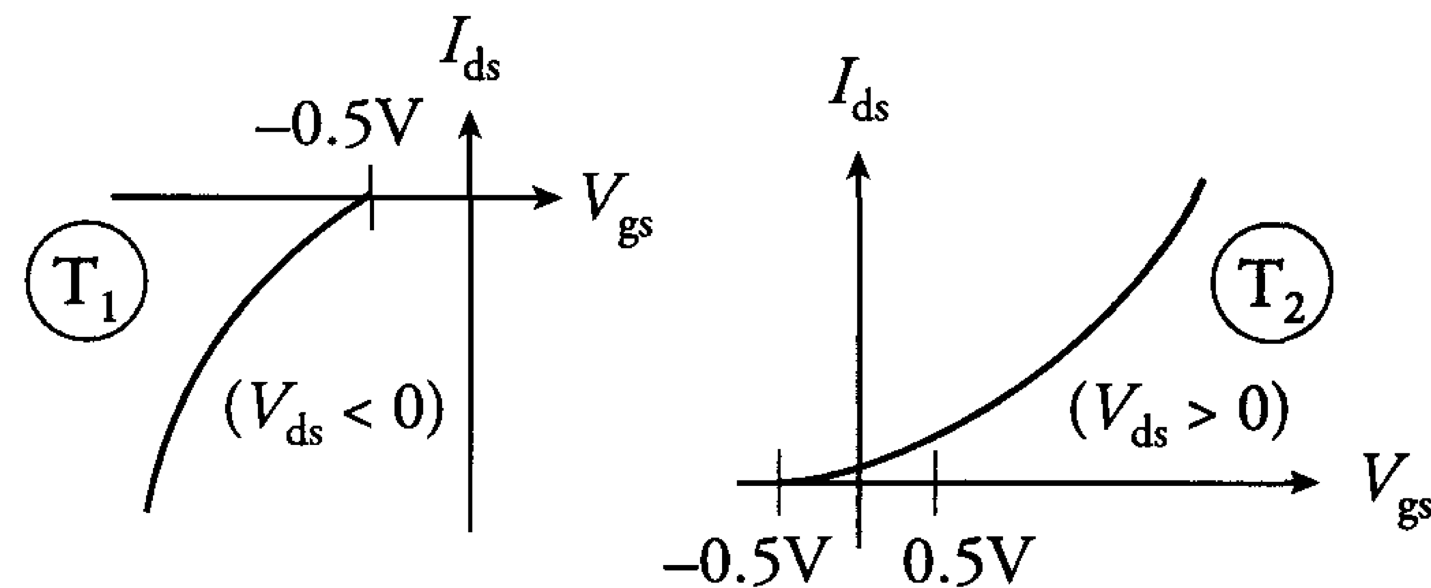
7. Transistors T_1 and T_2 in the adjacent circuit have the same aspect ratio $(\frac{W}{L})$ and a K-factor equal to zero.



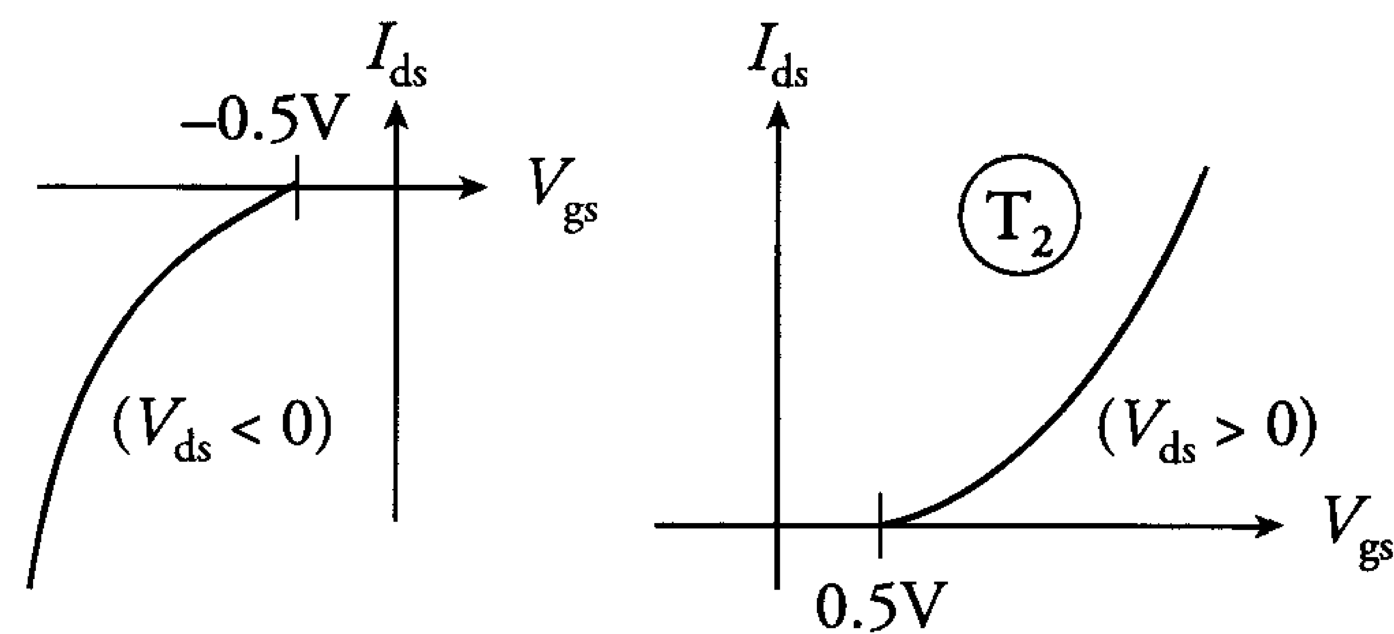
a) Sketch $V_{out}=f(t)$ when the $I_{ds}=f(V_{gs})$ characteristics for T_1 and T_2 are as follows:



b) Repeat (a) for the following characteristics:



c) Repeat (a) for the following characteristics:



All parasitic capacitances may be neglected in this exercise.

8. The following function must be implemented in a CMOS logic circuit: $z = c(\bar{a}\, b + a\, \bar{b})$

   a) Draw a circuit diagram of a static CMOS implementation of the required logic circuit. The required inverse signals must also be generated in this circuit.

   b) Adopt the approach presented in this chapter and draw the CMOS stick diagram and layout of the logic circuit
   Assume $(\frac{W}{L})_n = \frac{1.5}{0.25}$ and $(\frac{W}{L})_p = \frac{2.5}{0.25}$ (scale: $0.1\,\mu$m≡ 2mm).

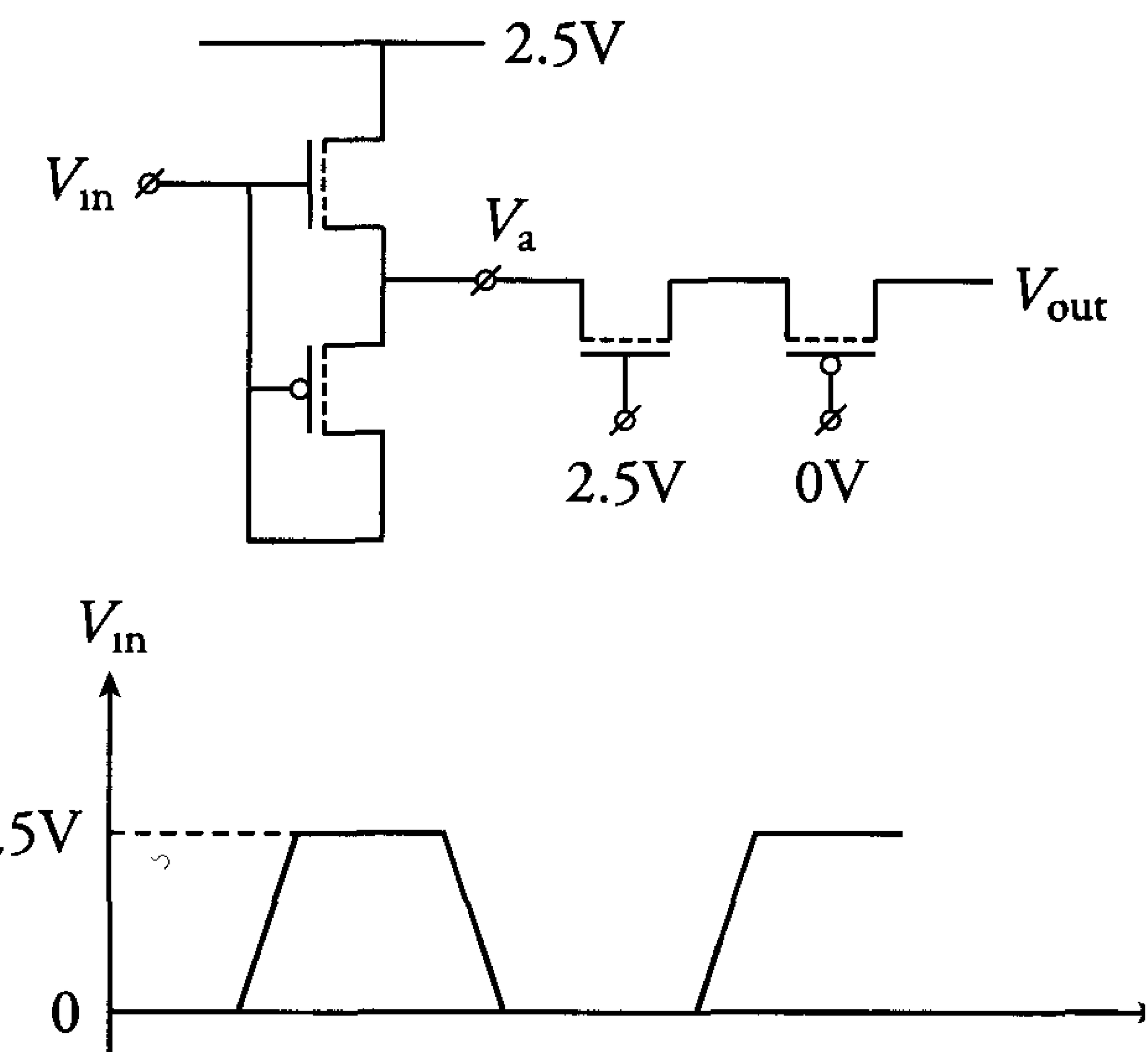
9. Consider the following logic function: $z = c + a\, b + \bar{a}\, \bar{b}$

   a) Rewrite this function such that it is optimised for implementation in MOS.

   b) Draw a circuit diagram of a static CMOS implementation.

   c) Adopt the approach presented in this chapter and draw the CMOS stick diagram and layout of the logic circuit
   Assume $(\frac{W}{L})_n = \frac{1.5}{0.25}$ and $(\frac{W}{L})_p = \frac{2.5}{0.25}$ (scale: $0.1\,\mu$m≡ 2mm).

10. The following values are given for the parameters in the adjacent circuit: $V_{X_n} = 0.5\,$V
$V_{X_p} = -0.6\,$V
$K_p = K_n = 0\,\text{V}^{-1/2}$
$V_{bb} = -2\,$V



Explain what happens to voltages V_a and V_{out} when V_{in} switches as shown. Draw V_{in}, V_a and V_{out} in one diagram.

Chapter 5

Special circuits, devices and technologies

5.1 Introduction

This chapter discusses a number of special circuits, devices and technologies. The circuits and devices can be used in digital, analogue and mixed analogue/digital applications. They are realised in various MOS technologies or their derivatives, which include the BICMOS technologies discussed in this chapter.

The chapter begins with an explanation of circuits that operate as image sensors. We distinguish *charge-coupled devices* (CCDs) as well as *CMOS image sensors*. Their ability to capture images finds its usage in all kinds of cameras. Their operation is based upon the conversion of light into electrons.

The second category of special devices covered in this chapter are MOS transistors capable of delivering high power. These MOS field-effect transistors, or *power MOSFETs*, are possible as a result of improvements in technology, which now enable the manufacture of transistors capable of withstanding high voltages and large current densities. Power MOSFETs obviously operate according to the same field-effect principle as ordinary MOS transistors. This principle is discussed in chapter 1.

Finally, devices based on mixed bipolar and CMOS technologies are discussed. These *BICMOS technologies* are relatively recent developments. They can enhance the performance of both digital and mixed analogue/digital circuits.

5.2 CCD and CMOS image sensors

5.2.1 Introduction

Charged-coupled devices (CCDs) are almost only used in imaging circuits. They basically operate by transferring a charge from one transistor gate to another in a 'channel'. CCD implementations include *surface-channel* (SCCD) and *buried-channel* (BCCD) devices. Also, for analogue applications, there must be a relationship between the size of the packet and the signal which it represents. The packet size must therefore be maintained during transfer. An alternative to CCD imaging is CMOS imaging, which is currently enjoying increased attention.

5.2.2 Basic CCD operation

CCD shift registers can be realised with 2-phase, 3-phase and other multi-phase clock systems. The operation of a CCD is explained below with the aid of the 2-phase SCCD structure shown in figure 5.1. A diagram of the 2-phase clocks ϕ_1 and ϕ_2 is also shown in this figure. The gates indicated by bold lines are polysilicon *'storage gates'*, under which charge is stored. The remaining gates are *'transfer gates'* created in a second polysilicon or metal layer. They lie on a thicker oxide than the storage gates and therefore have a much higher threshold voltage ($V_T \approx 2\,V$). The transfer gates serve as a barrier between the storage gates. Operation of the 2-phase SCCD is explained on the basis of the surface potential distributions under the gates.
Suppose the first and third storage gates contain a full and an empty charge packet, representing the logic levels '1' and '0', respectively.

The charge packet corresponding to the first storage gate is then full of electrons. This is represented by a full *'charge bucket'* under the gate in figure 5.1. The charge bucket corresponding to the third storage gate, however, is almost empty, i.e. it is practically devoid of electrons. At time point 1, both ϕ_1 and ϕ_2 are 'low' and the storage gates are separated from each other. At time point 2, ϕ_1 has switched from a low to a high level and the charge is transferred from the ϕ_2 storage gates to the ϕ_1 storage gates. At time point 3, both ϕ_1 and ϕ_2 are 'low' again and the charge is now stored under the ϕ_1 storage gates. The description of the shift behaviour at time points 4 and 5 is obtained by replacing ϕ_1 by ϕ_2 in the above descriptions for time points 1 and 2, respectively.

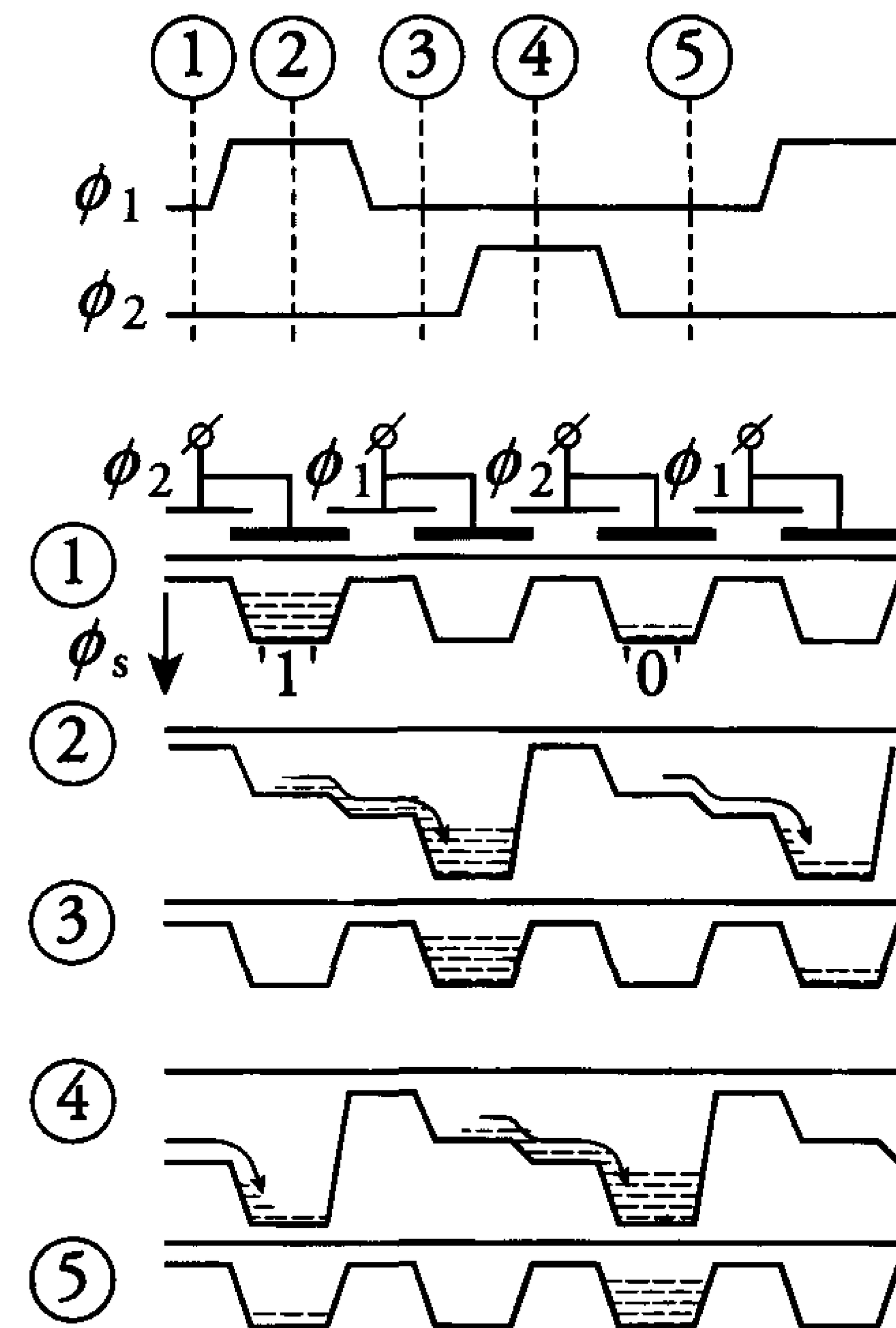


Figure 5.1: *The shift operation in a basic 2-phase SCCD*

A comparison of the time points 1 and 5 in figure 5.1 shows that the charge has been transferred from the first to the third bucket in one complete clock period. In fact, the charge is transferred from one CCD *'cell'* to another during a single clock period. Each cell clearly requires two storage elements which each comprise a bucket, a *transfer gate* and a *storage gate*. The two storage elements in a CCD cell are analogous to the master and slave latches in a *D-type flip-flop*. Clearly, the implementation of a 2-phase CCD register comprising 100k bits, for example, requires 200k storage elements on a chip. In practice, a better ratio between the number of storage elements and the number of register cells is obtained by using another type of clocking strategy.

The discussion of charge transfer in figure 5.1 is based on the assumption that one bucket was full with electrons and another one was empty. The operation of an SCCD clearly relies on the filling of these buckets. Figure 5.2(a) shows a simplified SCCD comprising some sensor cells and an output section.

In an image sensor photons reach the silicon surface through a lens. The silicon then converts the photons into electrons locally. A complete image is then captured in an array, which is read out by shifting (transferring) its contents to the CCD array output. The *charge transfer* in an SCCD occurs right at the silicon surface under the gates.

Unfortunately, the surface is inhomogeneous and therefore plagued by *surface states*. These surface states have a certain energy and can trap electrons which have higher energy. During charge transfer, the associated change in surface potential profile causes the surface states to release the trapped electrons. If this occurs before transfer is complete, then the released electrons will simply rejoin the rest of the electrons in the packet and *'transfer efficiency'* is maintained. However, if an electron is released from a surface state when transfer is complete, then it cannot rejoin its charge packet. This reduces transfer efficiency. The surface states continue to release the trapped electrons until a new charge packet arrives. The new packet will not be degraded by surface states that are still full when the packet arrives. The empty surface states will, however, be filled by the new packet and the process will repeat itself.

Clearly, transfer efficiency depends on the number of surface states. In previous generations of CCDs, transfer efficiency was increased by using a small charge to represent a '0'. This *'fat zero'* ensures that surface states remain filled. Transfer efficiency is also reduced by incomplete transfer of charge packets at high clock frequencies.

Leakage current accounts for another problem related to CCDs and, of course, to other dynamic memories as well. This *'dark current'* is caused by thermal generation of minority carriers and slowly fills the buckets of a CCD. The result is a *'maximum storage time'*, during which the data in a CCD will remain correct. In addition, dark current causes a fixed noise pattern on the data that is read from a CCD.

Both transfer efficiency and dark current largely determine the operating limits of a CCD. These factors therefore require considerable attention during CCD design.

The above section clearly indicates that surface states form an important limiting factor for the performance of SCCDs. These surface states are unavoidable. Therefore, the only way to improve performance is to realise a CCD in which storage and transfer of charge occurs in a channel which is 'buried' a short distance below the silicon surface. A buried n-channel can be realised by creating a thin n-type layer on top of a p-type substrate. Compare the SCCD and BCCD structures in figure 5.2(a) and (b) respectively.

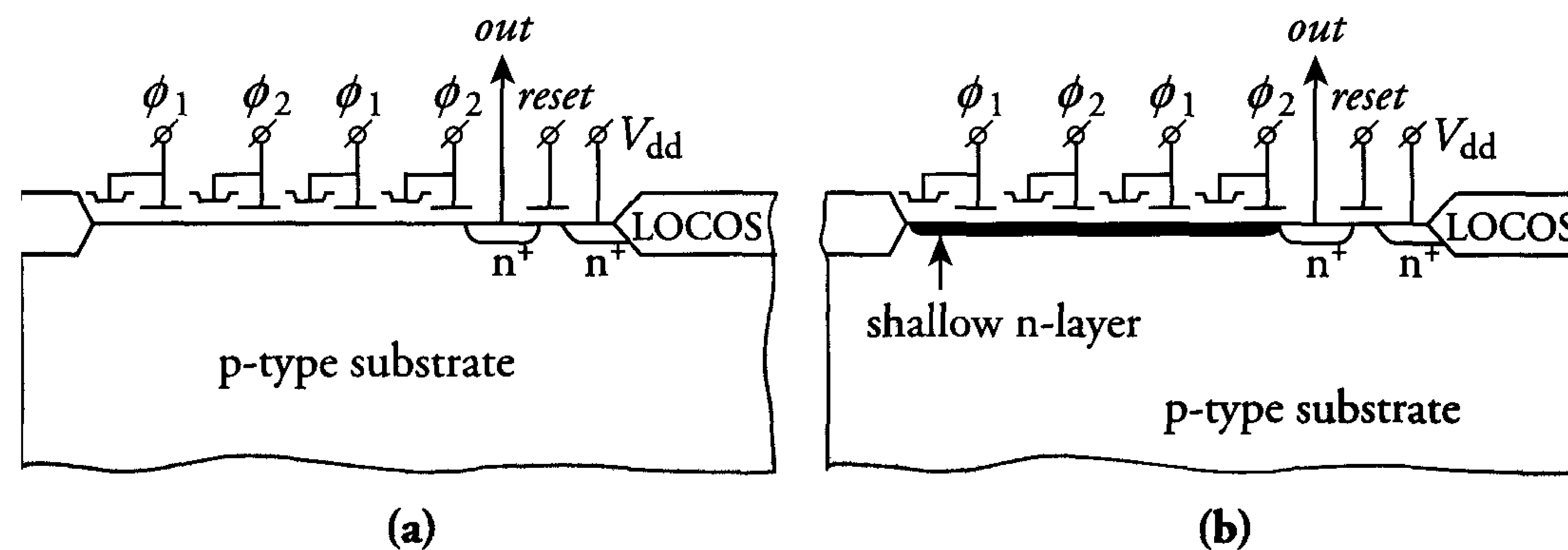


Figure 5.2: *(a) Basic structure of an SCCD and (b) a BCCD*

The operation of an SCCD is closely related to the characteristics of a MOS capacitor with a thick depletion layer. However, the operation of a BCCD is related to the characteristics of a MOS capacitor with a fully depleted layer. Therefore, the first requirement for the successful operation of a BCCD is that the thin n-type layer is fully depleted of electrons. This is achieved by using clock and control signals with an offset voltage. This voltage produces a potential maximum a short distance below the silicon surface. Electrons (representing data bits) injected into the device are stored at this potential maximum. The depleted n-type layer prevents the charge carriers from reaching the surface states and a high transfer efficiency is therefore achieved. The operation of a BCCD is otherwise identical to that of an SCCD.

Buried-channel CCDs were developed for two important reasons. The first is their *immunity* to surfaces states. The second is the increased *operating frequency* which they afford compared to surface-channel CCDs. The increase is caused by the fact that charge is transferred at a speed determined by the bulk mobility instead of the surface mobility. The maximum clock frequency of a BCCD is therefore twice that of an SCCD of equivalent dimensions. However, the definition of the buried channel in a BCCD requires an extra mask. BCCDs are also subject to many problems when their dimensions are reduced. In addition, it is inherently difficult to control the charge in a BCCD because it is stored at a distance from the gate which is longer than for an SCCD. Currently, all image sensor CCDs are implemented as BCCDs. Because of the large number of pixels, a lot of transfers are required. The immunity to surface states then outweighs the disadvantages of BCCDs.

The charge-coupled device principle can be used in both analogue and

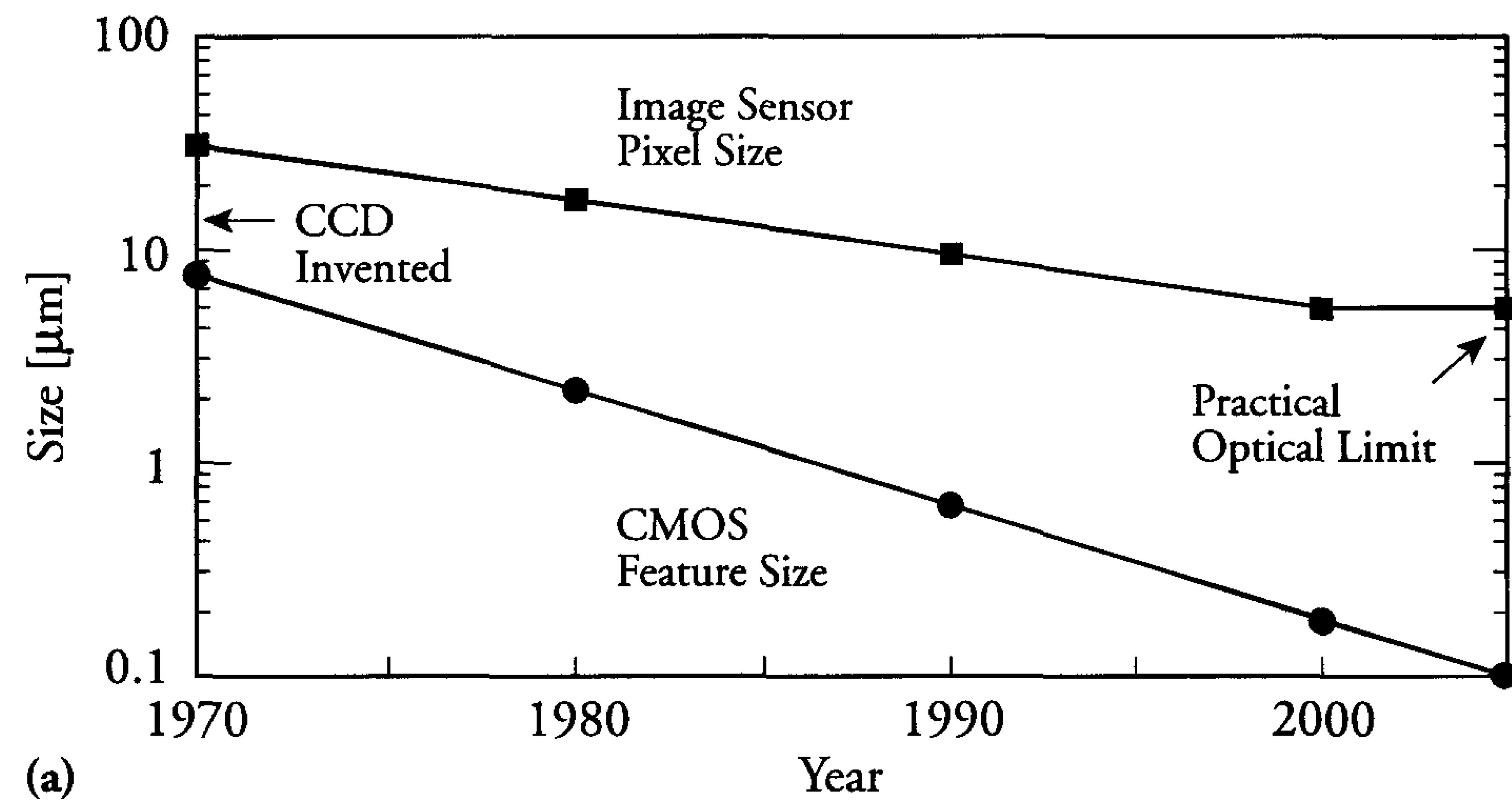
digital applications. As stated, the bulk part of the applications is in image sensors. Professional sensors now consist of 63 Megapixels. In video camera applications, both CCD sensor and storage device are integrated on one single chip, see figure 5.4.

5.2.3 CMOS image sensors

MOS image sensors already exist since the late 1960s. Due to problems with noise, sensitivity, scalability and speed, CCD sensors have become much more popular. In the early 1990s however, CMOS image sensors regained their popularity. The efforts were driven by low-cost, single-chip imaging systems solutions. Today the developments in, and applications of CMOS imaging have intensified so much that complete sessions at the major IC conferences, like IEDM and ISSCC, are devoted to them [14].

Another driving factor for an increased activity in CMOS image sensors is the continuous improvement in CMOS technology. Scaling of the sensor pixel size is limited by both optical physics and costs [12] and occurs at a lower pace than the scaling of the CMOS feature size, see figure 5.3(a). This allows to combine the CMOS image sensor with image processing on a single chip at relatively lower costs.

The ability to capture low-light images depends on the efficiency to convert incoming photons into electrons, which subsequently discharge the pixel capacitor. We distinguish between both passive and active pixels. An *Active Pixel Sensor (APS)* includes an active amplifier in every pixel. Figure 5.3 also shows three different pixels. When the pass transistor in figure 5.3(b) is accessed, the photodiode is connected to a bit line. Its charge is converted into a voltage by the readout circuit (amplifier) located at the bottom of a bit line. Due to the small pass gate, this single transistor pixel allows the smallest pixel size and consequently, the highest *fill factor* (ratio of sensor area to total area of sensor plus support electronics). The performance of a pixel was improved by adding active amplifier circuitry to the cell, see figure 5.3(c), resulting in average fill factors between 20% and 30%. The photogate APS in figure 5.3(d), integrates charge under the gate. Its readout architecture looks simular as in CCDs [12].



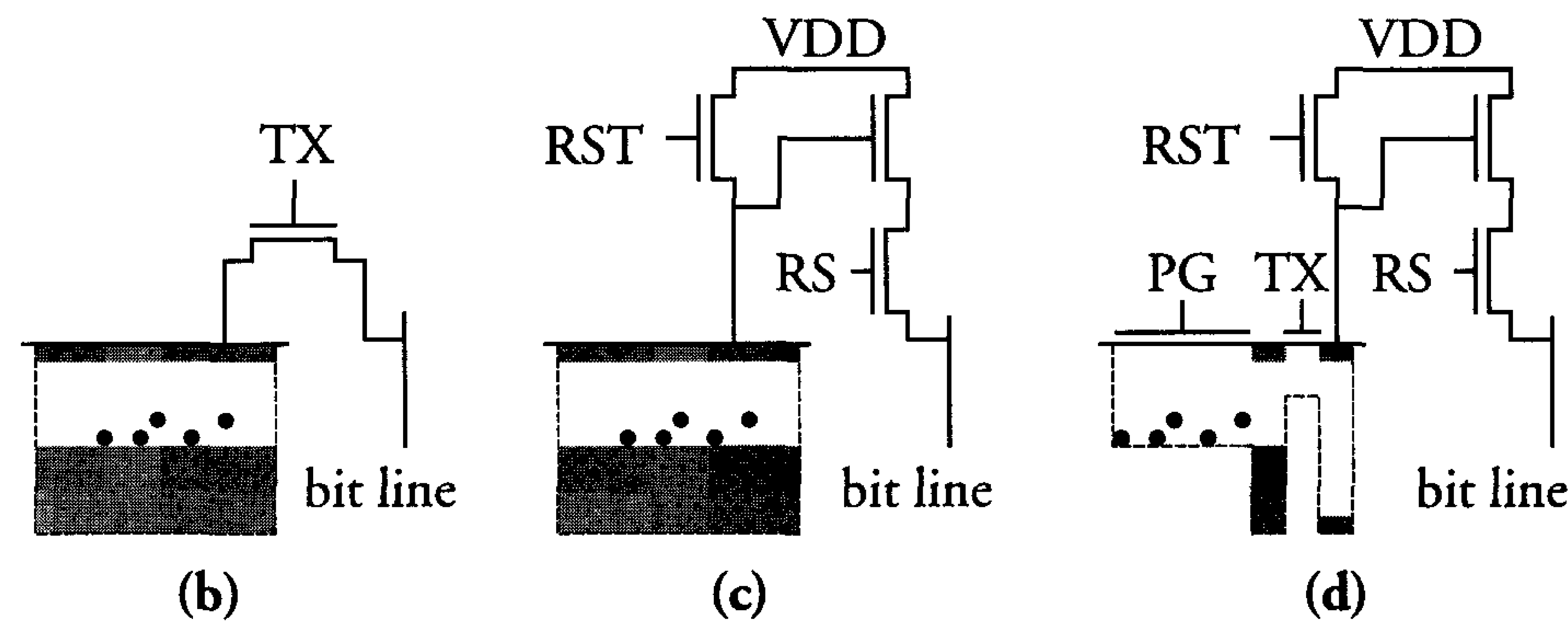


Figure 5.3: *(a) Scaling of MOS pixel and feature size, (b) passive pixel, (c) photodiode active pixel sensor (APS) and (d) photogate APS pixel. (source: [12])*

The low power consumption, high level of integration and low production costs allow CMOS sensors to be used in a variety of applications: multimedia, low cost cameras [13], PC camera, security and machine monitoring, video phone, fax, etcetera. More sophisticated and high-resolution imaging applications will become available as CCD and CMOS imagers continue to improve [14].

Figure 5.4: *A CCD image sensor (photo: PHILIPS)*

5.3 Power MOSFET transistors

5.3.1 Introduction

The invention of the bipolar junction transistor in 1947 provided the foundation for modern integrated circuits and power electronic circuits. The first power devices based on semiconductor technology were demonstrated by Hall in 1952. He used germanium stacked junctions to achieve a continuous forward current of 35 A and a punch-through voltage of 200V. Since about 1955, silicon has been preferred for power devices. By 1960, such junctions allowed the implementation of 500 V rectifiers. Currently, silicon rectifiers are available with continuous current ratings of 5000 A and reverse voltages of 6000 V. The application of MOS technology in *power transistors* has been a major focus point for the industry since the late seventies.

The prospects of high speed and high input impedance in many low-voltage applications are particularly attractive. *Double-diffused MOS transistors* were originally introduced during the mid-seventies. The *DMOS transistor* allowed increased performance without reducing the source-drain distance, whilst excessive electric fields were avoided. Originally, the introduction of DMOS power FETs was seen as a major threat to the bipolar power transistor. However, their advantages only render *power MOSFETs* suitable for a limited part of the power electronics application area.

The advantages of power MOSFETs compared to bipolar power transistors are as follows:

- Infinite current gain.
  The infinite impedance leads to a DC input current which is zero. This advantage is offset by the need for large driver circuits to drive the high input capacitance presented by a MOS power transistor.
- High switching speed.
  There is a considerable charge stored in the large base of a bipolar high-voltage transistor. This results in a switching time of about 0.5 μs compared to a few nanoseconds in MOS.
- *Safe operating area (SOA).*
  - The output current of a bipolar transistor has a positive temperature coefficient. For a constant base-emitter voltage, an increase in current results in a temperature increase. This in turn causes a larger current and a destructive *'hot spot'* is eventually created. One solution is to include a thermal stabiliser in the circuit by adding an emitter resistance. This *emitter ballasting* increases the SOA.
  - The transistor with the highest temperature in a group of parallel bipolar transistors will eventually take all the current. This is because of the positive temperature coefficient of the output current of a bipolar transistor. The transistors will burn out successively if the current is excessive. However, emitter ballasting can be used to prevent the above *'current hogging'* and facilitate parallel connection of bipolar transistors to increase driving capability.
  - The *output current* of a MOS transistor has a negative temperature coefficient. A higher temperature therefore leads to a higher on-resistance and to a reduced current. This self-regulating process ensures that MOS transistors can simply be connected in parallel.

The disadvantages of MOS power transistors compared to bipolar power transistors are as follows:

- A higher on-state voltage drop.
  This is because no injection of minorities occurs; it is a problem for thick high-voltage devices, even though a MOS device can be used up to $BV_{\mathrm{ds}} = BV_{\mathrm{cb}}$. The use of power MOS transistors was therefore generally limited to high-frequency and low-voltage (<200 V) circuits, where the on-state voltage drop is acceptable. For higher voltages, injecting devices such as MOS-controlled thyristors have been developed.
- Higher production costs.
  The width of the emitter of a bipolar high-voltage transistor may be between 50 μm and 100 μm. It can therefore be created without using accurate and fine lithography. A MOS high-voltage transistor, however, requires several accurate lithographic and processing steps to create the required double diffusion, gate oxide, polysilicon and planar high-voltage passivation.

During the past ten years, improvements in technology and yield have resulted in better performance for MOS power transistors. Breakdown voltages over 1000 V are now possible. The *breakdown voltage* V_{B} of a

power MOSFET is related to its typical resistance ($R_{on} \cdot Area$). Typical corresponding values might be ($R_{on} \cdot Area$) = $40\,\Omega{\cdot}\mathrm{mm}^2$ at $V_B = 1000$ V for one power MOSFET and ($R_{on} \cdot Area$) = $0.4\,\Omega{\cdot}\mathrm{mm}^2$ at $V_B = 100$ V for another. A power MOSFET of the first type with a die size of $10\,\mathrm{mm}^2$ has an on-resistance R_{on} of $4\,\Omega$. The power dissipated in such a device is determined by the product of its on-resistance and the square of its current, i.e. $P = R_{on} \cdot I^2$. In practice, power dissipation is limited by the maximum power rating of the power MOSFET's package. Figures between 100 W and 350 W have been realised for existing packages. The above power MOSFET with $R_{on} = 4\,\Omega$ could therefore be packaged as a device with a current capability of about 3 to 6 A. MOS power devices with die sizes of $200\,\mathrm{mm}^2$ have been reported in the literature. Large-area low-voltage devices are designed for use as synchronous rectifiers, replacing diodes in power supplies (e.g. in PCs and laptops). When the current levels of power devices exceed about 1 A at operating voltages in excess of 500 V, monolithic integration of the power devices with the rest of the circuit is no longer cost effective.

5.3.2 Technology and operation

The *vertical DMOS transistor* (VDMOST) shown in figure 5.5 is an example of a discrete power MOSFET.

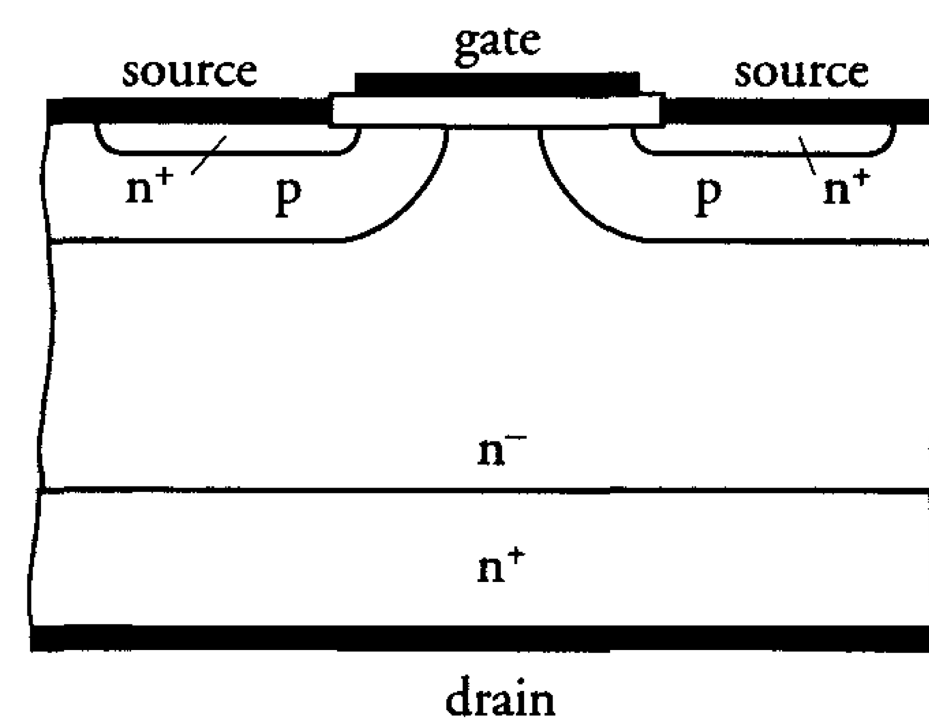


Figure 5.5: *Cross-section of a VDMOS discrete power transistor*

The threshold voltage of the above VDMOST is generally 2 to 3 Volts. When the gate voltage is increased from 0V to about 12V, the p-well area near the surface is inverted (see section 1.3). A channel then exists between the n^+ source and the n^- epitaxial layer. The charge carriers

will flow *vertically* from the source to the drain when a high voltage is applied to the latter. The drain voltage can vary from 50 V to 1000 V, depending on the application. VDMOS transistors are usually n-type rather than p-type because of their higher channel mobility. Because of the scaling of the gate oxide thickness, devices with a gate voltage of 5V and a threshold voltage of around 1.2V become available as well.

5.3.3 Applications

Power MOSFETs can be used as discrete power switches in fluorescent lamp ballasts and switch-mode power supplies. In electrical shavers, they are used both in the form of discrete devices and as part of larger integrated circuits, e.g. automatic supply voltage adaptors and battery chargers. Their high current capability makes power MOSFETs suitable for use in driver circuits, e.g. for stepper motors and plasma display panels, etc. Power MOSFETs are easily integrated in bipolar and BICMOS circuits because they do not inject minority carriers. The combination of low-voltage bipolar transistors and high-voltage lateral DMOS transistors of both n and p types facilitates production of analogue high-voltage circuits [3]. Examples include video output amplifiers [4] and [5].

5.4 BICMOS digital circuits

5.4.1 Introduction

Since the mid-eighties, a growing interest in the technology has resulted in a lot of commercially available BICMOS ICs. The *BICMOS technology* facilitates a combination of both bipolar and CMOS devices on a single IC and enables the simultaneous exploitation of the advantages of both device types.

Initially, the penalty of more complex processing restricted the use of BICMOS technologies to fairly specialised applications. Modern bipolar and CMOS technologies, however, have become very similar and both are quite complex. Processes are therefore possible in which a large number of steps for the manufacture of MOS and bipolar devices are identical. These BICMOS processes can be economically viable and no longer need be a technologist's nightmare. It is estimated that a BICMOS wafer will cost 20% to 30% more than a CMOS wafer. In several applications, this price increase will be offset by the performance enhancement.

Performance characteristics of BICMOS devices and their technology are explained below. Future expectations and market trends are also discussed. The discussions are mainly focused on VLSI applications.

5.4.2 BICMOS technology

There are several ways of obtaining a BICMOS process. It could, for instance, be based on an existing bipolar process or a completely new BICMOS process could be developed. The usual approach, however, is to start from a CMOS process. An associated advantage is that existing CMOS design and CAD tools can then be used for BICMOS designs. A BICMOS process based on an n-well CMOS process is considered here. This is a logical choice because of the considerable similarities between this BICMOS process and the n-well CMOS process discussed in chapter 3.

The development of the BICMOS process from an n-well CMOS process is explained with the aid of the cross-sections in figure 5.6. The source and drain diffusions are typically a few tenths of a micron deep. The depth of the n-well varies between two and three microns. The realisation of an npn transistor requires an additional p-type diffusion in the n-well. This diffusion forms the base of the npn transistor and is shown in figure 5.6.

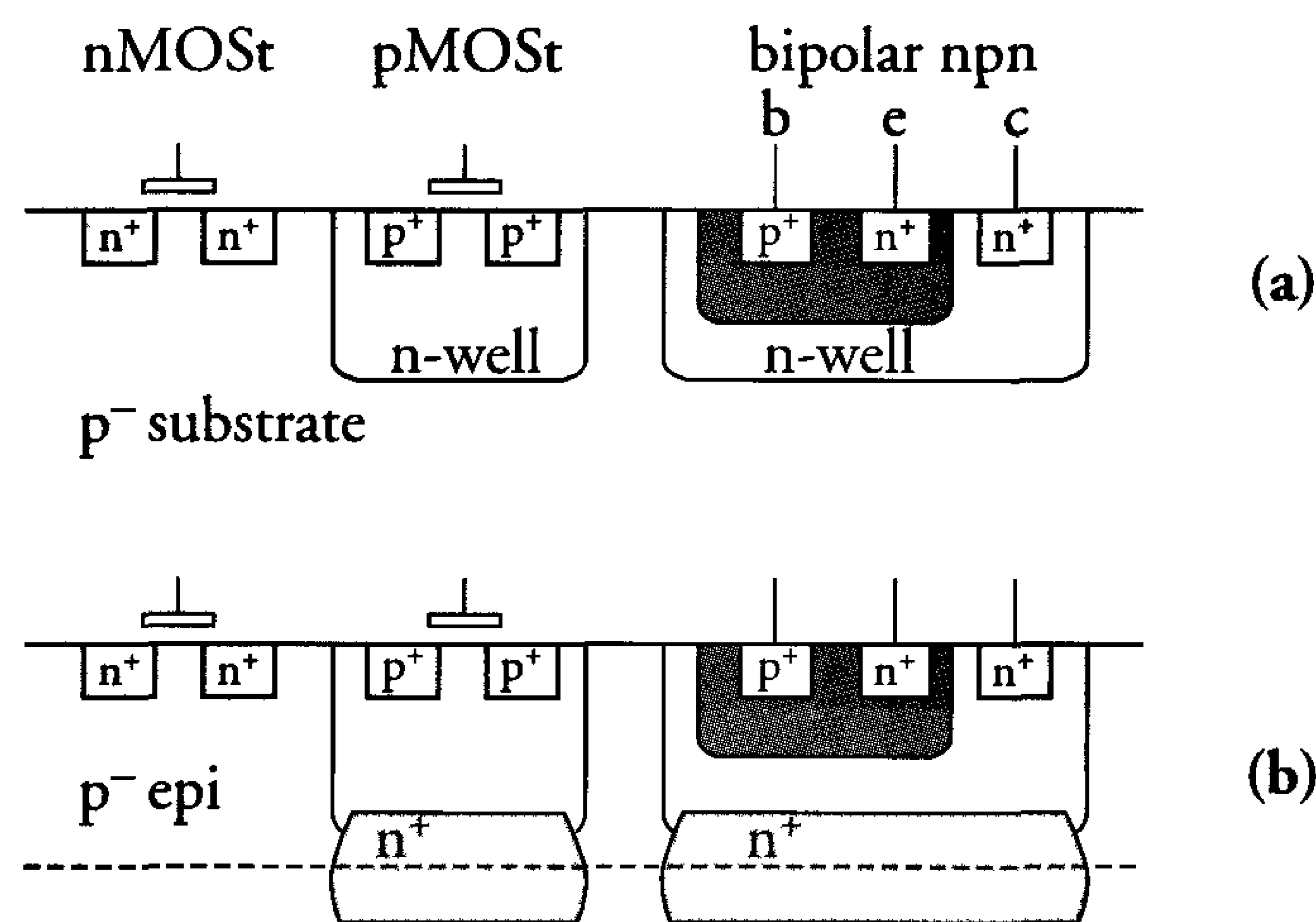


Figure 5.6: *The development of a BICMOS process from an n-well CMOS process*

The npn transistor exhibits a relatively high collector series resistance. This is also the base resistance of the pnp transistor in the *parasitic thyristor*, as discussed in chapter 9 (latch-up). This 'simple' structure is therefore rather susceptible to *latch-up*.

The above disadvantages are largely overcome when the structure shown in figure 5.6b is used. In the associated process, n^+ diffusions are created in the p-type substrate prior to the growth of a p^- epitaxial layer. The resulting 'buried layer' n^+ areas subsequently become part of the n-wells. The npn transistor obtained in this process is basically isolated and latch-up via the parasitic pnp transistor is largely prevented by the n^+ buried layer. The creation of the buried collector areas and the base diffusion requires two more masks than in a standard n-well CMOS process.

BICMOS technology is being developed by many companies and institutes. Significant progress has been achieved in terms of density and performance. Most manufacturers use structures similar to figure 5.6b. Refinements, of course, are also used. BICMOS circuits based on bipolar processes and p-well CMOS processes are also encountered. Standard VLSI BICMOS requires between 15 and 20 masks. Options may increase this number to about 25. The average extra cost of BICMOS compared with CMOS is about 30%, depending on the number of metal layers.

5.4.3 BICMOS characteristics

Its higher *gain factor* and lower *noise* generally renders bipolar technology more suitable than CMOS for analogue applications. However, CMOS is more attractive for digital control, storage and signal processing circuits because of its low quiescent power, reasonable performance and high packing density. The mixture of the two technologies offers unique possibilities in both analogue, digital and mixed analogue/digital applications.

BICMOS was first introduced in digital I/O circuits, where it provided increased output driving capability. It was subsequently applied in the peripheral circuits of SRAMs to shorten the access times. These circuits included sense amplifiers, word line and bit line drivers.

Low-voltage bipolar transistors and high-voltage lateral DMOS transistors, incorporating both n-type and p-type channels, are combined in some BICMOS processes. These processes allow the integration of truly analogue high-voltage circuits, such as the video output amplifiers mentioned in section 5.3.3.

The previously-mentioned applications of BICMOS technologies illustrate their potential benefits. However, in addition to the increase in costs compared to an average CMOS technology, there are other drawbacks associated with BICMOS. For instance, the CMOS digital parts of a BICMOS chip may generate considerable transient noise on the supply and ground lines. This 'bounce' is discussed in chapter 9. Considerable efforts are required to prevent it from entering analogue parts of the chip. Moreover, the reduced density of BICMOS logic limits its usage to critical functions on a VLSI chip. This reduces the potential performance advantage. The commercial use of BICMOS technology for digital ICs is therefore only justified when the additional costs are compensated by increased performance.

5.4.4 BICMOS circuit performance

BICMOS logic gates usually employ CMOS transistors to perform the logic function and bipolar transistors to drive the output loads. The two typical BICMOS implementations of a *NAND gate* shown in figure 5.7 illustrate this two-stage structure.

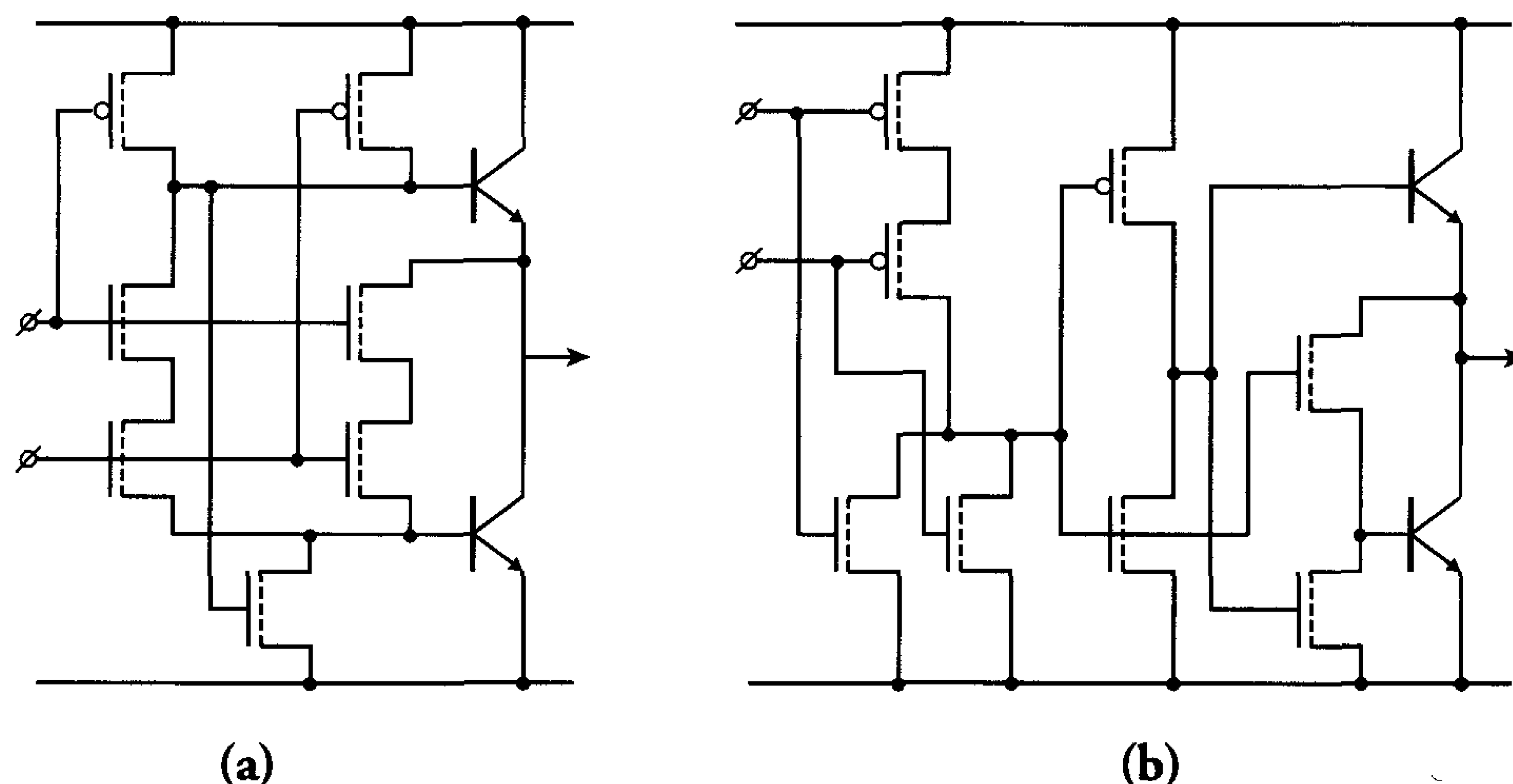


Figure 5.7: *Typical BICMOS implementations of a NAND gate*

The two-stage structure of a BICMOS logic gate leads to a larger propagation delay for an unloaded BICMOS gate than for its CMOS counterpart. The performance advantage of a BICMOS implementation over a CMOS implementation therefore only applies in the case of gates with larger fan-outs. Figure 5.8 shows a frequently published comparison of the propagation delay as a function of fan-out for typical CMOS and BICMOS NAND gates. The comparison was made for nMOS and pMOS transistor widths of $4\,\mu$m and $7\,\mu$m, respectively, in a process with a $0.35\,\mu$m gate length. The cross-over point lies between a fan-out of two and three. For higher fan-outs, the performance of a BICMOS circuit is better.

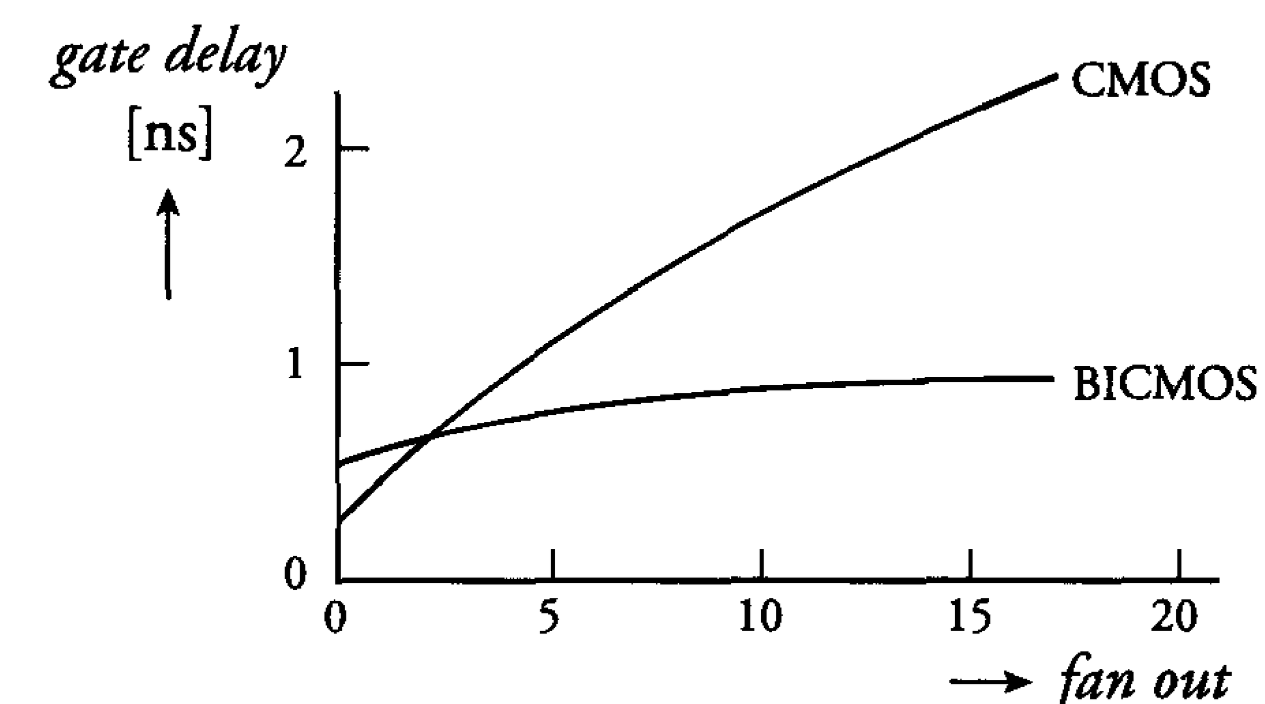


Figure 5.8: *Gate delays of typical CMOS and BICMOS NAND gates*

Figure 5.8 suggests that BICMOS is preferable to CMOS even for relatively low fan-outs. For large capacitive loads, the figure shows that the propagation delay can be reduced by a factor of 2.5 when BICMOS is used. However, the presented comparison does not account for the extra area required by the driver stage in the BICMOS implementation. A more representative comparison is obtained when the CMOS logic gate is also equipped with a CMOS output driver. The resulting comparison is shown in figure 5.9 for BICMOS and CMOS NAND gates implemented as NOR gates followed by bipolar and CMOS drivers, respectively. Such a comparison shows a dramatic reduction in speed advantage and reveals that BICMOS only affords a small performance improvement for gates with a high fan-out. In practice, this means that implementation of logic gates in BICMOS is not cost effective for low to medium speed applications. Its usage in VLSI circuits and *Application-Specific ICs* (ASICs) is therefore limited to circuits that have to drive large capacitances, e.g. driver and I/O circuits. BICMOS is also used in ICs that have to operate beyond the performance limits of CMOS.

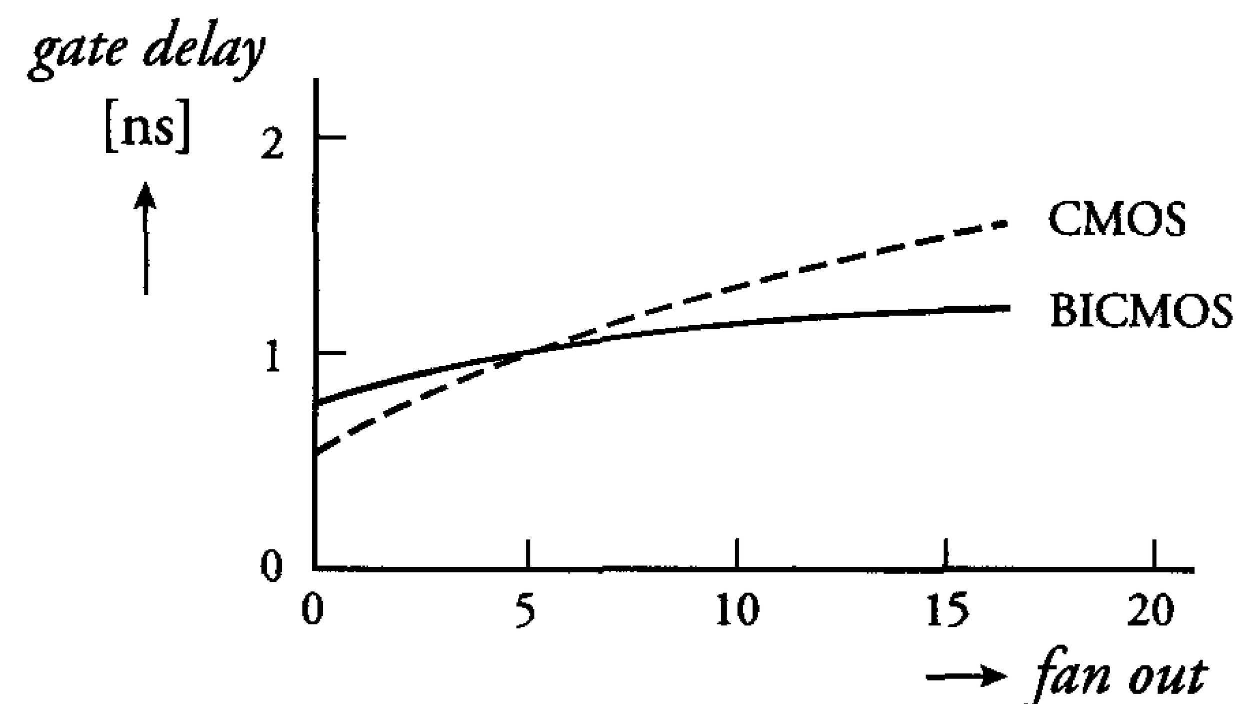


Figure 5.9: *Propagation delays of CMOS and BICMOS NAND gates implemented as NOR gates with CMOS and bipolar drivers, respectively*

Supply voltage dependence, temperature dependence and process parameter dependence are also important factors that must be included in a comparison of the performance of BICMOS and CMOS circuits. These factors are compared below.

CMOS current levels are quadratically reduced when the supply voltage is reduced. This results in a reduction of the speed of both CMOS and BICMOS circuits. Bipolar circuits, however, are also hampered by inefficient operation at lower supply voltages. Manufacturers of BICMOS ICs will therefore face a dilemma when supply voltage standards below 2.5 V become more accepted as minimum feature size decreases to below 0.25 μm. Innovative design may reduce this dilemma.

The influence of temperature on the performance of CMOS and BICMOS circuits is closely related to the different origins of transistor currents. In bipolar transistors, the current is caused by diffusion. This current is less affected by temperature than the MOS transistor drift current discussed in section 2.3. As a consequence, the switching speed of BICMOS is less dependent on temperature than that of CMOS.

It has been empirically found that variations in CMOS parameters caused by processing spread have a greater influence on circuit performance than variations in bipolar process parameters.

Finally, it should be noted that a BICMOS driver implementation shows a reasonable power dissipation advantage over a CMOS driver.

It is clear that the application of BICMOS technology is not trivial. This explains its limited application in digital products.

5.4.5 Future expectations and market trends

From a performance point of view, the future for BICMOS technologies originally looked promising. This was especially true in the high-end market for ASICs and SRAMs. However, a fair comparison of BICMOS and CMOS circuit performance reveals that the advantages afforded by BICMOS are really only significant in mixed analogue/digital circuits and in digital ICs in driver and I/O type of circuits.

In addition, the performance advantages are achieved at the expense of increased process complexity. This leads to increases of up to 30% in process costs. It is therefore obvious that application of BICMOS is only worthwhile for a limited part of the digital IC market. The temporary increase around 1996 was caused by plunging MOS memory prices (of DRAMs in particular) and to a growing BICMOS IC market (high demand for BICMOS-based Pentium™ chips).

This conclusion is verified by market research [10] (see figure 2 in the Preface), which shows that the use of BICMOS has increased from around 2% of the total IC market in 1991 to about 18% in 1996.

In the high-end part of the memory market, BICMOS circuits are only expected to be used in peripheral circuits to reduce access times. At voltages of 2.5V and below, the bipolar transistor gain is relatively low and BICMOS technologies lose their advantages. Thus, the total BICMOS market is expected to decline to about 4% of the total IC market in 2001.

5.5 Conclusions

A number of devices and technologies that can be used in both purely digital as well as mixed analogue/digital ICs are discussed in this chapter. Because this is the only link between the presented topics, no general conclusions are presented here. The reader is therefore referred to the application sections associated with the CCD and MOS power transistor topics and the section on future expectations and market trends associated with the BICMOS topic.

5.6 References

Charge-coupled devices (CCDs)

[1] C.H. Sequin and M.F. Tompsett,
'Charge transfer devices',
Volume supplement 8 of Advances in Electronics and Electronic Physics, Academic Press, New York, 1975

[2] J.D.E. Beyon and D.R. Lambeds,
'Charge-coupled devices and their applications',
McGraw-Hill, London, 1980

[2a] A.J.P. Theuwissen,
'Solid-State Imaging with Charge-Coupled Devices',
Kluwer Academic Publishers, 1995

[2b] Gerald C. Holst,
'CCD ARRAYS, CAMERAS and DISPLAYS',
JCD Publishing/SPIE Optical Engineering Press, 1998

MOS power transistors

[3] A. Ludikhuize, 'A versatile 250/300V IC process for Analog and Switching Applications',
IEEE Trans. on Electron Devices, Vol. ED-33, pp 2008-2015, December 1986

[4] P. Blanken, P. van der Zee,
'An integrated 8MHz video output amplifier',
IEEE Trans. on Consumer Electronics, Vol. CE-31, pp 109, 1985

[5] P. Blanken, J. Verdaasdonk,
'An integrated 150 V_{pp}, 12kV/μs class AB CRT-driving amplifier',
Digest of Technical Papers, 1989, ISSCC, New York

[5a] B.E. Taylor, 'Power Mosfet Design',
John Wiley & Sons, 1993

[5b] B.J. Baliga, 'Power ICs in the Saddle',
IEEE Spectrum, July 1995, pp 34-49

BICMOS

[6] P.A.H. Hart,
'BI(C)MOS Dream or Nightmare?',
Proceedings of the 17th ESSDERC '87 Conference in Bologna,
14-17 September 1987, pp 187-194

[7] G. Koetzke,
'VLSI Technology Trends',
Proceedings of Comp Euro '89 Conference in Hamburg

[8] B. Santo,
'BICMOS circuitry: the best of two worlds',
IEEE Spectrum, May 1989, pp 50-53

[9] A.R. Alvarez,
'BICMOS Technology and Applications',
Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993

[10] W.J. McClean,
'Status 1999, A report on the IC industry',
ICE corporation, Scottsdale, Arizona, 1999

[11] S.H.K. Embabi, A. Bellaouar, M.I. Elmasry,
'Digital BICMOS Integrated Circuit Design',
Kluwer Academic Publishers, 1993

CMOS Image sensors

[12] Eric R. Fossum,
'CMOS Image Sensors: Electronic Camera-On-A-Chip',
IEEE Transactions on Electron Devices, Vol. 44, October 1997

[13] Ulrich Ramader, et al.
'Single-Chip Video Cameras With Multiple Integrated Functions',
ISSCC Digest of Technical Papers, 1999, pp 306-307

[14] 'Image Sensor' Session at the ISSCC conferences:
ISSCC Digest of Technical Papers, 2000 and onwards

5.7 Exercises

1. A dynamic shift register can be implemented as discussed in the chapter on CMOS circuits. It can also be implemented as a charge-coupled device (CCD). What are the main differences between the former implementations and the CCD implementation? State advantages and disadvantages to support your answer.

2. Assume that the transfer of a logic '1' through an SCCD is represented by a full charge packet. Explain what happens if the temperature increases when a series of data bits consisting of a hundred '1's, one '0' and again a hundred '1's, i.e. 111...1111011111...111, is transferred through the device.

3. Explain the main differences between a low-voltage MOS transistor which operates at 2.5 V and a power MOSFET.

4. A gate array consists of a large number of fixed transistors and BICMOS input and output (I/O) circuits. A single BICMOS output circuit is too weak to drive a certain large output load capacitance. Explain how the driving capability can be improved.

5. Explain why BICMOS circuits exhibit a longer propagation delay than their CMOS counterparts for small capacitive loads and a shorter propagation delay for large capacitive loads.

Chapter 6

Memories

6.1 Introduction

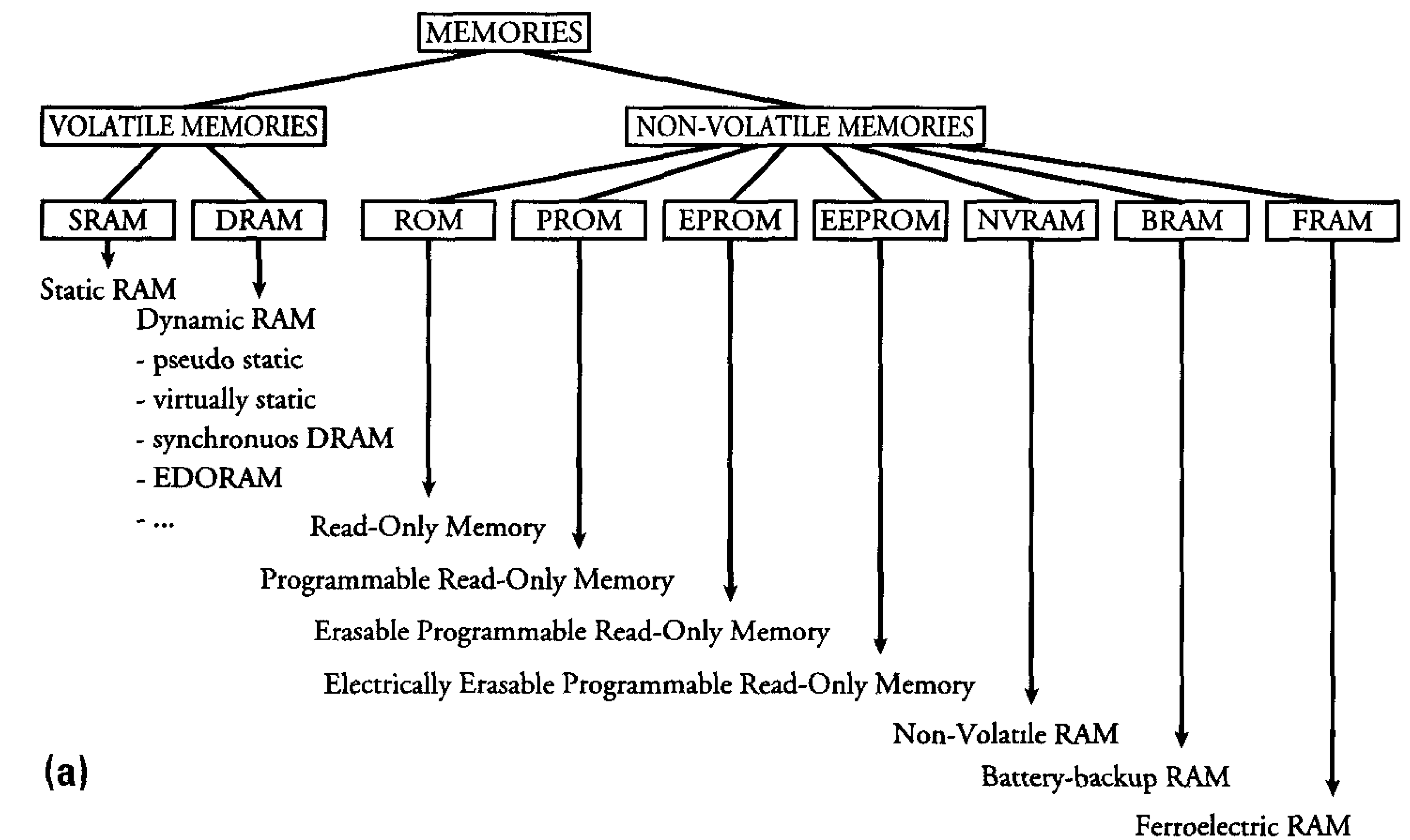
Memories are circuits for the storage of digital values. A memory may constitute a single IC or be part of a larger IC. These types are referred to as *stand-alone* and *embedded* memories, respectively.

The digital values in a memory are each stored in a *'cell'*. The cells are arranged in a *matrix* or *array*, which affords an optimum layout.

The *'data retention time'* of a memory is the period for which it keeps its data when the supply voltage is removed. Memories that lose their data when power is removed are referred to as *volatile*. Memories that retain their data in the absence of power are called *non-volatile* memories. A finer division of memories yields the following three types:

- Serial memory;
- Random-access memory (RAM);
- Read-only memory (ROM).

Figure 6.1 presents an overview of the various implementation possibilities for memories. This figure also shows the respective market shares in 1998 and the expected shares in 2001. The increased market share gained by the DRAMs is mainly the result of the rise of new high-speed architectures, which make them particularly suited for the growing high memory bandwidth applications such as hard disk drives, graphics boards and printers, etc.



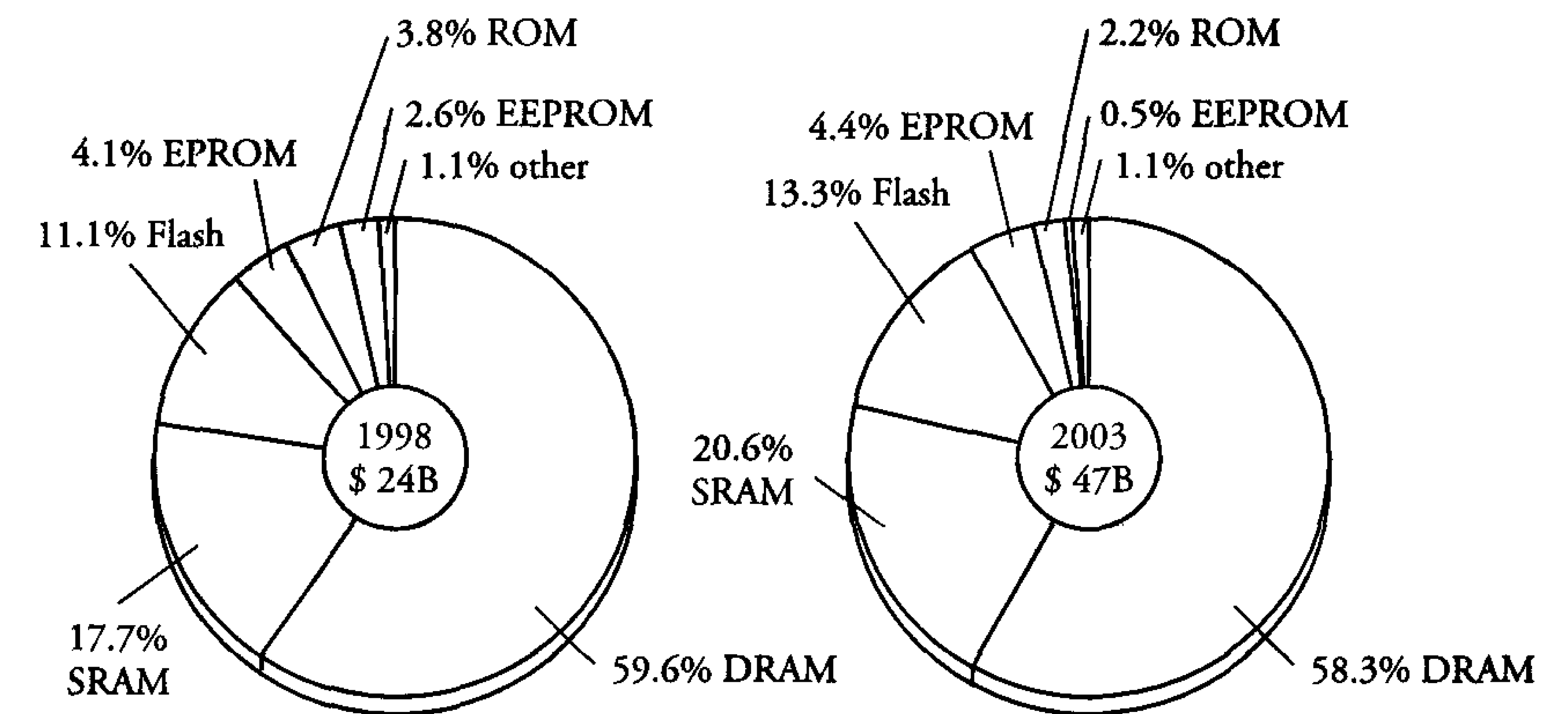


Figure 6.1: *(a) Overview of different types of memories.*
(b) Memory market shares in billions of dollars in 1998 and expected shares in 2003 (source: ICE).

Volatile memories include *'static'* and *'dynamic'* RAMs. Electrical feedback in the memory cell of a *static RAM* (SRAM) ensures that voltage levels are maintained and data is retained as long as the power supply remains. The data in a *dynamic RAM* (DRAM) memory cell is stored as a charge on a capacitor. Gradual leakage necessitates periodic refreshing

of the stored charge. A dynamic RAM that internally refreshes its own data is called a *pseudo-static* or *virtually static* RAM.

The cells in serial memories form one or more shift registers, which can each store a 1-bit data stream. The 'first in, first out' (*FIFO*) operation of shift registers ensures that data enters and leaves a serial memory in the same sequence. Examples of their use include delay lines in video applications.

The cells in a RAM or ROM array must have individual unique *'addresses'*. Alternatively, they may be connected in parallel groups. In this case, each group or *'word'* has a specific address. The capacity of a RAM or ROM that is divided into words is specified by the number of words and the number of bits per word. Examples are 64 M×4, 32 M×8 and 16 M×16. These three specifications all refer to a 256-Mbit memory, which can store over 2,000 newspaper pages or 4 hours of radio data.

The data in a ROM can only be read, whereas the data in a RAM can be written and read. The sequence in which data is read from a ROM or RAM is unrestricted. Therefore, *access* is in fact *random* in both cases. The term RAM, however, is generally only used to refer to memories that allow reasonably high frequency read and write operations at random locations.

A RAM requires both data and address inputs and data outputs. Figure 6.2 is a general schematic representation of an addressable memory.

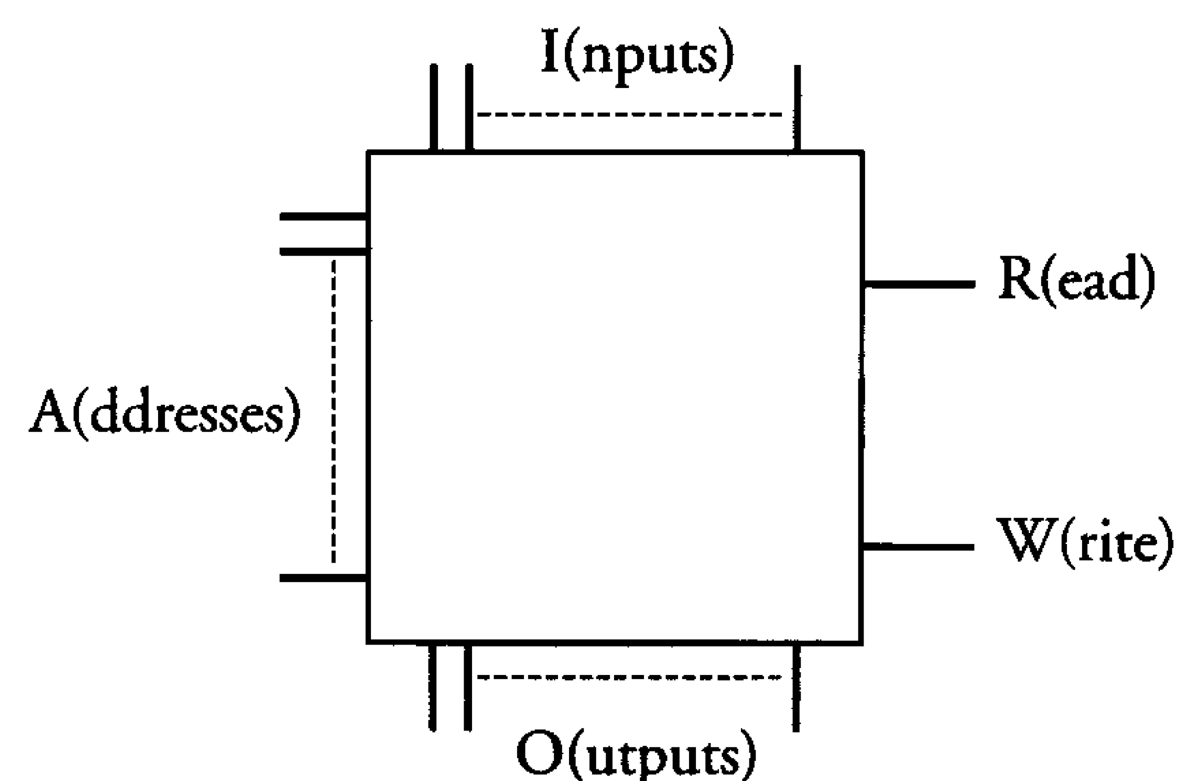


Figure 6.2: *General representation of a memory*

The memory shown is obviously a RAM. The read (R) and write (W) inputs are often combined in one single input which controls the mode

of operation. A ROM requires no data inputs but does require address inputs and data outputs. The schematic of a ROM is therefore obtained if the data (I) and W inputs in figure 6.2 are removed. The schematic of a serial memory is obtained if the address inputs are removed.

The *'access time'* of a memory is the time interval between the initial rising clock edge in a read operation and the moment at which the data is available at the output terminals. The *'cycle time'* of a memory is the minimum possible time between two successive accesses. The cycle time of an SRAM may be greater than, smaller than or equal to its access time, while the typical cycle time of a DRAM is about twice as long as the access time. This is because the accessed cells in a DRAM must be refreshed after each read and write operation. Although access times are often used for comparison of the different memories available from different manufacturers, cycle time comparison would be better for benchmarking purposes.

The various types of memories are discussed in this chapter. Their operation and properties are explained and possible applications are given. A brief discussion of the structure of a simple 4k-bit SRAM provides considerable insight into memory operation.

6.2 Serial memories

Serial memories are usually encountered in the form of static or dynamic shift registers. Modern *video memories* are an important exception. These memories are serial by nature and random access is therefore not required. However, they are implemented as DRAMs, in which cells are serially accessed. Such a memory is sometimes called a *video RAM* or *VRAM* (see section 6.3.4).

Serial memories may be implemented using the CMOS shift register cells presented in chapter 4. The extensive discussions on shift registers in chapter 4 makes further elaboration on serial memories unnecessary.

6.3 Random-access memories (RAM)

6.3.1 Introduction

Random-access memories can be subdivided into the two following classes:

- Static RAM (SRAM);
- Dynamic RAM (DRAM).

These two types of RAM are discussed separately below. The basic operation of a RAM is explained with the aid of a 4k-bit SRAM. A subsequent discussion of the major differences between SRAMs and DRAMs illustrates the considerable difference in their operation.

6.3.2 Static RAMs (SRAM)

A true static memory is characterised by the elapse of a certain time between a change on its address inputs and the presence of valid bits at its data outputs. Dynamic memories often require a considerably more complex pulse pattern with very stringent timing requirements.

SRAM block diagram

For most stand-alone SRAMs, every possible combination of address inputs is decoded in a random-access memory. A memory with n address inputs therefore contains 2^n addresses. An *SRAM* with twelve address inputs, for example, therefore has at least 4096 memory cells. Figure 6.3 shows the block diagram of such a 4k-bit SRAM. Its 4096 memory cells are organised in an array of 64 rows and 64 columns. Each row and column can therefore be addressed by 6 address inputs. In addition to an array of memory cells, an SRAM also requires control logic circuits. These circuits will now be described.

- A *row decoder* selects the *'word line'* x_i of the row in which the addressed cell is located. The row decoder is also known as an *x-decoder.*

- A *column decoder* selects the *'bit line select'* line y_j of the column in which the addressed cell is located. The column decoder is also known as a *y-decoder.* The addressed cell is located at the point of intersection of the selected row and column and is referred to as cell x_i,y_j. The y_j signal selects the *bit lines* b_j and $\overline{b_j}$ of the addressed cell.

- *Address buffers* connected to the address inputs drive the row and column decoders. The output lines of the row and column address buffers traverse the length and width, respectively, of the array. They therefore form large capacitive loads for the address buffers.

- The tri-state *data input buffers* drive data buses db and $\overline{db}$ when the memory is being written. These buffers drive the large capacitive load of the data bus line and the selected bit line. They must also be capable of forcing the memory cell into another logic state.

- A *sense amplifier* detects the contents of the selected cell via the complementary bit lines b_j and $\overline{b_j}$ and data bus lines db and $\overline{db}$. The detection must occur as rapidly as possible so that the access time is reduced to a minimum. The sensitivity of the sense amplifier may be as low as 50 to 100 mV. Currently, current sensing has become more popular to further increase speed, even without switching the bit lines.

- The tri-state *data output buffer* transfers the data from the sense amplifier to the SRAM output when the memory is being read.

The SRAM control signals

The control signals required in an SRAM are described below. For the sake of simplicity, the commonly-used *output enable* (OE) signal is omitted.

- The *write enable* ($\overline{WE}$) signal determines whether data is written to the selected cell or read from it. With writing, the bit lines are driven from the input and, with reading, the bit line signals are transferred to the output.

- The *chip select* ($\overline{CS}$) signal facilitates selection of a single SRAM when many are combined to form a large memory system. Such a system consists of one or more *memory banks.* The memories in

such a system are connected to common address and data buses. Although more than one memory (or even a complete bank) can be selected at the same time, only one at a time can put data on the data bus. The $\overline{CS}$ signal of the relevant memory is activated by decoder logic in the memory bank. This logic produces 'high' logic levels on the $\overline{CS}$ inputs of the remaining memories. Their output buffers are therefore placed in the high-impedance mode and the memories are isolated from the data bus.

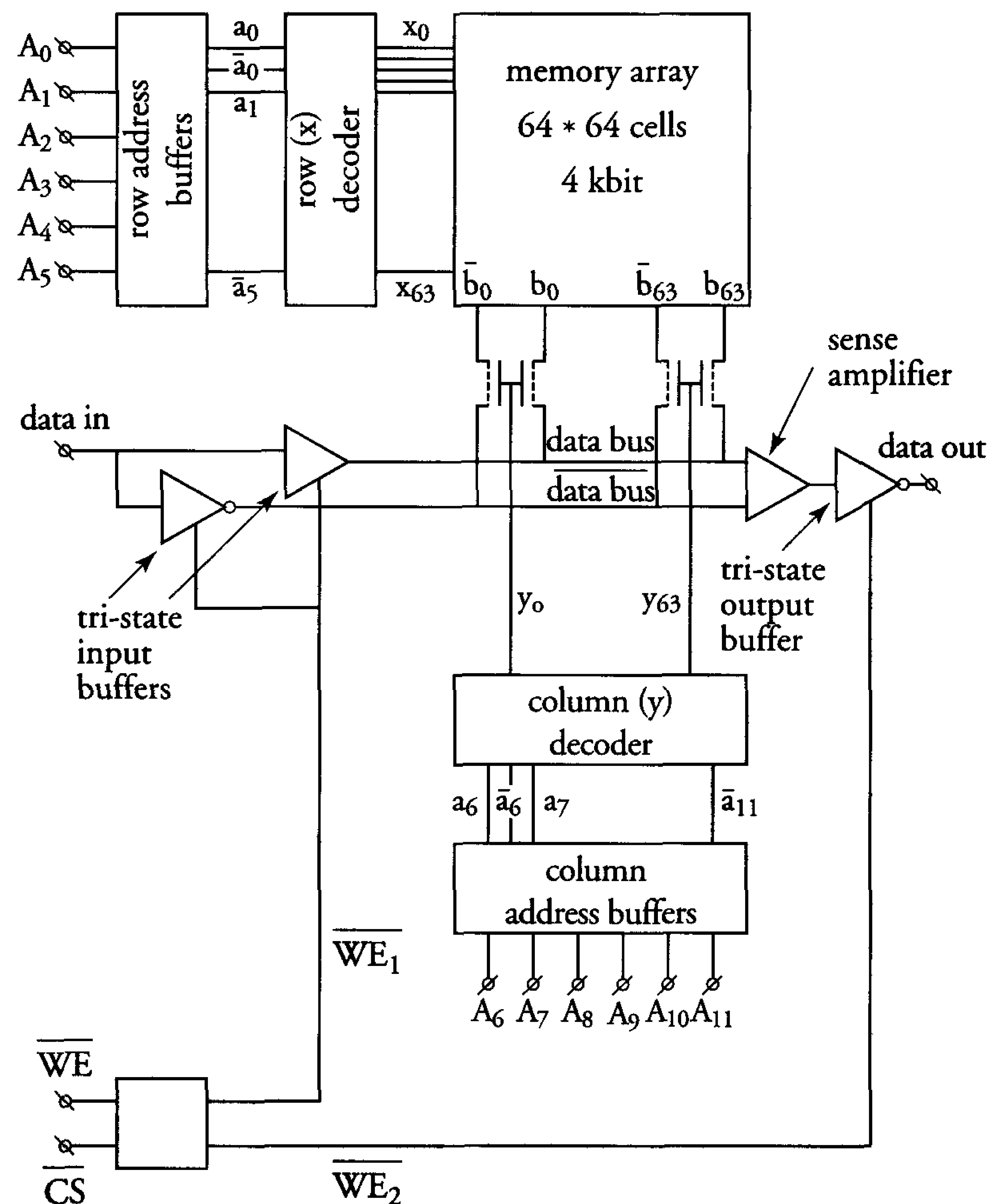


Figure 6.3: *Block diagram of a 4k-bit SRAM*

The SRAM read operation

The read operation performed in an SRAM is explained with the aid of an example in which data is read from the cell x_{35},y_{40}, see figure 6.3. The steps required to achieve this are as follows:

- The word line x_{35} is activated by placing the decimal value 35 on address inputs A_5 to A_0: $A_5A_4A_3A_2A_2A_1A_0 = 100011$.
- The bit line select signal y_{40} is activated by placing the decimal value 40 on the address inputs A_{11} to A_6: $A_{11}A_{10}A_9A_8A_7A_6 = 101000$.
- The $\overline{CS}$ signal is driven 'low' to select the memory.
- The $\overline{WE}$ signal is driven 'high' so that the information in the selected cell can be read via the selected bit lines, the sense amplifier and output buffer. The logic '1' on the $\overline{WE}$ signal activates the output buffer and places the tri-state input buffers in the high-impedance state. At the beginning of each read cycle, all bit lines b_i and $\overline{b_i}$ are precharged through clocked transistors to the high level (other precharge levels, such as half-V_{dd} or low (V_{ss}) levels are also used). If the value '0' is subsequently read from the selected cell, then bit line $\overline{b_{40}}$ remains 'high' while bit line b_{40} discharges slightly via the cell. The bit line voltage levels are transferred to the respective $\overline{db}$ and db data buses. The sense amplifier rapidly translates the resulting voltage difference to a logic '0', which is then transferred to the output via the buffer. A similar explanation applies when the value '1' is read from the selected cell.

The SRAM write operation

The write operation performed in an SRAM is explained with the aid of an example in which data is written to the cell x_{17},y_{15}, see figure 6.3. The steps required to achieve this are as follows:

- The word line x_{17} is activated by placing the decimal value 17 on the address inputs A_5 to A_0: $A_5A_4A_3A_2A_1A_0 = 010001$.
- The bit line select signal y_{15} is activated by placing the decimal value 15 on the address inputs A_{11} to A_6: $A_{11}A_{10}A_9A_8A_7A_6 = 001111$.

- The $\overline{CS}$ signal is driven 'low' to select the memory.
- The $\overline{WE}$ signal is driven 'low' so that the information on the data input can be written to the selected cell via the data input buffers and the selected bit lines. The value on the *db* data bus is then equal to the value on the data input while the $\overline{db}$ data bus has its inverse value. The logic '0' on the $\overline{WE}$ signal activates the input buffers and places the tri-state output buffer in the high-impedance state.

SRAMs are designed in a variety of synchronous and asynchronous architectures and speeds. An asynchronous SRAM is activated when an address change is detected. As a result, a clock signal is generated and stored data is accessed. However, this type of SRAM is limited in its speed. Therefore, the fastest SRAMs are generally synchronous. Controlled by one or more clocks, synchronous SRAMs show reduced access and cycle times, boosting their clock frequencies to the same height as those of the high-performance RISC processors and PCs. Improved performance can be achieved when several bits are selected simultaneously by a single address. In this *burst mode*, the address is incremented by an on-chip counter. Several burst addressing sequences can be supported, including those used in Pentium™ and PowerPC™ processors.

Static RAM cells

Access time is an important RAM specification and is mainly determined by the signal propagation time from the memory cell to the output. A satisfactory *access time* requires an optimum design of the memory cell, selection circuits, bit lines, sense amplifiers and output buffers. Possible *memory cell implementations* for SRAMs are discussed in detail below.

1. *Six-transistor/full-CMOS SRAM cell*
   Figure 6.4 shows a memory cell consisting of six transistors T_1 to T_6. Transistors T_1 to T_4 comprise two cross-coupled inverters which function as a latch. Pass transistors T_5 and T_6 provide access to the latch. During a write operation, the word line goes 'high' and the data on the bit lines is transferred to the latch through pass transistors T_5 and T_6. The word line also goes 'high' during a read operation. In this case, however, the contents of the latch cause a slight discharge on one of the precharged bit lines. The discharge takes place through the relevant pass transistor, T_5 or T_6, and inverter nMOS transistor, T_1 or T_3.

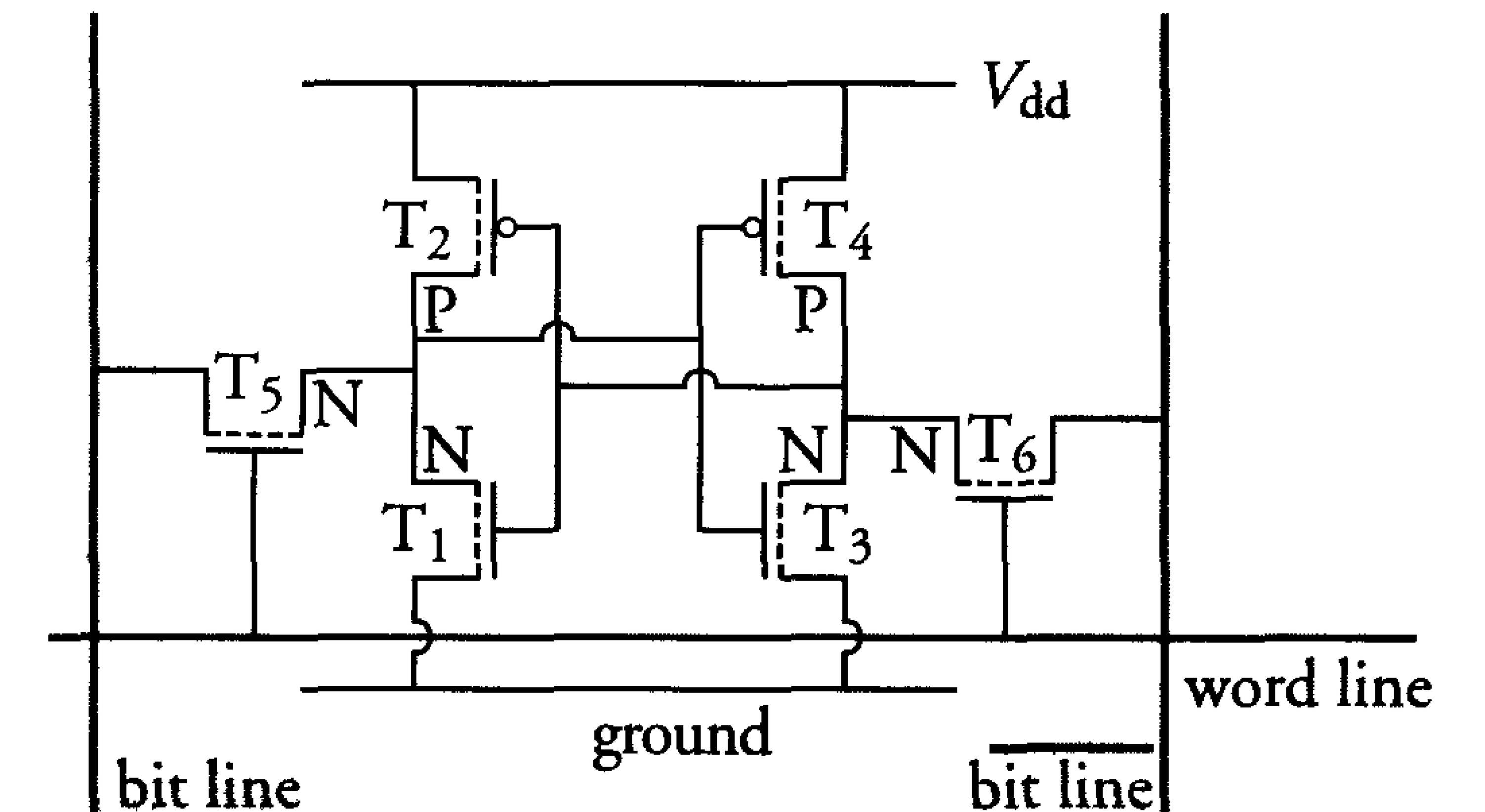


Figure 6.4: *Six-transistor static RAM cell*

A small voltage difference of about 50 mV between the two bit lines is sufficient for an SRAM sense amplifier to determine the logic level in the memory cell. This logic level is then transferred to the output pin via the output buffer.

The small leakage current in the parasitic diodes of the diffusions and a small sub-threshold current are the only currents that flow in the six-transistor cell when it is simply retaining data. Memories containing full-CMOS cells are therefore suitable for low-power applications. However, the relatively large distance required between nMOS and pMOS transistors requires quite a large chip area for this memory cell.

2. *Four-transistor/R-load SRAM cell*
   Figure 6.5 shows a memory cell consisting of four transistors. This cell contains two cross-coupled inverters with resistive loads. These types of inverters are discussed in section 4.2 and they lead to continuous static power dissipation in the memory cell. This dissipation is kept as low as possible by forming the resistors in an extra high-ohmic polysilicon layer. Typical values are 10 GΩ or more. This polysilicon layer necessitates a more complex manufacturing process than for the full-CMOS cell. An advantage of the four-transistor cell, however, is its reduced cell area, because the resistors are implemented in a second polysilicon layer and folded over the transistors.

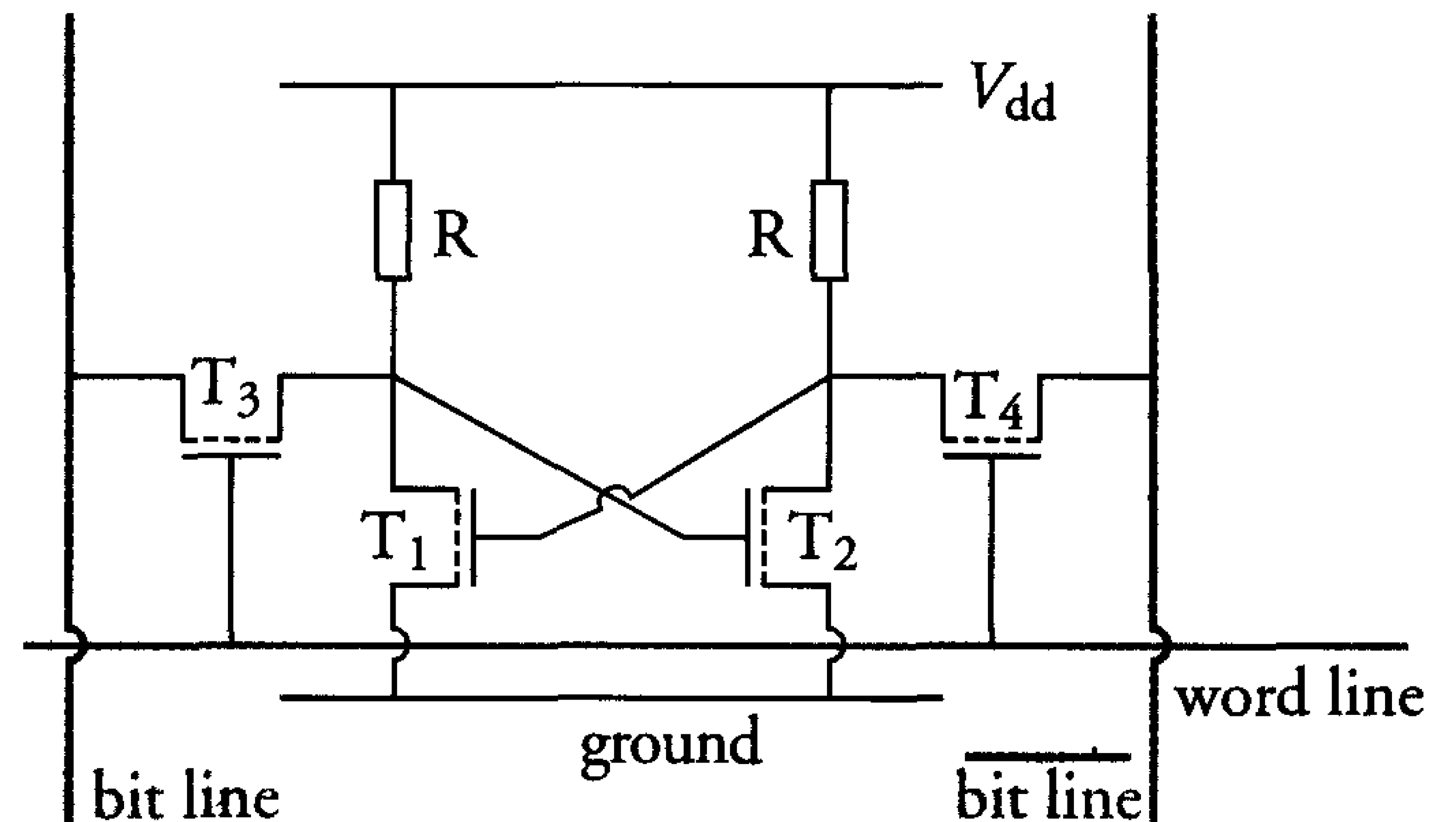


Figure 6.5: *Four-transistor static RAM cell*

The word lines in both the six-transistor and four-transistor memory cells are implemented in a stack of polysilicon and metal. The considerable parasitic capacitance and resistance of long word lines causes the cells furthest from a row decoder in an SRAM to exhibit a greater RC-delay than those closest to the decoder. This situation is often redressed by dividing the arrays of large memories into several smaller sections with separate row decoders between them. The resulting word lines have lower parasitic capacitance and resistance and their RC-delays are at least a factor four lower than for a single array. The silicides mentioned in chapter 3 are also used to reduce resistance of short polysilicon word lines and bit lines.

3. *Four-transistor loadless SRAM cell*
The introduction of a loadless four-transistor cell [14,15] allows a 35% area reduction using the same design rule. Comparing figure 6.5, in the loadless cell, the resistors R and the V_{dd} connection are completely removed and transistors T_3 and T_4 are replaced by pMOS transistors. This allows the cell nodes to store full-swing signals after writing. In the stand-by mode, bit lines are precharged to V_{dd} and the data is maintained in the cell when the leakage current of the pMOS transistors is more than an order of magnitude larger than that of the nMOS transistors. It is a promising alternative to the six-transistor cell, both for embedded as well as for stand-alone SRAMs.

6.3.3 Dynamic RAMs (DRAM)

The basic block diagram of a DRAM is quite similar to that of an SRAM. The main difference between an SRAM and a DRAM is the way in which information is stored in the respective memory cells. Most stand-alone DRAMs consist of n-type cells because of the high-performance requirements. DRAMs often use back-bias voltages to have a better control on the threshold voltage and to reduce junction capacitances. When DRAMs are embedded in a logic chip, p-type cells are often chosen, because the n-well in which the DRAM is located can then be separately connected to an additional positive back-bias to achieve the previous advantages.

Figure 6.6 shows the basic circuit diagram and a water model of a *single-transistor DRAM cell*, which is also called a *1 T-cell*. A logic '1' is written into the 1 T-cell by placing a high level on the bit line while the word line is active. The capacitor in the cell is then charged to a high level. This is also applicable with reverse polarities for p-type cells. The data in a cell is thus determined by the presence or absence of a charge on its capacitor. Parasitic junction leakage and transistor sub-threshold leakage cause this charge to leak away, just like the water in the pond evaporates as time progresses. The information in the cell must therefore be frequently refreshed.

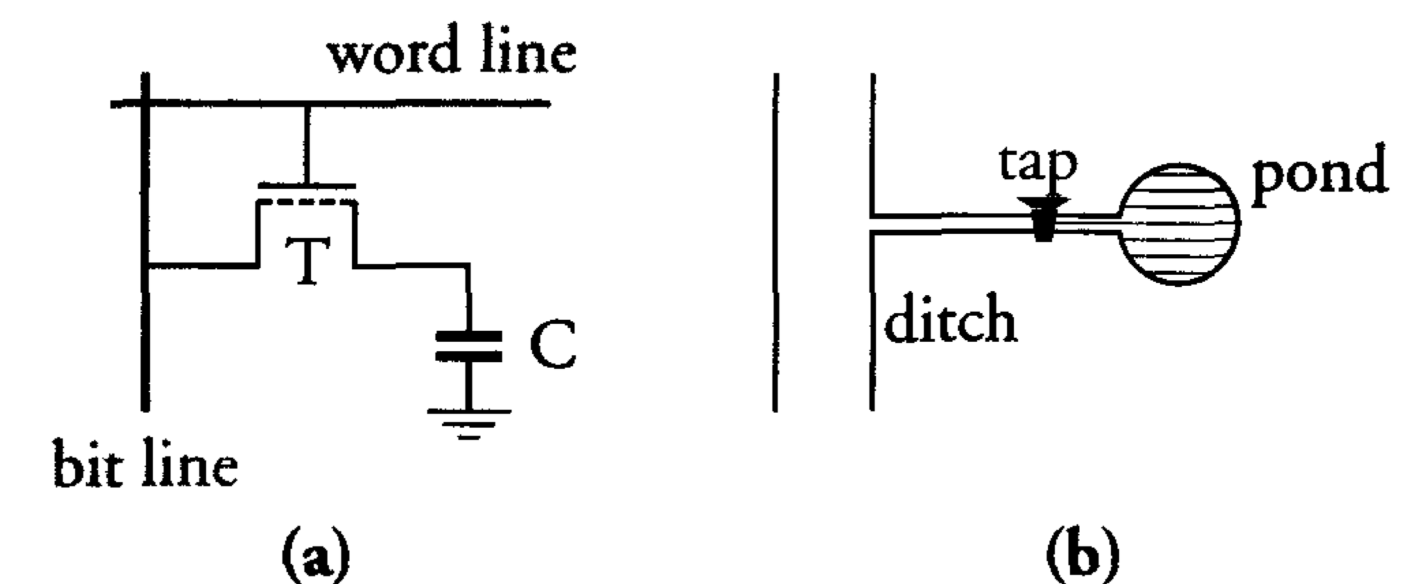


Figure 6.6: *(a) Circuit diagram of a DRAM cell (b) Water model of a DRAM cell.*

In addition to leakage, the information in a DRAM memory cell is also destroyed when it is read. This so-called *destructive read-out* (DRO) is caused by the cell capacitance being much smaller than the bit line capacitance. The cell contents must therefore be restored immediately after each read operation. For this reason, each bit line is equipped with a *refresh amplifier*, which consists of a sense amplifier and some restore

circuits. This sense amplifier detects the bit line level and writes its amplified value back into the cell. The operation is called a *row refresh* because it is done simultaneously for all cells that are addressed by an active word line.

In practice, the *refresh operation* must be performed every 2 to 256 milliseconds, depending on the cell size and the technology. A 64k-bit DRAM with an array comprising 256 words of 256 bits is considered as an example. If the data has to be refreshed every 2 ms and the cycle time is 0.4 μs, then it will take 0.1 ms to refresh all of the rows. This DRAM cannot therefore be accessed for approximately five percent of the time. This percentage is typically between one and five percent and is one of the reasons why DRAMs are more difficult to use than SRAMs.

The read operation in a DRAM requires a reasonable signal level on the bit line. This necessitates a minimum ratio between the cell capacitance (C) and the bit line capacitance (C_b). The required ratio depends on the following factors:

- the process tolerances, which cause sense amplifiers offsets,
- the sensitivity of the sense amplifier,
- the required noise margins.

Memory cell dimensions have become smaller for each successive generation of DRAMs. However, the cell capacitance has not scaled at all. Because of *α-particle sensitivity*, the cell charge has remained constant for five generations. The following ratio between C and C_b applies to recent generations of DRAMs:

$$\frac{1}{15} \leq \frac{C}{C_b} \leq \frac{1}{10}$$

Rather planar processes such as those described in chapter 3 are suitable for the implementation of DRAMs with capacities of up to 1 M-bit. A typical example of the *planar DRAM cell* used in these DRAMs is shown in figure 6.7.

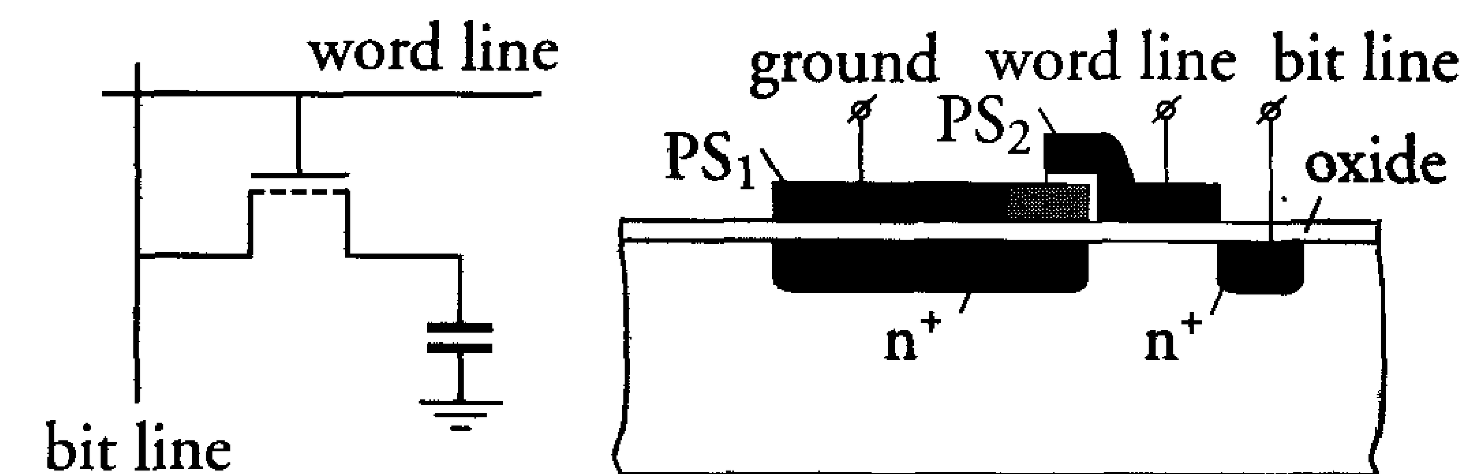


Figure 6.7: *The planar DRAM cell*

An unacceptably small capacitance renders planar cells unsuitable for current DRAMs. Three-dimensional cells which afford increased storage capacitance in a reduced planar surface area are therefore used for large DRAMs. These include the *stacked capacitance cell* (STC) and the *trench capacitance cell* shown in figure 6.8.

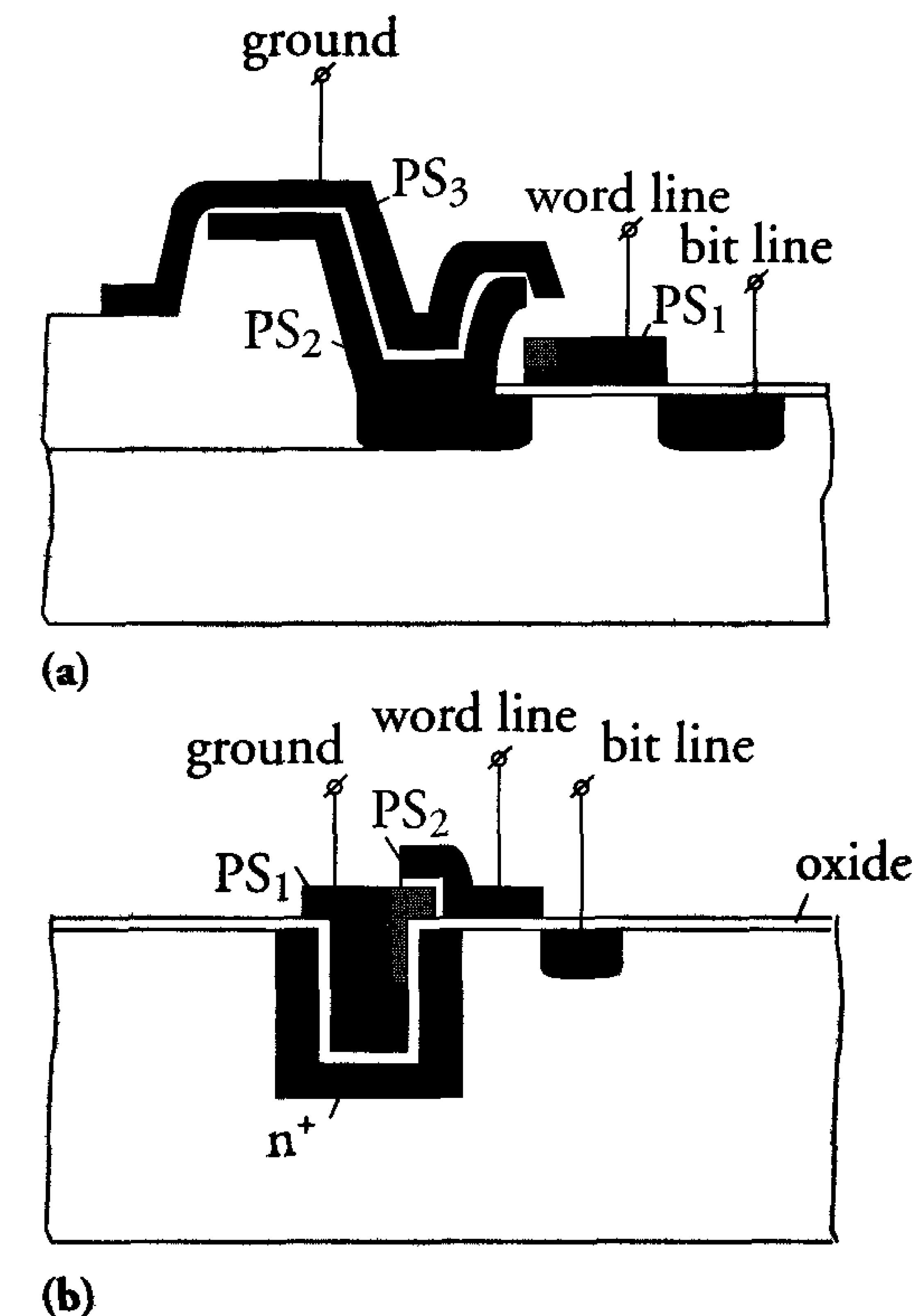


Figure 6.8: *(a) Stacked capacitance and (b) trench capacitance three-dimensional DRAM cells*

The manufacturing processes required for these cells are more complex than conventional processes. The STC cell has gained popularity and currently consists of a stack of metal and dielectric layers on top of the silicon. Another way of increasing DRAM cell capacitance is the use of high$-\epsilon$ dielectrics. Today, values of about 20 are used and much higher values are expected in the near future [16].
Despite associated processing and operational disadvantages, the DRAM has achieved a dominant market position. This is mainly because of the relatively low area per bit, which is generally 6 to 10 times lower than those of SRAMs. This leads to cost advantages of a factor of three to four.

General remarks on DRAM architectures

There are important differences between the basic DRAM and SRAM operation. Both SRAMs and DRAMs have similar secondary and sometimes even tertiary amplifiers in the I/O path.

The access time of a DRAM is approximately two to four times longer than that of an SRAM. This is mainly because most SRAMs are designed for speed, while DRAM designers concentrate on cost reduction.

DRAMs are generally produced in high volumes. Minimising the pin count of DRAMs by row and column address multiplexing makes DRAM operation slower but cheaper as a result of the smaller chip size. Because of the optimisation of DRAM cells for a small area, the higher DRAM processing costs can be regained by the larger number of dies on the wafer. The small cell area leads to significant charge leakage at high temperatures, which in turn necessitates a higher refresh frequency.

In addition to minimising cell area, other techniques are also used to reduce the total area of memories. One such technique reduces the number of bond pads on stand-alone DRAMs by multiplexing the row and column addresses through the same bond pads. This technique is illustrated in figure 6.9.

The $\overline{RAS}$ signal is used latch the row address and to start the memory cycle. It is therefore required at the beginning of every operation. Signal $\overline{RAS}$ is active low. The high period of the $\overline{RAS}$ signal is called the $\overline{RAS}$ precharge time. The $\overline{RAS}$ signal can also be used to trigger a refresh cycle. This is also called $\overline{RAS}$ Only Refresh (ROR). The $\overline{CAS}$ signal, which is also active low, is used to latch the column address and to start the read and write operation. It can also be used to initiate a $\overline{CAS}$ before $\overline{RAS}$ refresh cycle. There are minimum amounts of time specified for the active and inactive periods for both $\overline{RAS}$ and $\overline{CAS}$ signals. The period that $\overline{CAS}$ must be inactive is called the $\overline{CAS}$ precharge time. The $\overline{CAS}$ signal is not required to be active during a ROR refresh cycle.

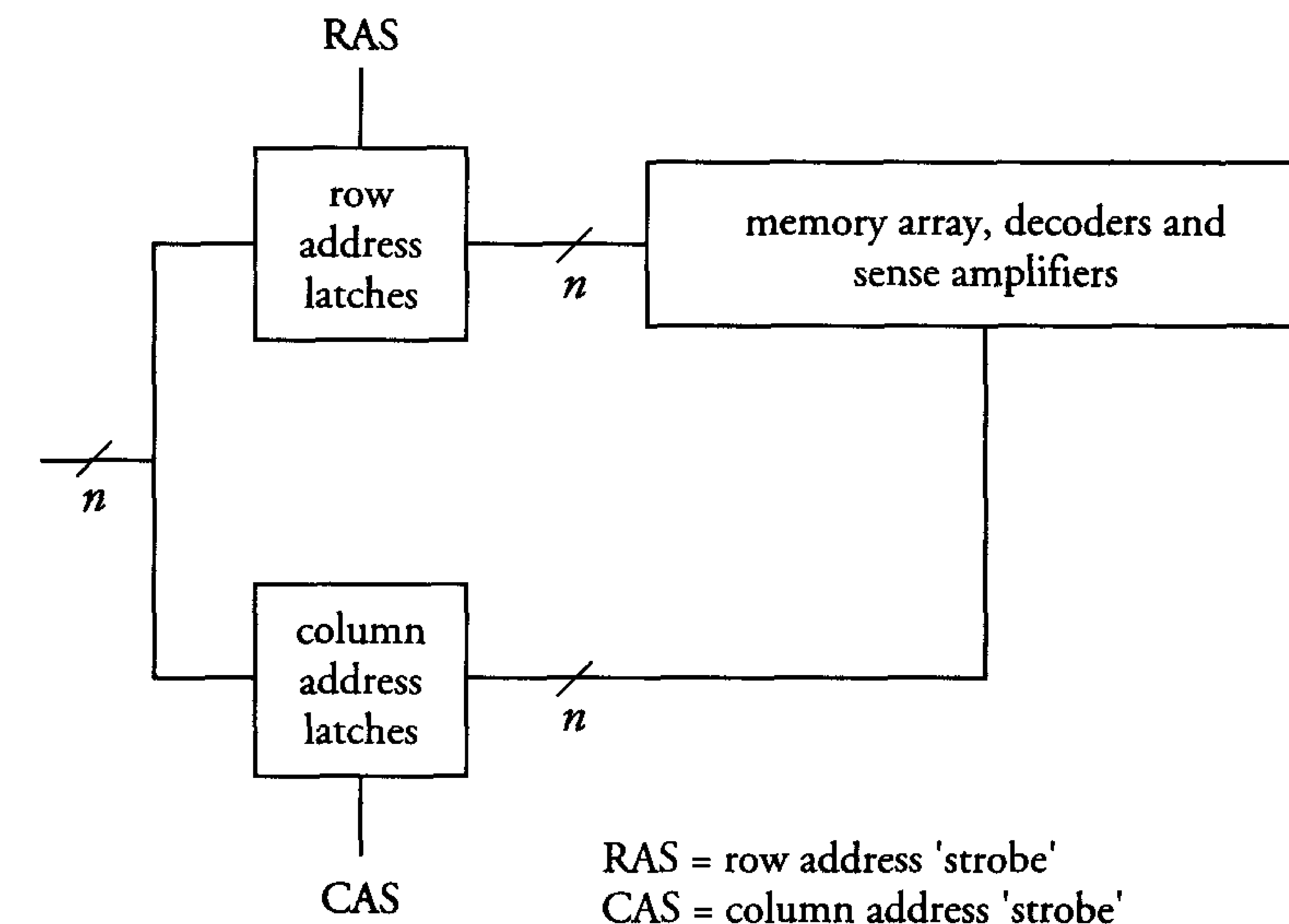


Figure 6.9: *Row and column address multiplexing in DRAMs*

Stand-alone SRAMs use separate bonding pads for the row and column addresses to achieve fast access times. The access time of a stand-alone SRAM is therefore considerably shorter than that of an equivalent stand-alone DRAM. This is illustrated in figure 6.10(a). This figure compares the access times of a stand-alone SRAM and a stand-alone DRAM, which uses row and column address multiplexing. The access time of the SRAM is only determined by the time interval t_1 whereas the total access time of the DRAM is determined by the sum of several set-up, hold and delay times. The improved DRAM access time in figure 6.10(b) is achieved by omitting the column address latches and implementing a *static column access.*

The data rate of a RAM is determined by the cycle time. This has already been defined as the minimum possible time between two successive accesses to a memory. The cycle time of an SRAM can be equal to its access time. In a DRAM, however, the cycle time is the sum of the access time, the precharge time of the bit lines and the refresh time. This holds for full random access. In page mode (or EDO), precharge and refresh times do not add to the (page mode) cycle time. Therefore,

page mode cycle times are about two to three times shorter than full random-access cycle times.

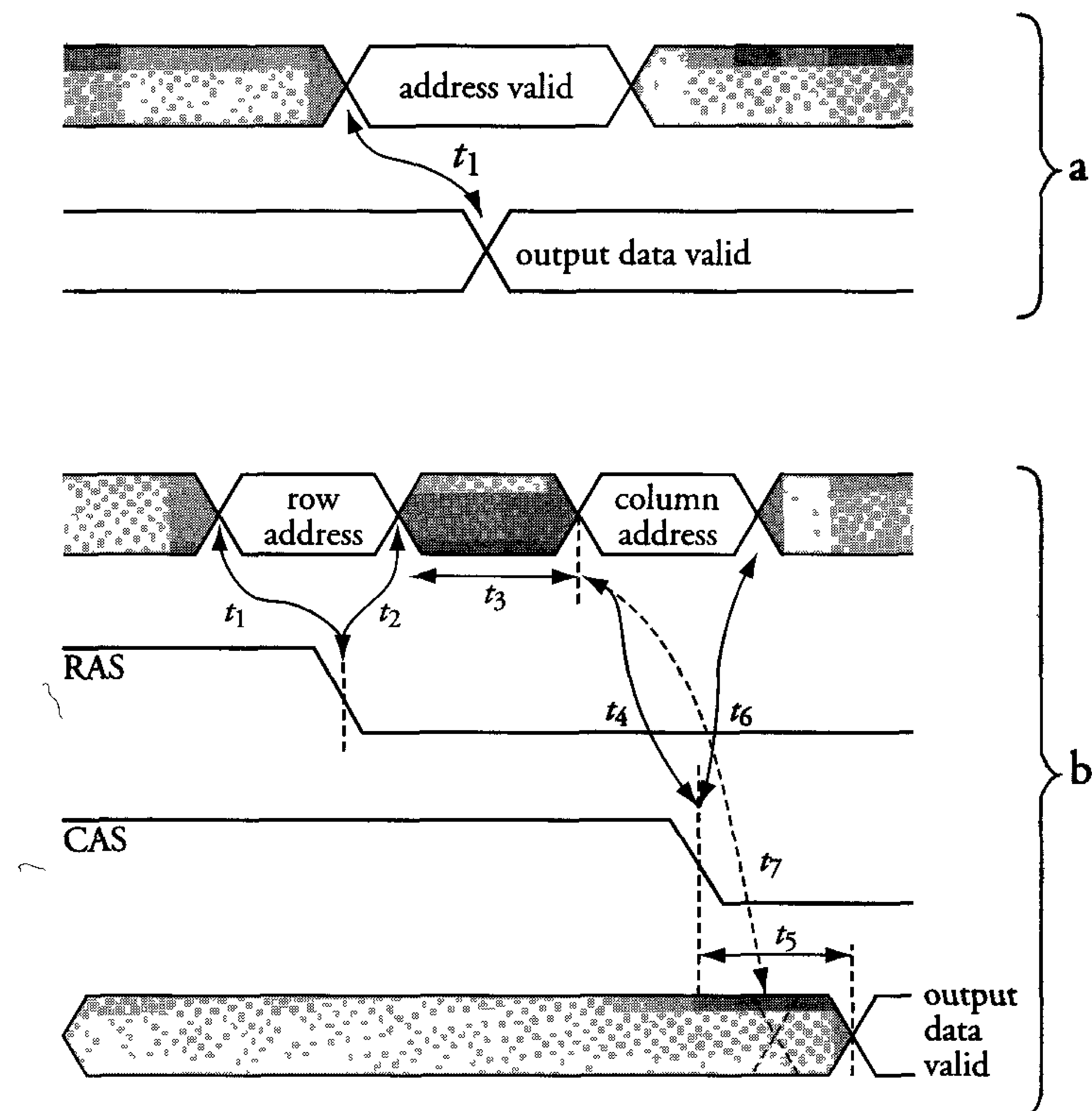


Figure 6.10: *Access times of (a) an SRAM: access time=t_1 and (b) a DRAM: access time=$t_1 + t_2 + t_3 + t_4 + t_5$ or improved access time=$t_1 + t_2 + t_3 + t_7$*

6.3.4 High-performance DRAMs

The increased overall performance of systems, which used DRAMs for storage, required the DRAM performance to increase at the same pace. Several solutions have been developed to improve DRAM performance during reading. These relatively new generation DRAMs include *Fast Page Mode (FPM)* , *Extended Data Out (EDO)* Mode, burst data using *synchronous DRAMs (SDRAM)* and *Rambus DRAM (RDRAM)*.

All four approaches are based on the ability to access complete pages without requiring the start of a new memory cycle. A *page*, which represents all the memory cells that share a common row address, can have a length of as many as several kbits. The total number of words in a page equals the total number of column addresses on a row. The drawback of page mode is the segmentation of the data, increasing the chance that the required data will not be on the accessed page. Particularly graphics applications benefit most from page mode access.

Another advantage of page mode architectures is their lower power consumption, because there are no sense and refresh currents during page mode access. Most DRAMs are asynchronous; these include conventional DRAMs, FPM and EDO RAMs. A memory operation is initiated on the arrival of input signals.

The differences between a synchronous and an asynchronous DRAM involve more than just the presence or absence of a clock signal. With SDRAMs, for instance, a precharge cycle is independent of a *RAS*, to allow multiple access on the same row. Internally, a refresh cycle is identical to a read cycle. No column addresses are needed during refresh, since no output data is required. FPM DRAM, EDO DRAM, SDRAM and RDRAM are all based on the same core memory. Therefore, their internal timing diagrams look very similar. The differences are mainly determined by how they communicate with the outside world. These differences include the speed at which address and control signals can propagate through the DRAM and the speed at which data propagates from the DRAM to the memory controller [7]. In the following, a brief overview of the different high-speed DRAM architectures is presented.

Fast Page Mode DRAM

In an FPM DRAM, the column address set-up starts as soon as the column address is valid, so that the column address can be latched at the falling edge of *CAS*. This is different from conventional page modes in which a column address access was initiated by the falling edge of the *CAS* signal. It was, therefore, required to wait with the column address set-up until the falling edge of *CAS*. In this way, a reduced page cycle can be achieved in comparison to conventional page mode DRAMs.

Extended Data Out DRAM

The EDO DRAM architecture looks very similar to the FPM DRAM. However, it contains an additional register that holds the output data. This allows the data from the current read cycle to be present at the

outputs beyond the start of the next read cycle. The possibility to "overlap" output data with input data of a next cycle results in a 30% speed improvement over comparable page mode DRAMs. Current EDO DRAMs contain a single bank architecture and must therefore process memory operations serially. A memory operation cannot start before the previous one is completed.

Synchronous DRAMs

When the transfer of address, data and control signals to a DRAM is synchronised by the system clock, such a DRAM is called a synchronous DRAM. Both SDRAMs and RDRAMs have such synchronous architectures and interfaces. Different synchronous DRAM architectures are presented here.

1. SDRAM architectures
   In an SDRAM, in addition to a given external starting address, the next column addresses during a burst are generated by an on-chip counter, while an asynchronous DRAM requires the memory controller to generate a new column address for each access. SDRAMs and RDRAMs are generally built with multiple memory banks (two, four...). Each bank is a memory of its own [8], allowing individual and parallel operation for maximum performance. SDRAM architectures use burst features to accommodate fast external transfer at increasing burst rates. Synchronous DRAMs (SDRAM, SGRAM and RDRAM) use the system clock as their clock input. Therefore, they are targeted at matching (or half) clock speeds of commonly-used PC systems (100 MHz, 133 MHz, 166 MHz, ...etc). At a burst rate of 100 MHz, the individual bits of the burst length are supplied at 10 ns intervals. Many SDRAMs can also operate in a random-access mode, in which they show similar timing as FPM or EDO DRAMs. SDRAMs may have 64-bit or even 128-bit wide I/O formats. Besides commodity DRAM applications, this allows them to also serve in applications with extremely high memory bandwidths. For this purpose, an SDRAM architecture includes: burst feature, more than one memory bank for parallel memory operation and a clocked or synchronous interface. Particularly graphics applications (which are characterised by high-speed and wide I/O buses) require extremely high bandwidths. Video RAMs (VRAMs) and Synchronous Graphics RAMs (SGRAMs) are specially designed for graphics applications.

2. Video RAM architectures
   As the pixels on a computer terminal or a TV are refreshed serially, the first Video RAMs provided continuous streams of serial data for refreshing the video screen. The standard DRAM had to be extended with a small serial access memory and a serial I/O port to support the storage of video pictures [9]. However, all VRAMs still have the original standard random-access DRAM port also available. During a serial read, the DRAM array is accessible via the DRAM port for a separate read or write operation. Special features, such as block write and flash write, etc. are supported by additional circuits. However, the rapid rise of special SDRAM architectures, such as SGRAMs will cause the VRAMs to fade in the near future.

3. SGRAM architectures
   SGRAM architectures are very similar to those of a VRAM. They contain similar additional hardware, such as registers and mask registers to support block write and write-per-bit functions. This results in faster and more efficient read and write operations. These features are supported by special registers and control pins. Colour registers are mainly used to store the colour data associated with large areas of a single colour, such as a filled polygon [9]. The data in these colour registers can be written in consecutive column locations during block-write operation. Write-per-bit allows the masking of certain inputs during write operations; it determines which memory locations are written.

   SGRAMs lag by about a factor two in memory capacity behind commodity DRAMs. 64-bit wide SGRAMs are being developed. A major difference with a VRAM is the additional synchronous interface of the SGRAM. Current SGRAMs are available at clock speeds exceeding 100 MHz. A trend in increasing the memory's bandwidth is the use of *Double Data Rate (DDR)* I/Os, which are already available in SDRAM designs. In the DDR mode, both the falling and rising edges of the clock are used to double the data throughput. This feature can push the SGRAM's graphics peak bandwidth up to several Gbytes/s. The popularity of SGRAMs has increased such that it is currently used in about 70% of all graphics systems. Another DRAM version, called the Rambus^{TM} DRAM (RDRAM), is gaining popularity as well, particularly in graphics applications.

4. RDRAM architectures
   The RDRAM (particularly the *Direct RDRAM*) provides high bandwidth for fast data transfer between the memory and the programming parts in a system. The Rambus^{TM} interface is licensed to many DRAM manufacturers and, at certain engineering fees, they can get customised interfaces to their existing products. Because of the high bus clock rates and the use of DDR, RDRAMs claim extremely high bandwidths, competing with that of SDRAMs and require fewer data lines than the wide-word DRAM. The Direct RDRAM has only little overhead on a standard DRAM architecture and offers several modes from power-down (only self-refresh) to selective powered-down memory blocks [10].

   An alternative to the Direct RDRAM is the Concurrent RDRAM, which can operate two memory banks simultaneously. By using interleaved data, it offers about the same bandwidth. A similar approach, which also offers a high bandwidth, is SYNCLINK, which is supported by JEDEC. There are several other memory types which offer high to extremely high bandwidths. This offers system designers a wide choice in creating the optimum solution for their particular application.

Currently, DRAMs have reached the gigabit level, with laboratory versions available of 1 Gbit and 4 Gbit in development. As the application area increases, the hunger for increased densities and higher speeds will drive the complexity and performance of SDRAMs and DRAMs to incredibly high levels.

6.3.5 Error sensitivity

The logic value in a RAM cell may change as a result of *α-particle radiation*. These α-particles may come from impurities in the metal layer (e.g. aluminium) or from the package. They may even arrive from outside the chip. Their effect is to generate a relatively large number of electron-hole pairs, which may randomly discharge memory cells. This random loss of stored information occurs in both DRAM and SRAM cells. DRAMs are particularly prone to the resulting *'soft errors'*, which become more influential as densities increase and stored charges decrease. SRAMs based on CMOS technology have low power consumption and reduced susceptibility to α-particles.

Memories can also be covered with polymide to protect them against external α-particle radiation. This reduces soft-error rates by a factor of 1000 or more. This does, of course, not apply to α-particles from metal or silicon of the die. This is one of the reasons why the cell charge is not decreased every new DRAM generation.

6.3.6 Redundancy

Stand-alone memories are sold in very high volumes and must therefore be very cheap to produce. Methods to achieve a low price include yield-improvement techniques which may, for example, result in a yield in excess of 70 % for areas greater than 100 mm^2. However, many stand-alone memories have one or more cells that do not function properly. For this reason, most stand-alone memories include several redundant memory rows and/or columns which can be used to replace defective cells. The faulty cells are detected by means of memory testers and a laser beam is used to isolate their corresponding rows or columns. This so-called *laser-fusing* technique is also used to engage spare rows and columns. This technique may be used to improve the yield by a factor of as much as 20 to 50 during the start-up phase of current new process generations.

6.4 Non-volatile memories

6.4.1 Introduction

As discussed in section 6.1, a non-volatile memory keeps its stored data when the memory is disconnected from the supply. Non-volatile memories include ferroelectric RAM, ROM, PROM, EPROM, EEPROM and flash (E)EPROM. In the following paragraphs, their basic operation is discussed in some detail, including their fundamental properties.

6.4.2 Ferroelectric RAM (FRAM)

Ferroelectric RAM technology has "almost been available" for quite some time, mainly as a result of too optimistic expectations. The basic concepts of *FRAM* operation have been known since the fifties. However, with the focus on the costs and the quality of silicon memories, hardly any progress in FRAM technology has been made since then.

The first FRAM realised on silicon was unveiled in 1988 [11]. It contained 256 bits, which were built up from a six-transistor, two-capacitor array per cell. Compared to DRAM technology, this FRAM consumed a lot of area. Using a two-transistor, two-capacitor cell in 1992, current densities up to 4 Mbit are in development, with many different standard interfaces, in submicron technologies with one-transistor, one capacitor per bit. This basic cell looks very similar to a basic DRAM memory cell, see figure 6.11.

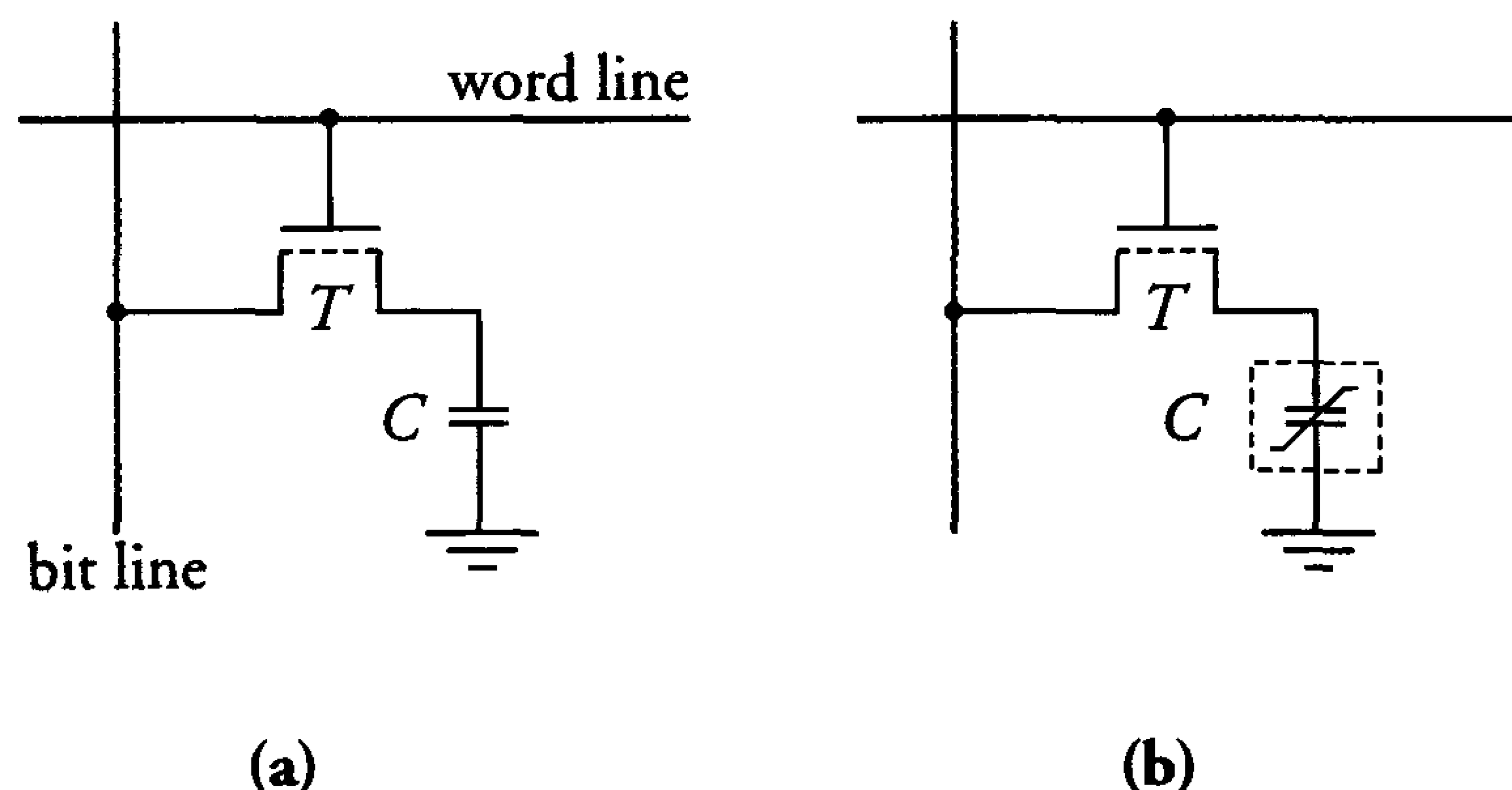


Figure 6.11: *(a) Basic DRAM cell, (b) basic 1T, 1C FRAM cell*

The operation of a DRAM cell is discussed in section 6.3.3. The operation of an FRAM cell is based on the polarisation state of its ferroelectric capacitor. The dielectric material used in this capacitor belongs to a certain class of dipole materials, which are called Perovskite crystals. By applying an electric field across this dielectric, these crystals polarise. This polarised state is maintained after the electric field is eliminated. The dielectric is depolarised when an electric field of the opposite direction is applied.

During a read operation, an electric field is applied across the capacitor. Similar to a DRAM, the current is detected by a sense amplifier. When the dipoles switch state, the sense current is higher. Again similar to a DRAM, the data in a FRAM cell is destroyed during reading (*Destructive Read-Out (DRO)*). The cell contents must therefore be rewritten (refreshed) immediately after each read operation. A complete read cycle includes a precharge period, a read operation and a rewrite operation. Because of higher dielectric constants, an FRAM's cell charge density is higher than that of DRAM cells, allowing smaller cell sizes.

Advances in FRAM technology have resulted in trench capacitor and stacked capacitor architectures, analogous to DRAM technology evolution. Currently, a dozen manufacturers are offering or developing FRAMs. Basically, an FRAM operation depends on voltages rather than currents. This makes FRAMS particularly suited for low power applications. FRAMs are therefore considered as the ideal memory for emerging applications such as smart cards and RF identification [11]. Potential applications include digital cellular phones and Personal Digital Assistants (PDAs). Compared to EEPROM and flash memories, the number of read/write operations (endurance cycle) for FRAMs is several orders of magnitude higher (up to 10^{10} to 10^{12})

6.4.3 Read-Only Memories (ROM)

A ROM is in fact a random-access memory which is written during the manufacturing process. The information is therefore lasting and non-volatile. It can be read but it can never be altered. With the exception of the write facility, the architecture of a ROM is similar to that of a RAM. Subsequent discussions are therefore restricted to the different techniques for writing the information during the manufacturing process. The ROM memory cells required by each technique are examined separately.

Different processing layers could, in principle, be used to store information in a ROM. Two popular choices are the diffusion and contact layers. ROM cells and structures based on the corresponding ACTIVE and CONTACT masks are discussed below.

ROM cell with the information in the ACTIVE mask

Figure 6.12 shows the structure of a ROM which is programmed by means of the ACTIVE mask, see section 4.6. The ROM cell is enclosed by a dashed line in the figure. An example of the layout of such a cell is given in figure 6.13.

All bit lines in the ROM in figure 6.12 are precharged when ϕ is 'low'. The V_{S_1} line is switched to ground when ϕ goes 'high'. The cell enclosed by a dashed line is read when the corresponding word line WL_3 goes 'high'. Bit line bl_2 will then be discharged if ACTIVE is present in the cell. Otherwise, bl_2 will remain charged. The information in the ROM is therefore stored in the ACTIVE mask, corresponding to the presence or absence of a memory transistor at the selected cell position.

Figure 6.14 shows a photograph of a ROM array based on the cell of figure 6.13.

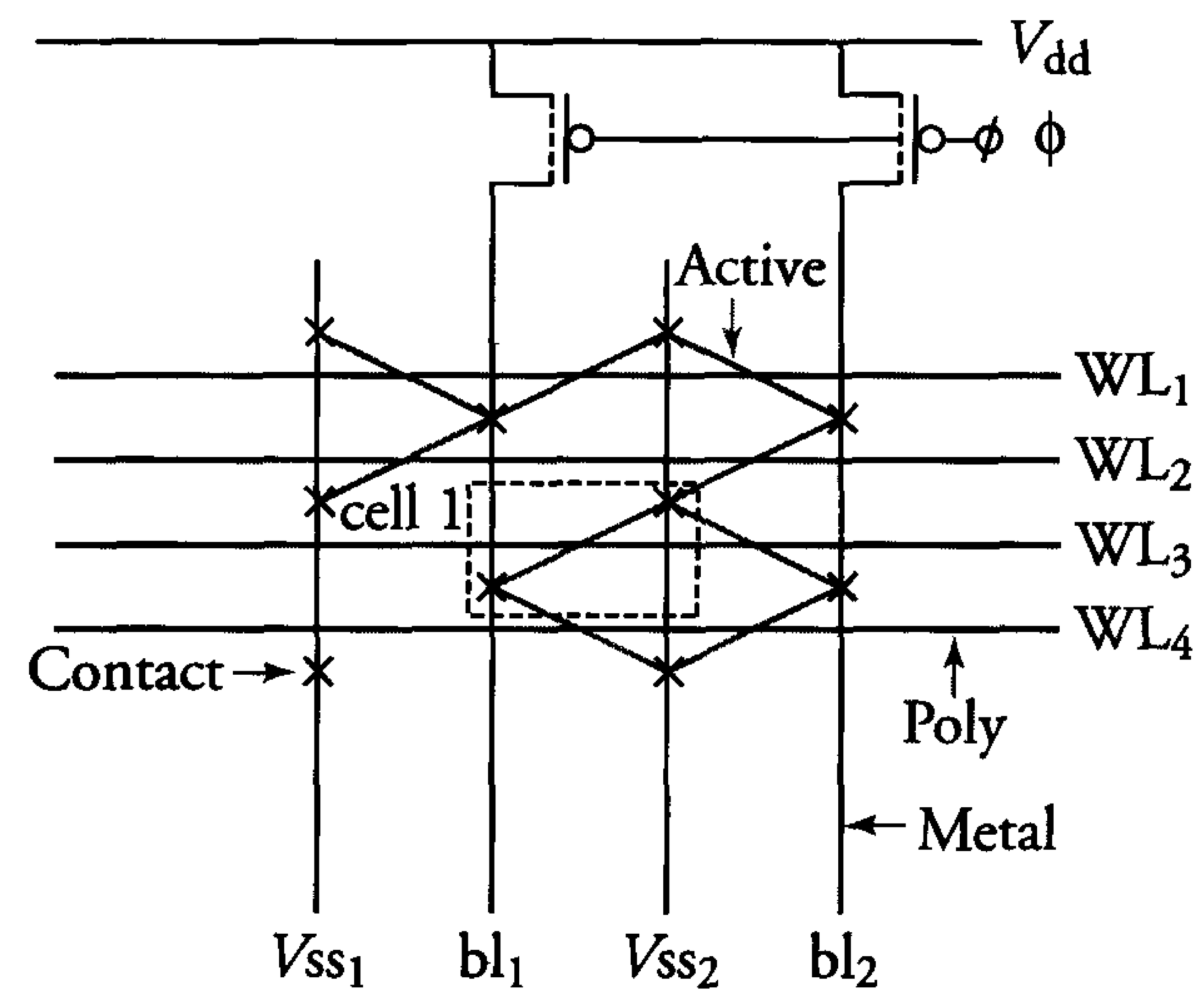


Figure 6.12: *ROM with information in the ACTIVE mask*

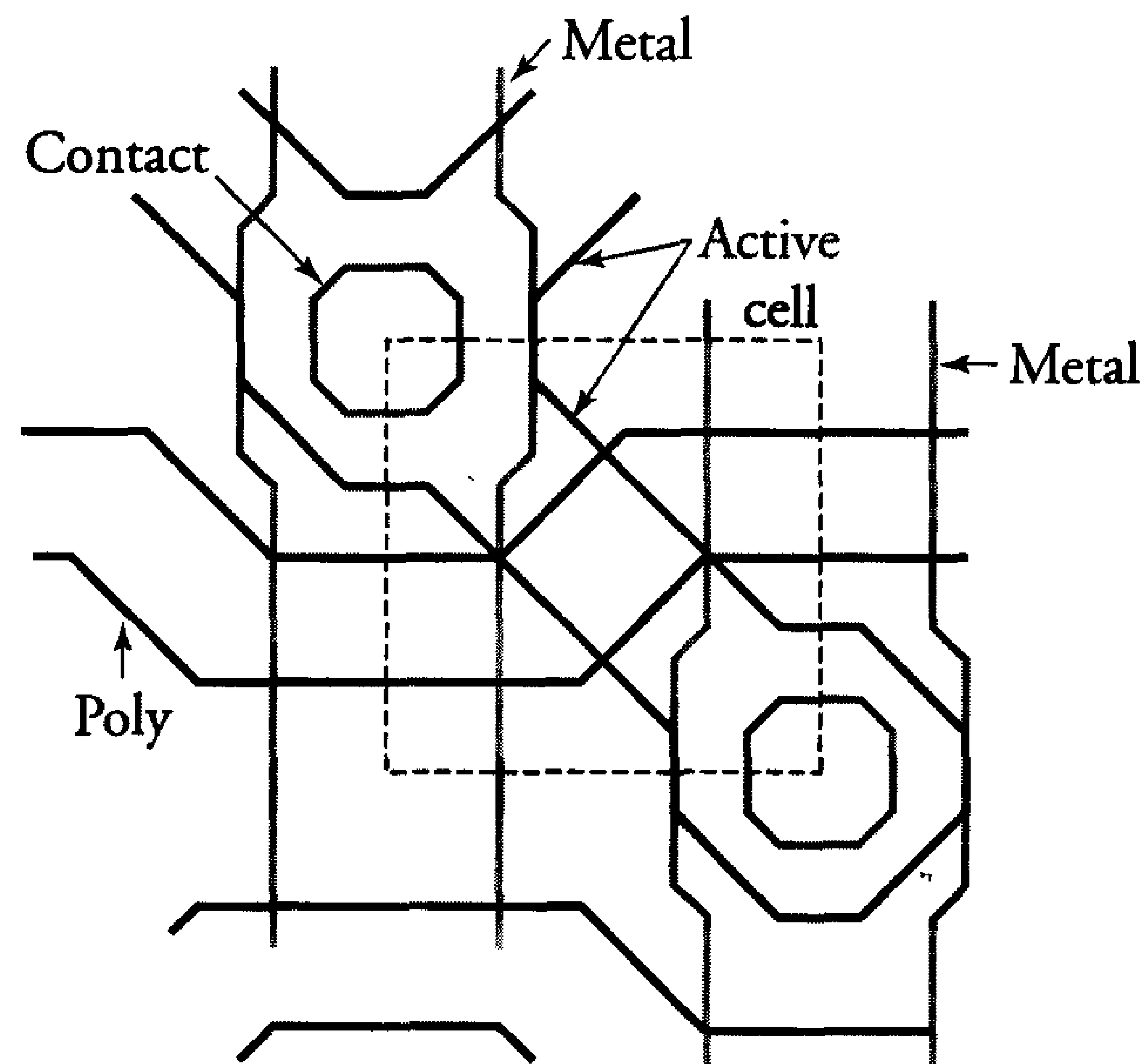


Figure 6.13: *Layout of an ACTIVE-mask programmed ROM memory cell*

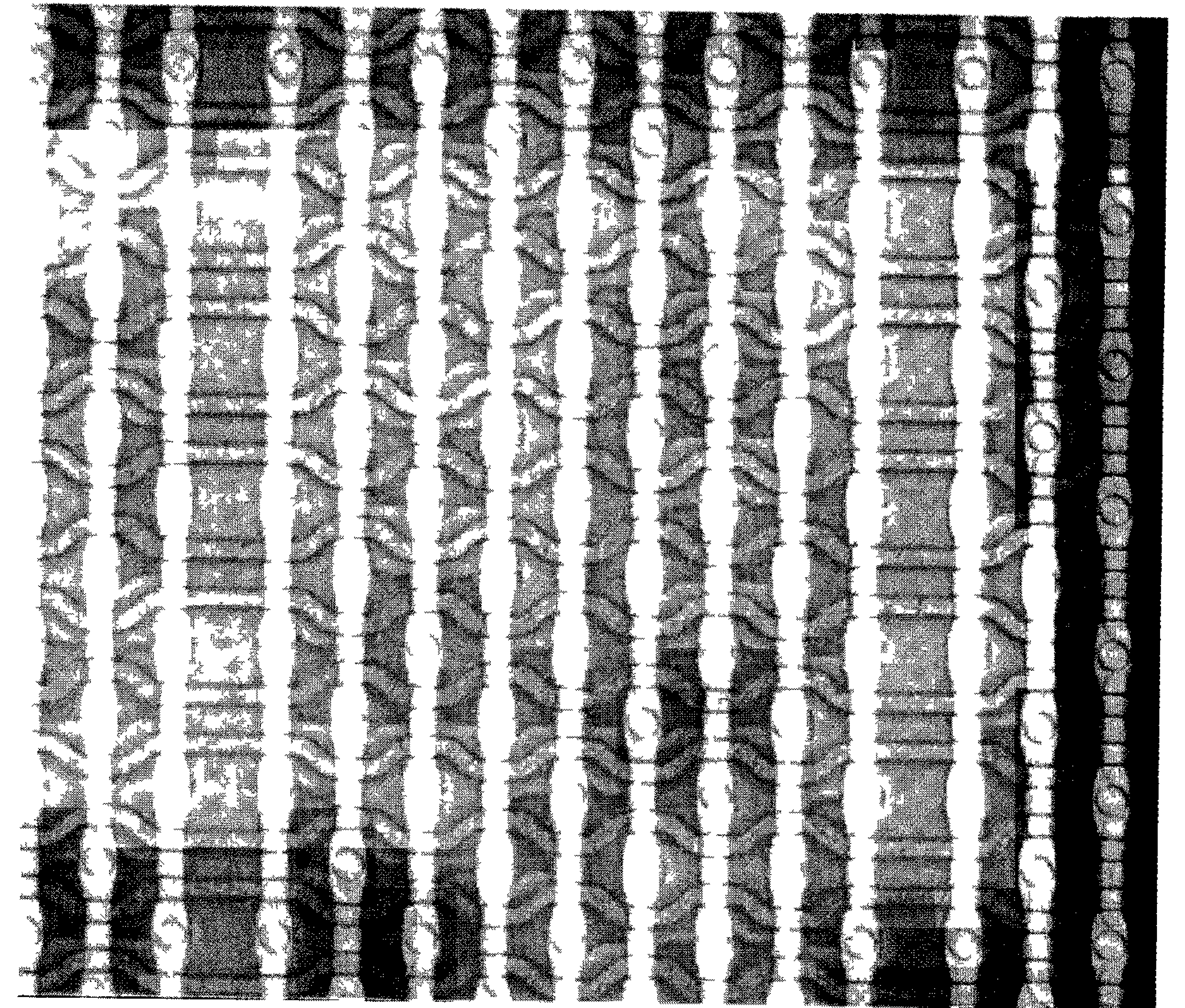


Figure 6.14: *Photograph of an array of ROM cells (photo: PHILIPS)*

ROM cell with the information in the CONTACT mask

Figure 6.15 shows the structure of a ROM which is programmed by means of the CONTACT mask. All bit lines in this ROM are precharged through the pMOS transistor when ϕ is 'low'. A word line is activated when ϕ goes 'high'. The bit lines of cells connected to the selected word line and containing a CONTACT hole are then discharged. The CONTACT hole in the cell locally connects the aluminium (METAL) bit line to the drain of a transistor, which has its source connected to a grounded diffusion (ACTIVE) track. The series resistance of the ACTIVE tracks is reduced by means of an extra aluminium ground line which is implemented every 8 to 10 bit lines.

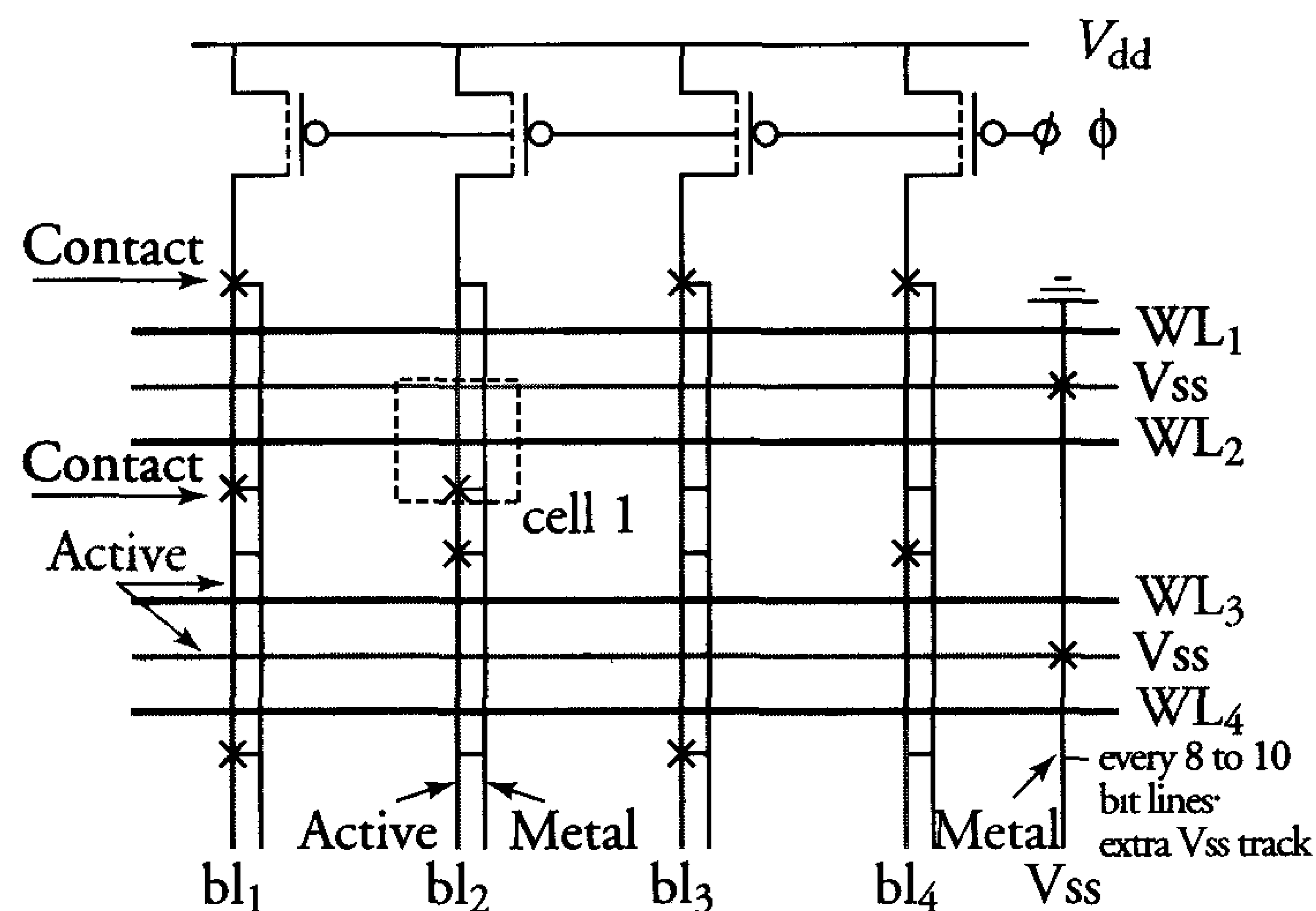


Figure 6.15: *ROM with information in the CONTACT mask*

Comparison of the ACTIVE-mask and CONTACT-mask programmed ROM cells

A fair comparison of the chip area of the ACTIVE-mask and CONTACT-mask programmed ROM memory cells requires the inclusion of a suitable fraction for the area of the extra ground line in the latter cell. This gives the following values for a 0.25 μm CMOS process:

ROM cell in figure 6.12: $1\,\mu m^2 \leftrightarrow$ ROM cell in figure 6.15: $1.5\,\mu m^2$

Although the second cell is the larger of the two, it has the advantage that its information is programmed in the CONTACT mask. This mask is used in one of the last steps in the manufacturing process. Therefore, ROMs which store information in the CONTACT mask can be largely prefabricated. Now, only a small number of manufacturing steps are required to realise a ROM with specific contents. In contrast, the ACTIVE mask is usually one of the first in the manufacturing process. The *turn-around time* between order and delivery is therefore much shorter for a ROM with its information in the CONTACT mask than for a ROM with information in the ACTIVE mask. Therefore, in multi-metal layer processes, the programming is increasingly done in the latest VIA mask.

There are some other types of ROM as well. In a *serial ROM*, a NAND type of structure is used to discharge the bit line. In such a ROM, a V_T-implant is used for program storage (enhancement or depletion type of memory transistor). The series connection of the cells allows a much smaller number of contacts. This results in a small area, but also in a relatively low speed of operation.

6.4.4 Programmable Read-Only Memories

Introduction

The three different types of programmable Read-Only Memory are PROM, EPROM and EEPROM. Respectively, these ROMs are *programmable*, *electrically-programmable* and *electrically-erasable programmable*. They are programmed by users rather than during manufacturing. Although they are programmed by users, these memories are still called read-only memories because the number of programming/erasing cycles is rather limited in normal usage.

PROMs (Programmable Read-Only Memories)

A *PROM* is a read-only memory which can be programmed only once by the user. Each cell contains a fuse link which is electrically blown when the PROM is programmed. PROMs are usually manufactured in a bipolar technology. The fuses are then implemented in a nickel-chromium (NiCR) alloy. The resulting cell is relatively large and is about four times the size of a ROM cell. The wish for rewritability has decreased the PROM's market share to almost zero in favour of erasable architectures. These are discussed in the following sections.

EPROMs

Figure 6.16(a) shows a schematic representation of an *EPROM* memory cell.

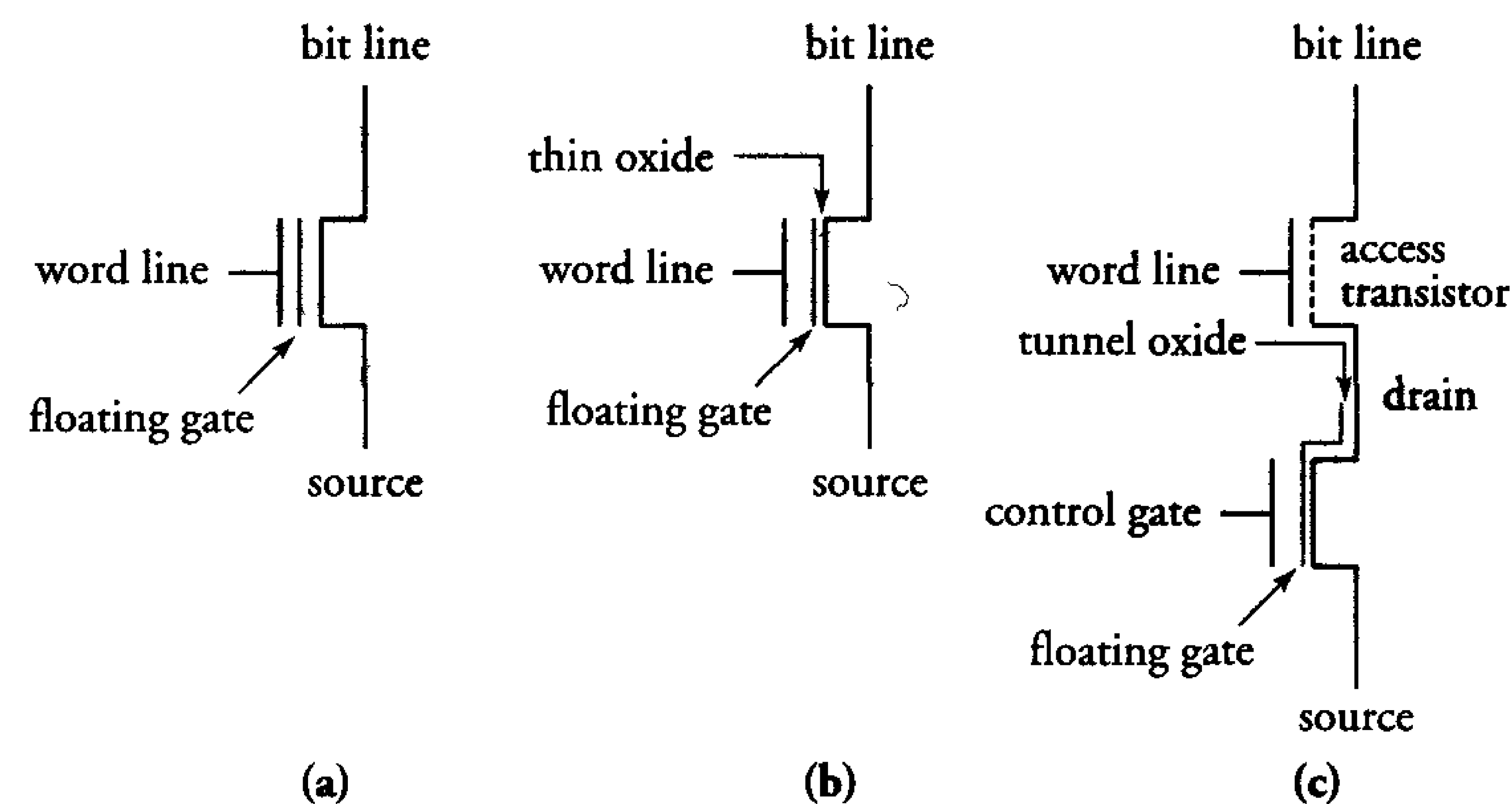


Figure 6.16: *Schematic representation of (a) an EPROM cell, (b) a flash (E)EPROM cell and (c) a full-featured EEPROM cell*

The data in this cell, as in an EEPROM cell, is represented by the presence or absence of charge on the *'floating gate'* of the memory transistor. The floating gate is charged by means of a large electric field between the transistor's source and drain. This accelerates electrons in the channel to very high energy levels. Some of the resulting *'hot electrons'* (see section 2.7) penetrate through the gate oxide to the floating gate. Sufficient charge is collected on the floating gate when high drain-to-source voltages of over 4 V (in a $0.25\,\mu$m process) and gate-source voltages of about 8 V are applied. This causes currents of the order of 0.5 mA in the cell. The number of programming/erasing cycles in an EPROM is limited (100 to 1000 cycles). The second power supply required for in-system programming of an EPROM, because of the high current requirement, is usually 12.5 V, which will probably be reduced in the near future. Alternatively, an EPROM can be removed from a system and programmed in a special PROM programmer.

EPROMs are erased by exposing the cells to ultraviolet (UV) light. This is done through the transparent windows in EPROM packages. In many applications, EPROMs are only programmed once. They are therefore also available as *one-time-programmable* (OTP) devices in cheap standard plastic DIL packages with no transparent windows.

6.4.5 EEPROMs and flash memories

Floating-gate PROM structures, which allow electrical erasing and programming, were developed at the end of the seventies. The two resulting categories are electrically-erasable PROM (EEPROM) and flash memories.

EEPROM

Unlike with flash memory, EEPROM data can be changed on a bit-by-bit basis. This is also called a *full-featured EEPROM*, whose basic cell architecture is shown in figure 6.16. Because of the separate access transistor in the cell, EEPROMs feature relatively low bit densities compared to EPROM and flash memories. This transistor allows selective erasure of cells. Erasure is often done per byte.

Figure 6.17 shows a cross-section of the storage transistor of a full-featured EEPROM cell.

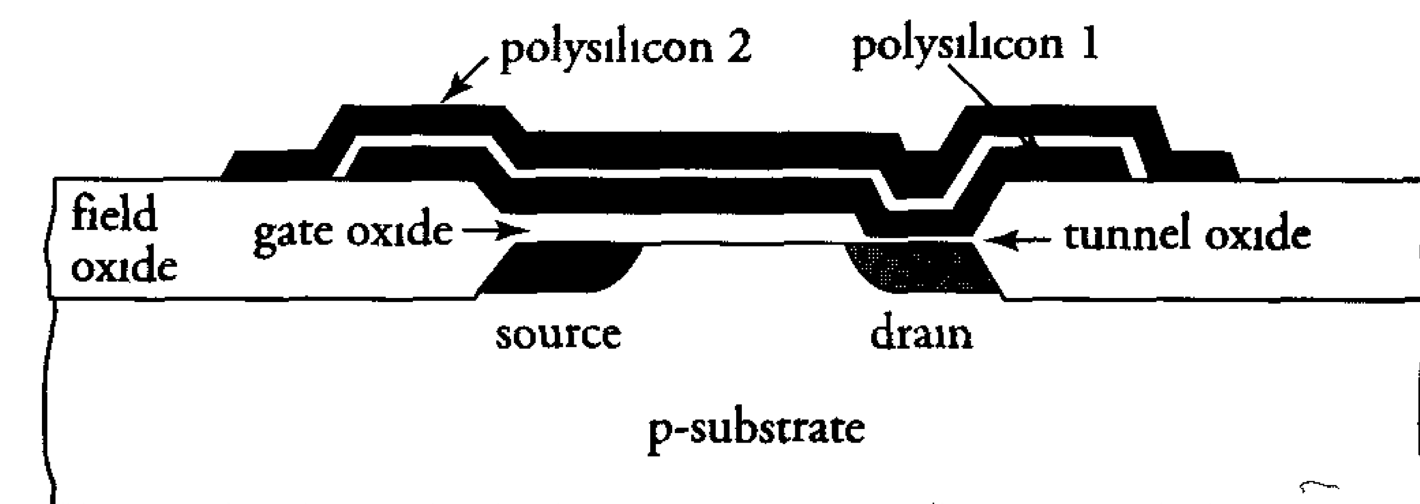


Figure 6.17: *Example of floating-gate EEPROM cell*

Data storage and erasure are achieved by moving electrons through a small thin oxide tunnel region between the floating gate and drain. This is done by applying a high electric field of about 10 MV/cm across the tunnel oxide, which induces so-called *Fowler-Nordheim (FN) tunnelling*. The cell is programmed when a voltage of about 12 to 15 V is applied between the gate and drain (substrate or source, depending on the technology). Now, electrons tunnel through the thin oxide and produce a positive charge on the floating gate. This decreases the threshold voltage of the memory transistor. The cell is erased by applying a reverse-biased voltage, which causes the electrons to flow to the floating gate. Therefore, the memory transistor in an erased cell has a high threshold voltage. The small currents involved in the tunnelling mechanism used

in full-featured EEPROMs facilitate on-chip generation of the 12 to 15 V for programming and erasing the memory.

An important characteristic of a full-featured EEPROM is the variation in memory transistor threshold voltage associated with successive programming/erasing cycles. Eventually, the threshold-voltage difference between a programmed and an erased cell becomes too small. This imposes a limit on the number of times that a cell can be erased and programmed. The plot of the threshold-voltage variation is called the *endurance characteristic*, see figure 6.18 for an example. The threshold difference enables more than 100,000 programming/erasing cycles for the individual cells.

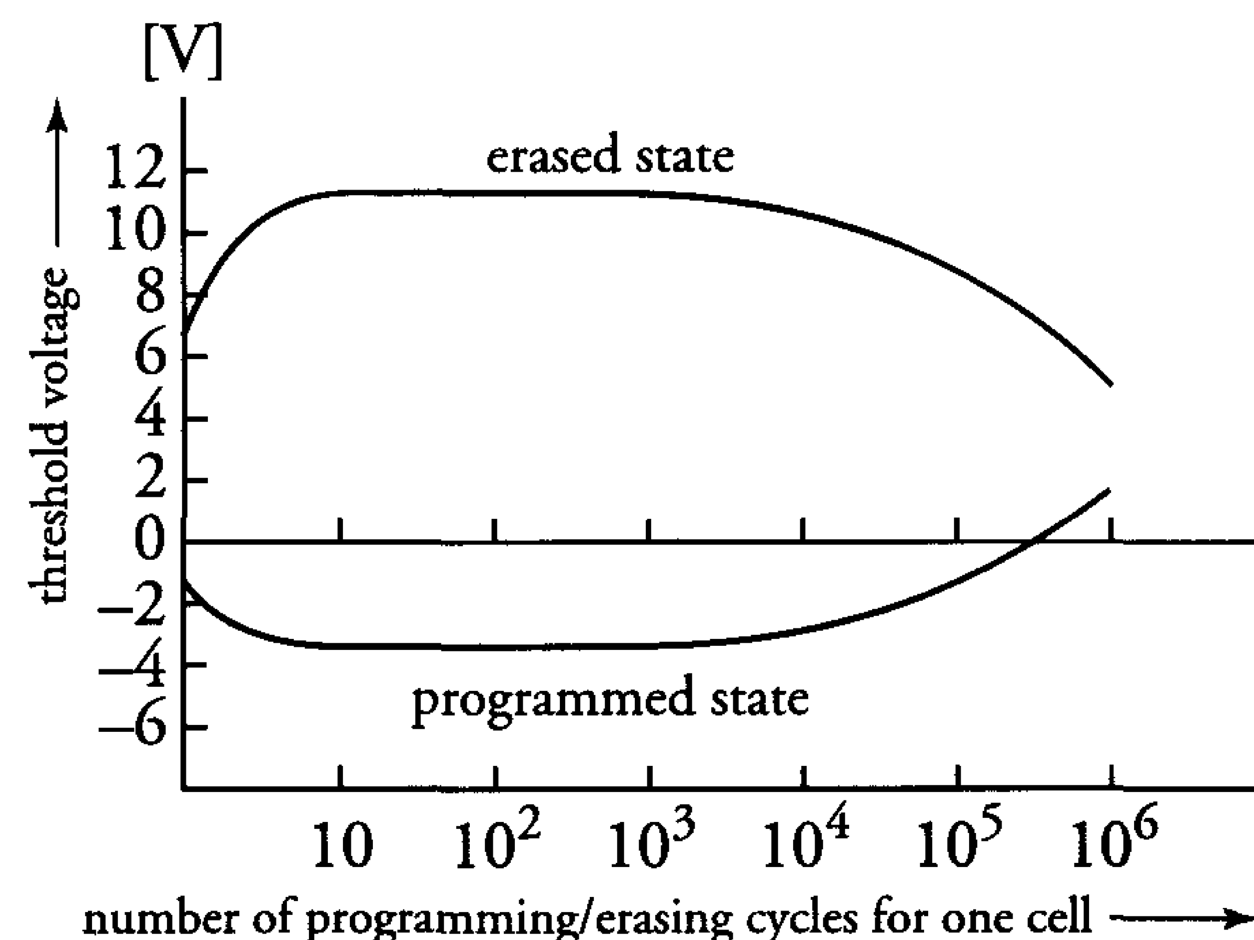


Figure 6.18: *Endurance characteristic of a full-featured EEPROM cell*

The *data retention time* of all EEPROMs is more than ten years. The various applications [12] of EEPROMs: conventional consumer applications, universal remote controls, cordless telephones, garage door openers, cameras, automotive, home audio and video and smart cards. Attention is also focused on cellular telephones and pagers. Innovative features have been added to EEPROMs, such as Block Lock which allows users to combine alterable data with secured data.

EEPROM technology is facing increased competition from flash memory, which allows much higher densities, as a result of the absence of the bit-by-bit change feature of an EEPROM, see figure 6.16.

Flash memory

A *flash memory* is in fact an EPROM or EEPROM in which the complete memory or complete sectors (blocks) of memory can be erased simultaneously. The two main techniques that can be distinguished are flash EPROM and flash EEPROM. Flash EPROMs are programmed using the hot-electron effect used in EPROMS. Relatively high electric fields and associated currents are needed for a reasonably efficient programming. All flash memories use FN tunnelling for erasure. Flash EEPROMs are both programmed and erased by means of this FN tunnelling. As this technique directly collects or removes charge to or from the floating gate, only low-current levels are involved. This results in low-power and high-efficiency operation and requires only relatively small on-chip charge pumps. Figure 6.19 shows a cross-section of a flash EEPROM storage cell, in which the cell is programmed by charge which tunnels through the gate oxide.

Figure 6.19: *Cross-section of a flash EEPROM*

Not all flash suppliers have adopted FN tunnelling for both programming and erasing. A large part of flash memories is used for code storage, which does not need high programming efficiency, but requires a reliable programming/erasing cycle [13].

All suppliers of flash memories for data- and file-storage applications specify more than 10^4 programming/erasing cycles. Depending on the technology and memory density, flash endurance can go as high as 10^6 cycles. Many flash memories are applied in high-volume consumer products, which require cost-effective, high-density solutions. Scaling of technologies, combined with the use of STI and stacked floating-gate structures, is one way to achieve Gbit densities.

Another approach is to store more bits of data in one single memory cell. In such a MultiLevel Cell (MLC), different amounts of electron charge on the floating gate may represent one of four possible combinations of two bits. During a read cycle, the control gate is set to a high level and the current through the cell is inversely proportional to the charge on the floating gate. Current sensing requires three differential

sense amplifiers, which each compare the cell current with that from one of three reference cells. The outputs of these sense amplifiers directly represent the stored cell data. Multilevel storage has been known for quite some time. However, reduced noise margins and increased design complexities created a lack of commercial interest until recently. The first *multilevel-storage* memory has been delivered since 1991. In a serial-EEPROM technology, analogue data samples were directly stored at a resolution of 256 levels in each cell, without the need for conversion of the analogue sample to binary words. The first commercial multilevel flash memory products were announced at the end of 1996. These first products stored two bits in one cell. Four bit cells are also in development.

The flash memory is penetrating many markets which were previously dominated by magnetic discs, ROMs, EPROMs an EEPROMs. Being able to double the density of flash memory would speed this process up even more.

6.4.6 Non-volatile RAM (NVRAM)

A *non-volatile RAM* combines SRAM and EEPROM technologies. This kind of memory is sometimes called a *shadow RAM.* Read and write actions can be performed at the speed of an SRAM during normal operation. However, the RAM contents are automatically copied to the EEPROM part when an on-chip circuit detects a dip in power. This operation is reversed when power returns. An NVRAM therefore combines the retention time of an EEPROM with the high performance of an SRAM.

6.4.7 BRAM (battery RAM)

A *BRAM* comprises an SRAM and a battery which provides sufficient power to retain the data when the memory is not accessed, i.e. when the memory is in the *stand-by mode.* The battery is used when power is absent. An SRAM is chosen because of its low stand-by power consumption. The battery is included in the BRAM package and the data retention time is close to 10 years.

6.5 Embedded memories

The integration of complete systems-on-a-chip (SOC) include the combination of logic circuits (logic cores) with memories (memory cores). There are several reasons to put memories and logic on the same chip. In many cases this is (and will be) done to:
- offer higher bandwidth
- reduce pincount
- reduce system size
- offer a more reliable system
- reduce system power

Also the low cost of interconnect at chip level may be a good reason to embed memories or other cores. The diagram [17] in figure 6.20 shows the relative cost of interconnect as a function of the distance from the center of the chip. It clearly shows that the chip level interconnect is by far the cheapest one.

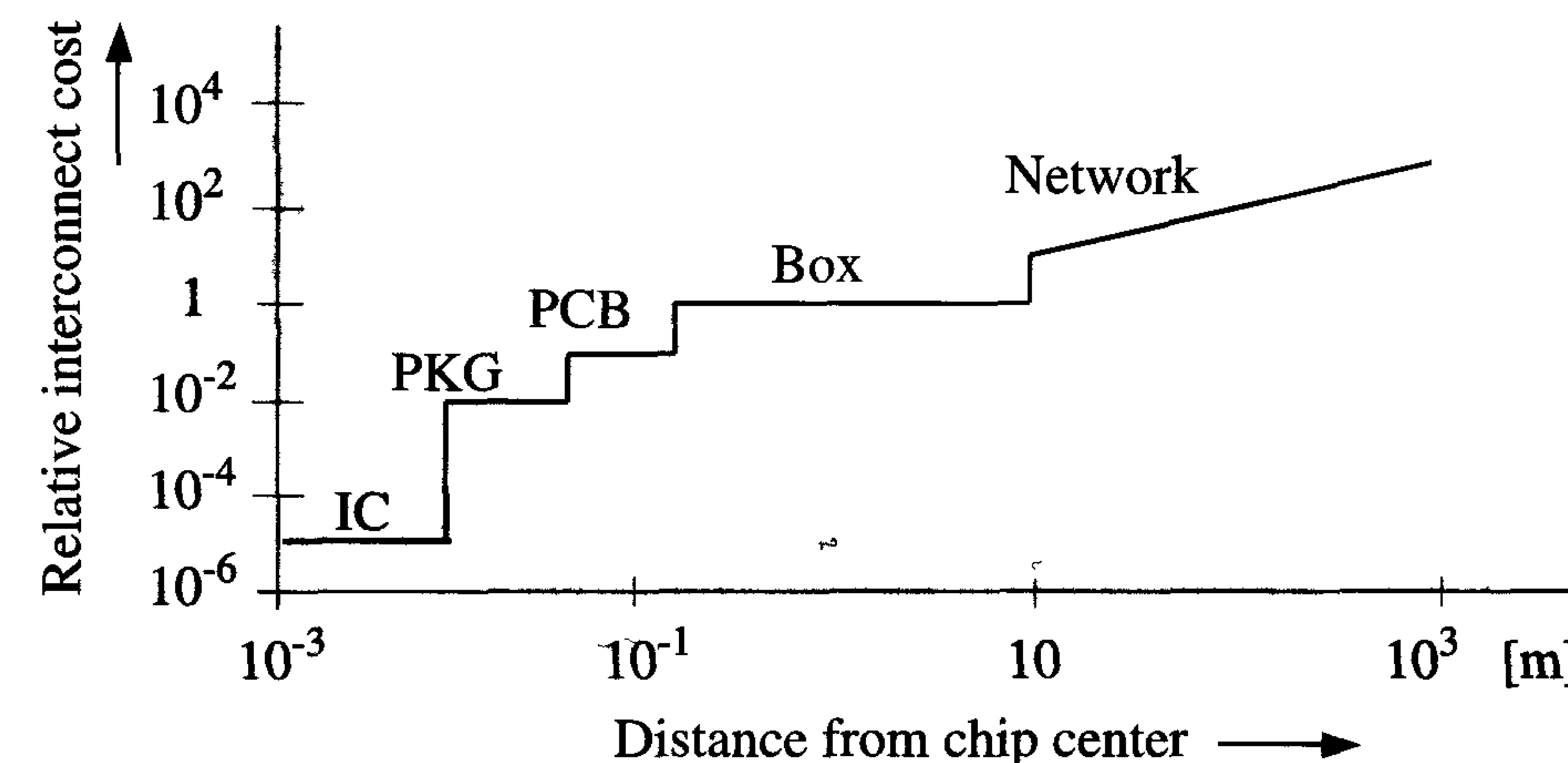


Figure 6.20: *Relative cost of interconnect*

Another reason to embed memories is to fill the design productivity gap. Figure 6.21 shows this gap with respect to the growth in IC complexity according to the ITRS roadmap [8]. The solid line represents the number of logic transistors per chip. The dotted line shows the design productivity. Many of the transistors made available by the technology, but unused by the design, may be used to increase the amount of embedded memory.

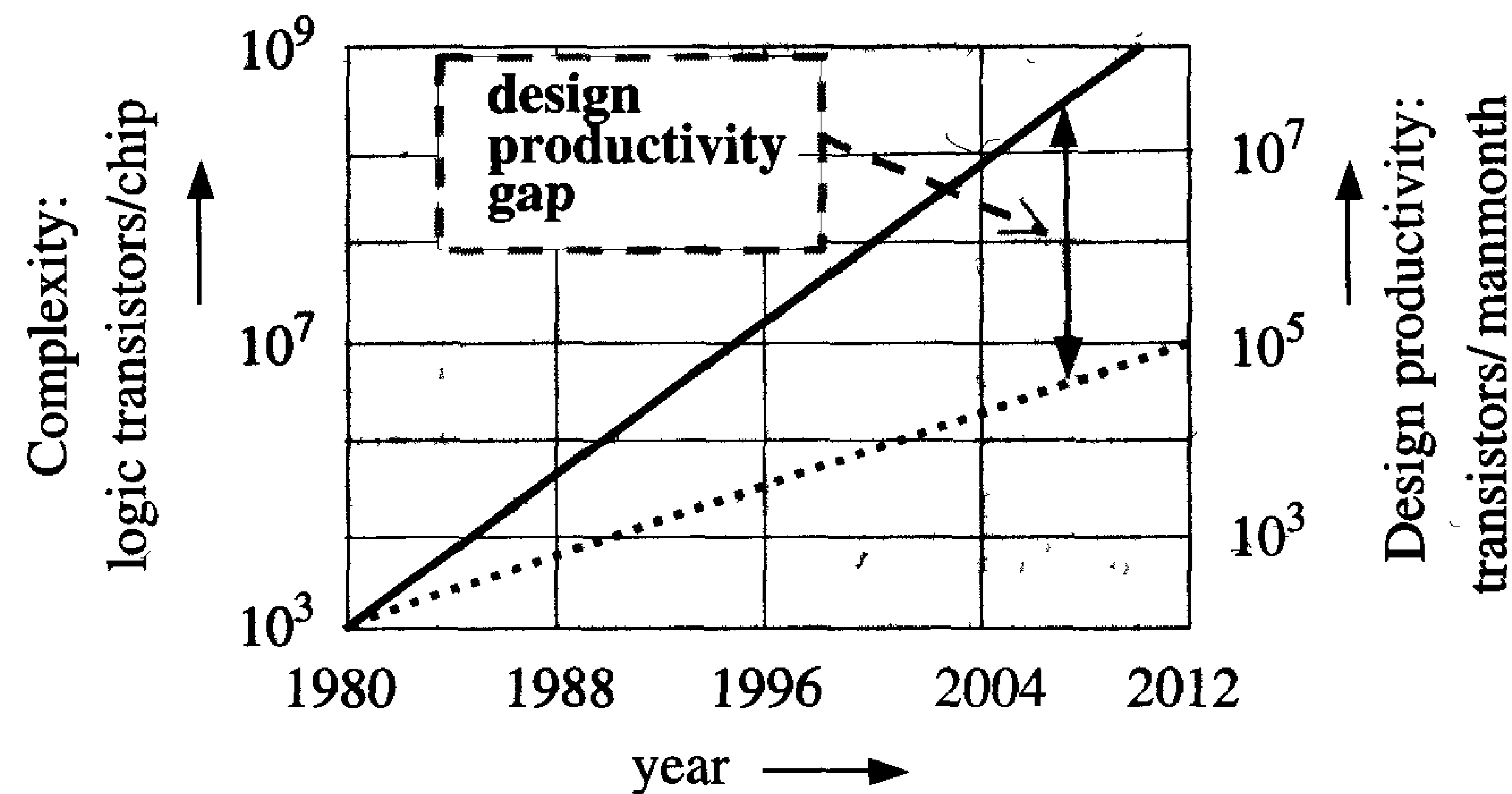


Figure 6.21: *The design productivity gap with respect to the growth in IC complexity*

Basically, there are three different approaches to implement a SOC.

The first one is to embed memories in a logic chip, integrated in a logic-based process (*embedded memory*). These memories could be SRAM, DRAM or even EEPROMs. For the latter two, additional processing steps are needed.

Higher levels of design and the increasing need for higher bandwidths drives the combination of huge memory capacities with processors on a single chip. These chips require the highest possible density of the memory blocks. This is the second approach: to embed logic (processors etc.) in a memory (mostly DRAM) process (*embedded logic*). A DRAM in a logic-based process will not be as compact as in a DRAM-based process, because this process has been optimised for it. Analogous to this, logic in a DRAM-based process will not be as compact as in a logic-based process, partly because DRAM processes use fewer metal layers than current logic processes.

However, the emerging graphics market requires very high speed DRAMs (see section 6.3.4) at limited power consumption, which drives the need for merged DRAM + logic processes (*Merged Memory Logic (MML)*). This is the third approach: to achieve the required logic density, an increased number of metal layers is added to a DRAM-based process. The decision to start from a DRAM with *embedded logic*, or from a logic process with embedded DRAM depends largely on the required memory capacity, the complexity of the logic part, the yield and the possible integration of IP cores.

Testing is a problem that arises with the merging of huge memory blocks with logic on a single chip. In a large-complexity, stand-alone memory, true memory performance can be measured because of the accessibility of the memory through the I/O pads. When such complex memories are embedded, direct accessibility through the pads is obviously less, because this is often done by multiplexing to I/O pads. BIST techniques are required to minimise testing costs and wafer handling.

Many vendors are currently developing efficient processes for the production of ASICs with large DRAM capacities. Most of them are major DRAM vendors. Although trench processes have been abandoned by most vendors since the $0.5\,\mu$m generation, they seem to be advantageous when DRAM is merged with logic.

As trenches lie below the silicon surface, the planarisation is much easier than when the capacitor is built from a stack of metal and dielectric layers on top of the silicon. This is particularly true when large memory areas are neighbouring logic with five to seven layers of metal. While Toshiba/IBM/Siemens are focusing on trenches, NEC is moving to a merged process without trenches. Toshiba's 256-Mbit DRAM technology forms the basis for a merged DRAM/ASIC technology that allows the integration of 128 Mbit of DRAM with about 400,000 gates on one single die. If this trend continues, pure ASIC vendors will have to find an alternative. Subcontracting to MML foundries may be an option for some of them. Others may offer their own, less area-efficient solutions. Customers usually make a trade-off between the relative size of the DRAM cells and the flexibility of DRAM core configuration. The pace at which MML devices will penetrate emerging graphics and other memory-intensive applications cannot be predicted. We will just have to wait and see.

6.6 Classification of the various memories

Table 6.2 provides an overview of the different types of memories with respect to some important parameters that characterise them. The numbers in table 6.2 are orders of magnitudes and may vary between different memory vendors. The characteristic values of these parameters render each type of memory suitable for application areas. These areas are summarised in table 6.1.

Table 6.1: Application areas for the various memory types

Memory type	Application areas
SRAM	Super-fast systems, low-power systems, cache memories in PCs, workstations, telecommunication, multimedia computers, networking applications, cellular telephones, supercomputers, mainframes, servers, embedded memories
DRAM	Medium to high speed, large computer systems, low-cost systems, large volumes, PC, hard disk drives, graphics boards, printer applications, PDAs, camcorders, embedded memories, embedded logic
FRAM	Low-power, non-volatile applications, smart cards, RF Identification, replacement of non-volatile RAM and potentially high-density SRAM
ROM	large volumes, video games, character generators, laser printer fonts, dictionary data in word processors, sound source data in electronic musical instruments embedded memories
EPROM	CD-ROM drives, modems, code storage,embedded memories
EEPROM	Military applications, flight controllers, consumer applications, portable consumer pagers, modems, cellular and cordless telephones, disk drives, printers, air bags, anti-lock braking systems, car radios, smart card, set-top boxes, embedded memories
FLASH	Portable systems, communication systems, code storage, memory PC cards, BIOS storage, digital cameras, ATA controllers, flash cards, palm tops, battery powered applications, digital cellular phones, embedded memories, MP3 players
NVRAM BRAM	Systems where power dips are not allowed, medical systems, space crafts, etc, which require fast read and write access

6.7 Conclusions

The MOS memory market currently represents about one third of the total IC market. This indicates the importance of their use in various applications. Most applications have different requirements on parameters such as memory capacity, power dissipation, access time, retention time and reprogrammability, etc. Modern integrated circuit technology facilitates the manufacture of a wide range of memories that are each optimised for one or more applications. The continuous drive for larger densities is leading to ever-increasing memory capacities and the limits are not yet in sight. The DRAM market shows the largest volumes and, not surprisingly, the highest demand for new technologies. This resulted in the presentation of the first 1 Gbit DRAM in 1995, ISSCC conference.

This chapter presents the basic operating principles of the complete range of memory types. Their characteristic parameters are compared in table 6.2 and their application areas are summarised in table 6.1.

Table 6.2: Characteristics of different memory types

DEVICE		*SRAM*	*DRAM*	*FRAM*	*ROM*	*PROM*	*EPROM*	*EEPROM*	*FLASH*	*NVRAM*	*BRAM*
components per cel		6	1.5	1.5	1	1.5	1	2.5	1	9	6
cell area		4 - 6	1.5	1.5	1	4	1.5	4	1.5	8	6
chip area		4 - 4.5	1.5	1.5	1	3	1.5	4	1.5	8	4.5
max. nº of programming/ erasing cycles		∞	∞	10^{10} - 10^{12}	1	1	10^4	10^5	10^4 - 10^6	SRAM : ∞ EEPROM: 10^5	∞
programmming time		10 - 100 ns	30 - 100 ns	150 - 200 ns	-	10 - 100 ms	5 - 10 µs	1 - 10 µs per byte or 100 µs (page mode)	5 - 10 µs	= SRAM or = EEPROM 5-10 ms all in parallel	= SRAM
access time		20 - 100 ns	30 - 100 ns	150 - 200 ns	10 - 100 ns	5 - 20 ns (bip.)	20 - 150 ns	5 - 150	20 - 150 ns	= SRAM	= SRAM
retention time	*no power supply*	0	0	> 10 years	∞	∞	> 10 years	> 10 years	> 10 years	> 10 years	> 7-10 years
	power supply	∞	2ms								

6.8 References

Information about memories is usually confidential and is often proprietary. Many of the relatively few books available on the subject are therefore outdated. This reference list is therefore limited to a recently published book and the titles of an interesting journal and digests on relevant conferences.

General

[1] B. Prince,
'Semiconductor Memories: A Handbook of Design, Manufacture and Application',
John Wiley & Sons, New York, 1996

[1a] W.J. McClean,
'Status 1999, A report on the IC industry',
ICE corporation, Scottsdale, Arizona, 1999

DRAM memory cells

[2] N. Lu,
'Advanced Cell Structures for Dynamic RAMs',
IEEE Circuits and Devices Magazine, January 1989

Further reading

[3] 'IEEE digest of technical papers of the International Solid State Circuit Conference'.
The ISSCC is held every year in February in San Francisco.

[4] IEEE Journal of Solid-State Circuits.

[5] IEDM Digest of technical Papers, since 1984.

High-performance memories

[6] B. Prince,
'High Performance Memories',
John Wiley & Sons, New York, 1996

[7] C. Hampel,
'High-speed DRAMs keep pace with high-speed systems',
EDN, February 3, 1997, pp 141-148

[8] C. Green,
'Analyzing and implementing SDRAM and SGRAM controllers',
EDN, February 2, 1998, pp 155-166

[9] *www.chips.ibm.com/products/memory.*
'Understanding Video (VRAM) and SGRAM Operation'

[10] D. Bursky,
'Graphics-Optimized DRAMs deliver Top-Notch Performance',
Electronic design, March 23, 1998, pp 89-100

[11] B. Dipert,
'FRAM: ready to ditch niche?',
EDN, April 10, 1997, pp 93-107

[12] B. Dipert,
'EEPROM, survival of the fittest',
EDN, January 15, 1998, pp 77-90

[13] B. Dipert,
'Data storage in a Flash',
EDN, July 3, 1997, pp 65-77

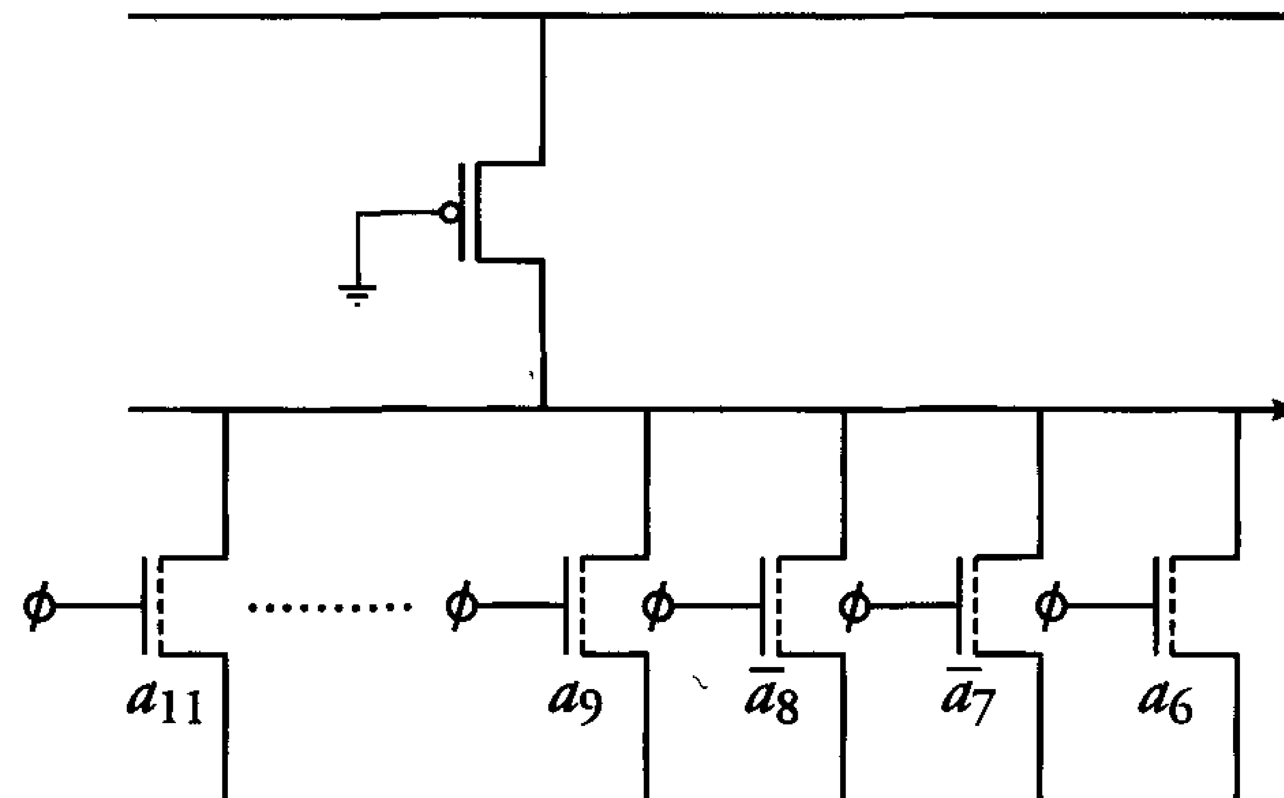
[14] K. Noda, et al.
'A 1.9 μm^2 Loadless CMOS Four Transistor SRAM Cell in a 0.18 μm Logic Technology',
IEDM Digest of Technical Papers, December 1998, pp 643-646

[15] K. Takeda, et al.
'A 16 Mb 400 MHz loadless CMOS 4-Transistor SRAM Macro',
ISSCC Digest of Technical Papers, February 2000

[16] K. Kishiro, et al.
'Structure and Electrical Properties of Thin Ta_2O_5 Deposition on Metal Electrodes',
Japan Journal on Applies Physics, vol. 37, 1998, pp 1336-1339

[17] J.S. Mayo,
Scientific American, 1981

6.9 Exercises



1. Assume that the column decoder in figure 6.3 is implemented in nMOS as shown in the adjacent figure and the column address is $a_{11}a_{10}a_9a_8a_7a_6$=010110.

   a) Describe the data flow in figure 6.3 during a read operation when word line x_{20} is also selected.

   b) What is the major disadvantage of such a decoder?

   c) What would be the problem if this decoder were implemented in static CMOS?

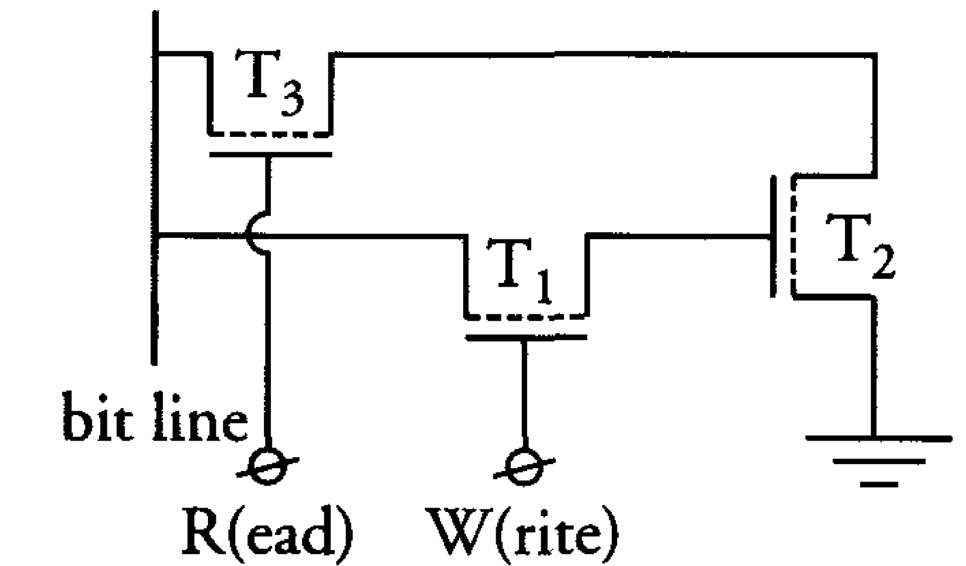
2. Describe the major differences between the ROM realisations of figures 6.12 and 6.15. Explain their relative advantages and disadvantages.

3. Why does a stand-alone flash EPROM usually require one more power supply than a full-featured EEPROM?

4. Table 6.2 gives a summary of some important memory parameters.

   a) Explain the difference in chip area between a non-volatile RAM and an SRAM.

   b) Explain the difference in access times between an SRAM and a DRAM.



5. The adjacent figure shows a dynamic memory cell which consists of three transistors. This is a so-called *3 T-cell*.

   a) Explain the operation of the 3 T-cell.

   b) What can be said about the read-out data after one write and read cycle?

   c) Comment on the size of the storage nodes in the 3 T-cell and the 1 T-cell?

6. What is a multilevel flash memory?
   What would be their main problem for future process generations?

7. Explain the difference between an embedded memory and embedded logic.

Chapter 7

Very Large Scale Integration (VLSI) and ASICs

7.1 Introduction

The continuing development of IC technology during the last couple of decades has led to a considerable increase in the number of devices per unit chip area. The resulting feasible IC complexity currently allows the integration of complete systems on single chips, which may comprise tens to hundreds of millions of transistors.

Consequently, the design of such chips no longer simply consists of the assembly of a large number of logic gates. This poses a problem at a high level of design: how to manage the design complexity. Besides this, measures must also now be taken for parasitic effects (see chapters 2, 9 and 11), which may reduce chip performance dramatically.

The design of a modern complex integrated circuit requires a considerable number of man-years. Such ICs combine signal processing capacity with microprocessor or microcontroller cores. The dedicated signal processing parts take care of the computing power (workhorse), while the microprocessor or controller serves to control the process and possibly performs some low performance computation as well. The development of such heterogeneous systems on one or more ICs, for instance, may require tens of man-years.

A significant amount of total IC turnover is generated in the "low-end market". This market consists of low-complexity ICs and was orig-

inally controlled by the large IC vendors. During the eighties, however, a change took place and the low-end market is now dominated by *Application-Specific Integrated Circuits* (ASICs). These are ICs which are realised for a single end-user and dedicated to a particular application. ASICs therefore implement customer-specified functions and there are various possibilities for the associated *customisation*. This can be an integral part of an IC's design or production process or it can be accomplished by programming special devices.

ASICs do not include ICs whose functionality is solely determined by IC vendors. Examples of these *"Application-Specific Standard Products"* (ASSPs) include digital-to-analogue (D/A) converters in compact disc players. Actually, *User-Specific Integrated Circuits* (USICs) would be a more appropriate name for ASICs. The use of USICs would clearly be preferable because it emphasises the fact that the IC function is determined by the customer's specification and not simply by the application area.

The turn-around time of an ASIC is the period which elapses between the moment a customer supplies an IC's logic *netlist* description and the moment the vendor supplies the first samples. The *turn-around time* associated with an ASIC depends on the chosen implementation type. A short turn-around time facilitates rapid prototyping and is important to company marketing strategies. In addition, ASICs are essential for the development of many real-time systems, where designs can only be verified when they are implemented in hardware.

Suitable computer aided design (CAD) tools are therefore essential for the realisation of modern ICs. These tools should allow efficient use of methods that reduce the complexity of the design task and reduce time to market. The design process is discussed on the basis of an ASIC design flow. The various implementation possibilities for digital VLSI and ASICs are discussed and factors that affect a customer's implementation choice are examined. Market trends and technological advances in the major ASIC sectors are also explained.

7.2 Digital ICs

Digital ICs can be subdivided into different categories, as shown in figure 7.1. ASICs can be classified according to the processing or programming techniques used for their realisation. A clear definition of the types and characteristics of available digital ICs and ASICs is a prerequisite for the subsequent discussion of the trends in the digital ASIC market. Figure 7.1 presents quite a broad overview of digital ICs but excludes details such as the use of *direct slice writing* (DSW) or masks for IC production. Several terms used in this figure and throughout this chapter are explained on the next page.

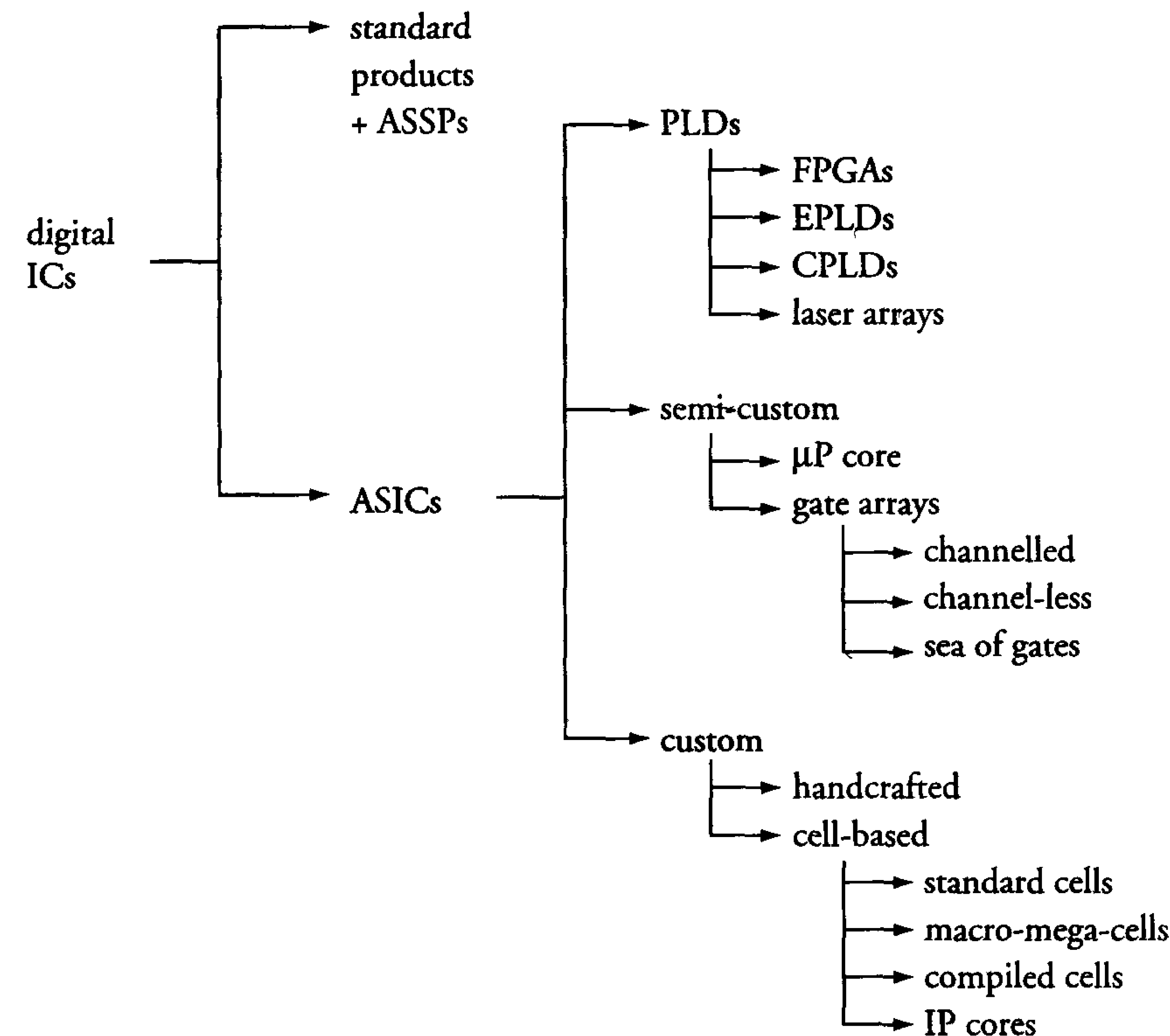


Figure 7.1: *An overview of digital ICs*

Definitions:

ASSP: *Application-Specific Standard Products* are ICs that are suitable for only one application but their availability is not restricted to a single customer. Examples include video ICs for teletext decoding and ICs for D/A conversion in compact disc players.

Core: Pre-designed industry (or company) standard building block: RAM, ROM, microprocessor (e.g. ARM, MIPS and Sparc), etc.

Custom: A custom IC is an IC in which all masks are unique for a customer's application. The term *full-custom IC* is often used to refer to an IC in which every sub-circuit is a new handcrafted design. In this book, full-custom ICs fall under the category of custom ICs. Cell-based custom-IC designs are based on standard cells, *macro cells*, *mega cells* and possibly *compiled cells*. Macro and mega cells, or cores are large library cells like multipliers, RAMs, ROMs and even complete microprocessors and signal processors. Compiled cells are automatically generated by modern software libraries. These cells are used for dedicated applications and are generated as a function of user-supplied parameters.

The customisation of PLD-based ASICs takes place after IC manufacture. Customisation of custom and semi-custom ASICs, however, is an integral part of IC manufacture. The turn-around time of ASICs from database ready to first silicon varies enormously and depends on circuit complexity and the customisation technique. This time can range from a few hours for a PLD to between 6 and 12 weeks for a custom design.

IP: Intellectual Property. With the complexity of ICs exceeding tens of millions of transistors, the traditional way of designing can no longer be continued. Therefore, the concept of *Virtual Component* has been introduced by the Virtual Socket Interface Alliance (VSIA), which is an international forum trying to standardise reusable cores, concepts, interfaces, test concepts and support, etc. Licensing and royalty issues of IP must also be addressed. This standardisation is a prerequisite to fully exploit the potentials of design reuse. The cores (or IP) can be represented in three forms.

A *soft core* is delivered in the form of synthesizable HDL, and has the advantage of being more flexible and the disadvantage of not

being as predictable in terms of performance (timing, area, power). Soft cores typically have increased intellectual property protection risks because RTL source code is required by the integrator.

Firm cores have been optimised in structure and in topology for performance and area through floor planning and placement, possibly using a generic technology library. The level of detail ranges from region placement of RTL sub-blocks, to relatively placed data paths, to parameterised generators, to a fully placed netlist. Often, a combination of these approaches is used to meet the design goals. Protection risk is equivalent to that of soft cores if RTL is included, and is less if it is not included.

Finally, *hard cores* have been optimised for power, size or performance and mapped to a specific technology. Examples include netlists fully placed, routed and optimised for a specific technology library, a custom physical layout or the combination of the two. Hard cores are process- or vendor-specific and generally expressed in the GDSII format. They have the advantage of being much more predictable, but are consequently less flexible and portable because of process dependencies. The ability to legally protect hard cores is much better because of copyright protections and there is no requirement for RTL.

Figure 7.2 is a graphical representation of a design flow view and summarises the high level differences between soft, firm and hard cores.

PLD: *Programmable Logic Devices* are ICs that are customised by blowing on-chip fuses or by programming on-chip memory cells. PLDs can be customised by end-users themselves in the field of application, i.e. they are *field-programmable* devices (FPGA). The customisation techniques used are classified as *reversible* and *irreversible*. PLDs include *erasable* and *electrically erasable* types, which are known as *EPLDs* and *EEPLD*, respectively. The former are programmed using EPROM techniques while the EEPROM programming technique is used for the latter devices. These programming techniques are explained in section 6.4. Complex PLDs (CPLDs) are often based on the combination of PAL^{TM} and PLA architectures.

Reuse: Future design efficiency will increasingly depend on the availability of a variety of pre-designed building blocks (IP cores). This *reuse* not only requires easy portability of these cores between different ICs, but also between different companies. Standardisation is one important issue, here (see IP definition). Another important issue concerning reuse is the quality of the (IP) cores. Similar to the Known-Good Die (KGD) principle when using different ICs in an MCM, we face a Known-Good Core (KGC) principle when using different cores in one design. The design robustness of such cores must be so high that their correctness of operation will always be independent of the design in which it is embedded.

Semi-Custom: These are ICs in which one or more but not all masks are unique for a customer's application. Many semi-custom ICs are based on 'off-the-shelf' ICs which have been processed up to the final contact and metal layers. Customisation of these ICs therefore only requires processing of these final contacts and metal layers. This results in short turn-around times. A gate array is an example in this semi-custom category.

Standard product: Standard products, also called *standard commodities*, include microprocessors, memories and standard-logic ICs, e.g. NAND, NOR, QUAD TWO-INPUT NAND, etc. These ICs are produced in large volumes. Their availability is unrestricted and they can be used in a wide variety of applications. They are often put into a product catalogue.

Usable gates: The number of gates that can actually be interconnected in an average design. This number is always less than the total number of available gates (gate array).

Utilisation factor: The ratio between that part of a logic block area which is actually occupied by functional logic cells and the total block area (gate array and cell-based designs).

	Design flow	Representation	Libraries	Technology	Portabillity
Soft not predictable very flexible	system design RTL design	behavioural RTL	N/A	technology independent	unlimited
Firm flexible predictable	floor planning synthesis placement	RTL & blocks netlist	reference library • footprint • timing model • wiring model	technology generic	library mapping
Hard not flexible very predictable	routing verification	polygon data	process-specific library & design rules • characterised cells • process rules	technology fixed	process mapping

Figure 7.2: *Graphical representation of soft, firm and hard cores (source: VSIA)*

7.3 Abstraction levels for VLSI

7.3.1 Introduction

The implementation of a complete system on one or more ICs starts with an abstract system level specification. This specification is then analysed and transformed into a set of algorithms or operations. Next, an architecture that performs these operations must be chosen. A signal processor serves as an example. The chosen processor must perform an adaptive FIR filter. As a consequence, this processor must repeatedly fetch numbers from a memory, multiply or add them and then write the result back into the memory. Such a chip may contain several ROM or RAM memory units, a multiplier, an adder or accumulator, data and control buses and some other functional modules.

The design of an IC comprises the transformation of a specification into a layout. The layout must be suitable for the derivation of all process steps required for the manufacture of the IC's functional modules and their interconnections. Clearly, the *design path* starts at the top (or system) level and ends at the bottom (or silicon) level. This *'top-down'* process is illustrated in figure 7.3.

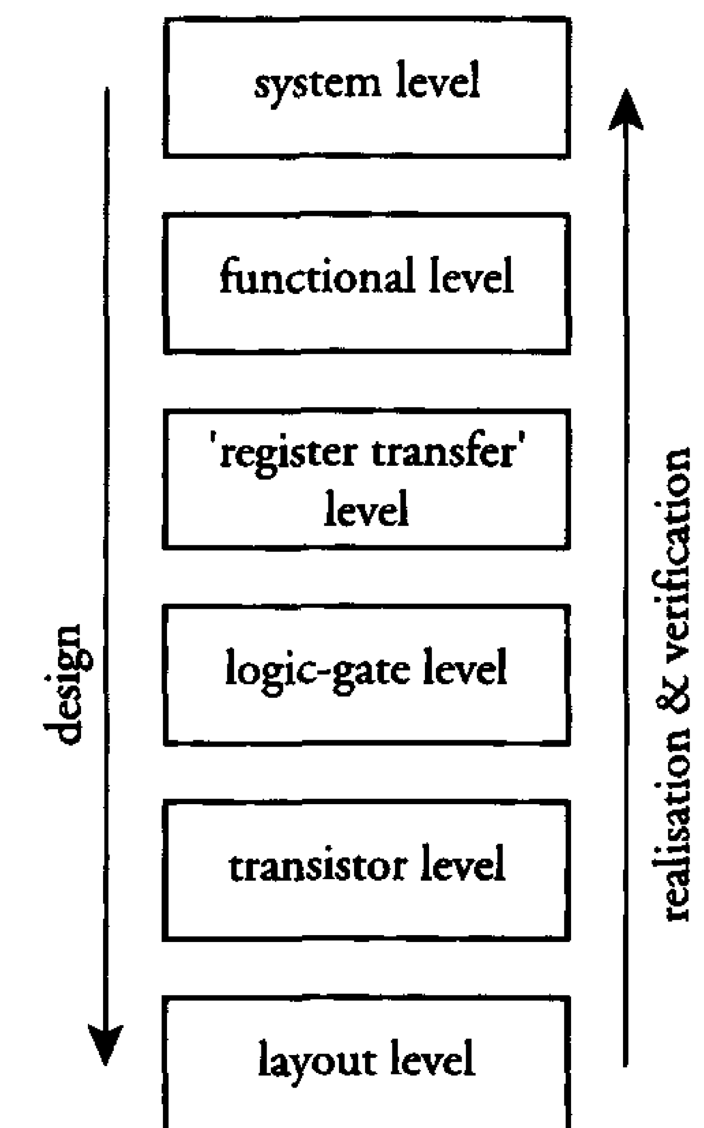


Figure 7.3: *Abstraction levels in the design and implementation/verification paths of VLSI circuits*

The various design phases are accompanied by several different *abstraction levels*, which limit the complexity of the relevant design description. The top-down design path allows one to make decisions across abstraction levels and gives high level feedback on specifications. The *'bottom-up'* path demonstrates the feasibility of the implementation of (critical) blocks. This process begins at the layout level of a single part and finishes with the verification of the entire IC layout. The abstraction levels that are used in the design path are described on the following pages. Table 7.1 shows the design complexity at these levels of abstraction.

Table 7.1: Design complexity at different levels of abstraction

Level	Example	Number of elements	
system	heterogeneous system	10^7-10^8	transistors
functional	signal processor	10^5-10^7	transistors
register	digital potentiometer	10^3-10^5	transistors
logic gate	Library cell (NAND, full adder)	2-30	transistors
transistor	nMOSt, pMOSt	1	transistor
layout	signal processor	10^8-10^9	rectangles

7.3.2 System level

A system is defined by the specification of its required behaviour. Such a system could be a multiprocessor system and/or a *heterogeneous system*, consisting of different types of processing elements: microprocessor cores, signal processor cores and memories, etc. Figure 7.4 shows a heterogeneous system, containing a signal processor, a microprocessor, embedded software and some glue logic (some additional overall control logic). The transformation of a system into one or more ICs is subject to many constraints on timing, power and area, for example.

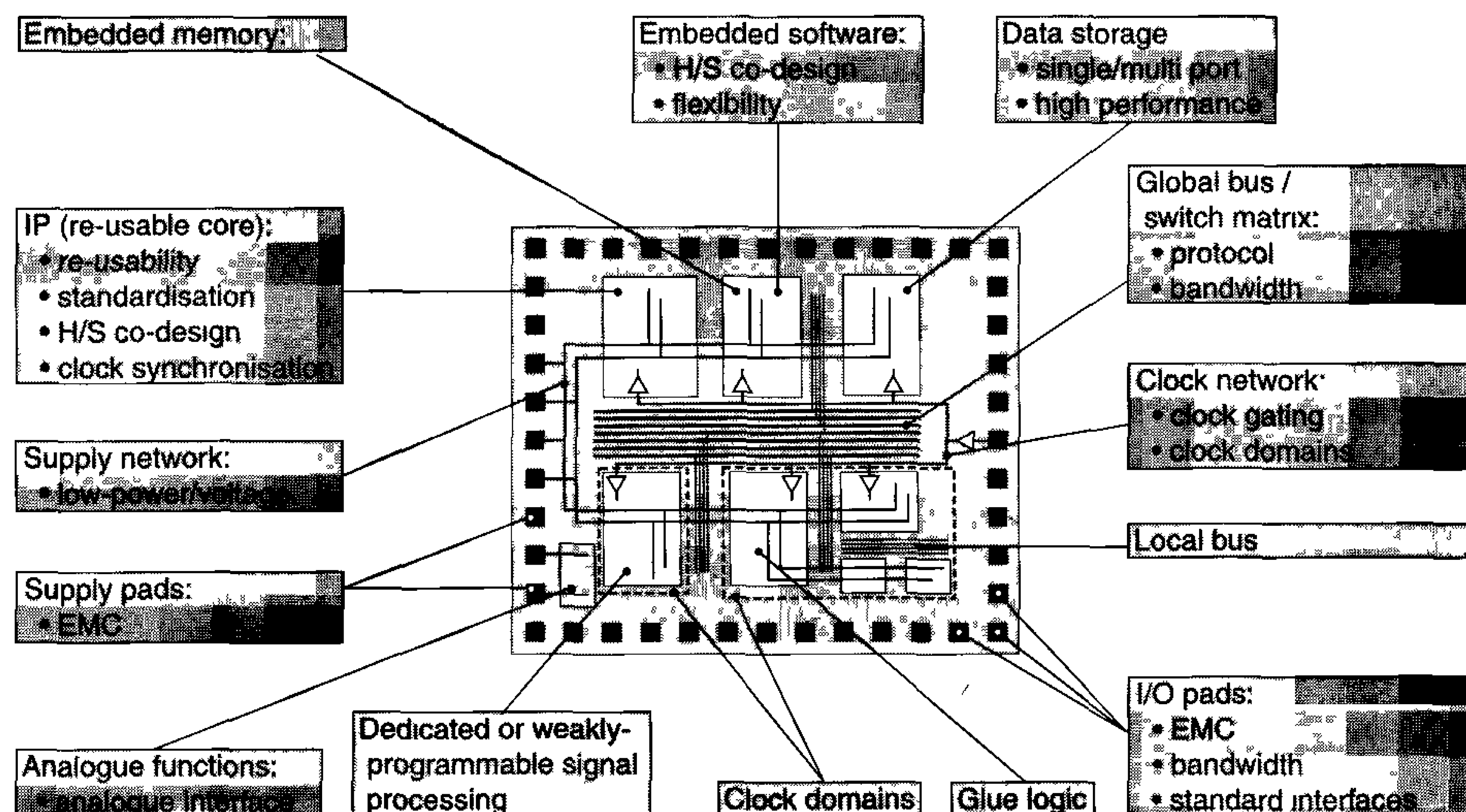


Figure 7.4: *Systems on a chip; an example of a heterogeneous system*

Decisions regarding functions that are to be implemented in hardware or software are made at the system level. Filter sections, for example, are frequently programmed in software. A system-level study should also determine the number of chips required for the integration of the chosen hardware. It is generally desirable to sub-divide each chip into several sub-blocks. For this purpose, *data paths* and *control paths* are often distinguished. The former is for data storage and data manipulation, while the latter controls information flow in the data path, and to and from the outside world. Each block in the data path may possess its own *microcontrol unit*. This usually consists of a decoder which recognises a certain control signal and converts it into a set of instructions.

The block diagram shown in figure 7.5 represents a description of a

signal processor at the system abstraction level. The double bus structure in this example allows parallel data processing. This is typically used where a very high data throughput is required. Data can be loaded into the Arithmetic Logic Unit (ALU) simultaneously from the ROM and the RAM. In this type of architecture, the data path and control path are completely separated. The control path is formed by the program ROM, which may include a program counter, control bus and the individual microcontrol units located in each data path element.

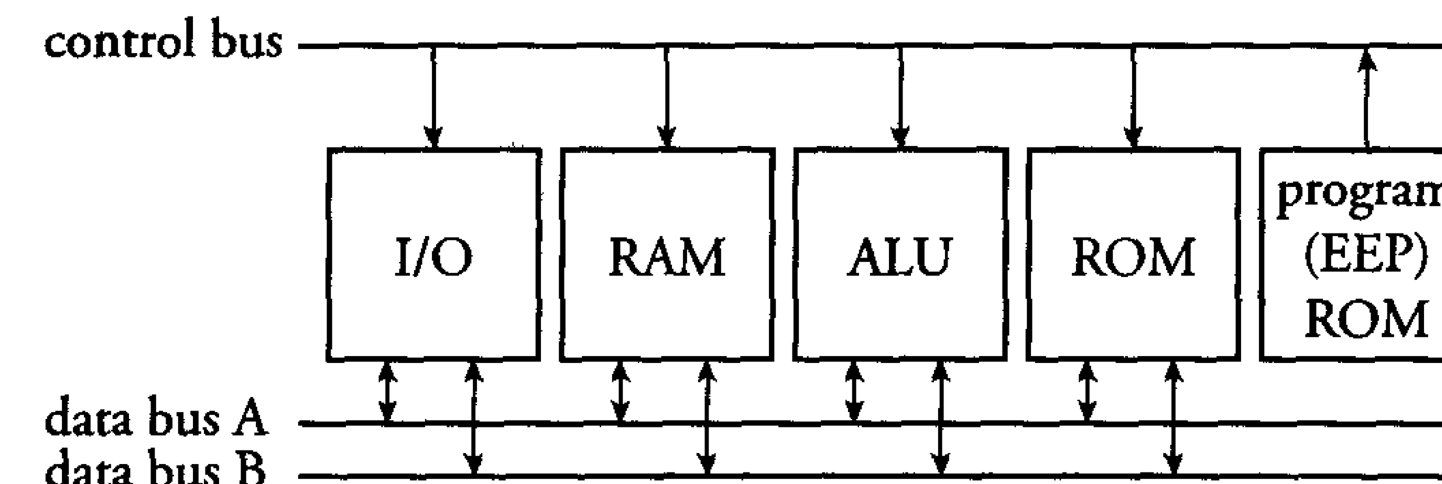


Figure 7.5: *Block diagram of a signal processor*

Other system implementations may not show such a clear separation of data and control paths.

7.3.3 Functional level

A description at this level of abstraction comprises the behaviour of the different processing elements and cores of the system. These elements may include a microprocessor core, a signal processor core, a RAM, a ROM and the I/O element, etc.

RAMs, ROMs and I/O elements are usually not very complex in their behaviour, which makes them suited for a description at this level. As a result of the simplicity of their behaviour, however, they are mostly described in the next, lower level of abstraction, the RTL level. Let us take the signal (or micro)processor core as an example at his level of abstraction. Both a microprocessor core and a signal processor core can be a system themselves. They are often composed of descriptions at the lower RTL level.

However, there are some tools in research or development which allow a description of such a processor at functional level. The maturity and ease of use of these tools is not yet such that they are common part of current design flows.

The chosen signal processor may consist of different processing units: ALU, digital potentiometers and again some memories: ROM and/or RAM. Actually, these processing units are functions as well, so the RTL level and functional level often do not just include one level of design. They might even show some overlaps (see also figure 7.12).

7.3.4 RTL level

RTL is an abbreviation for Register-Transfer Language. This notation originates from the fact that most systems can be considered as collections of registers that store binary data, which is operated on between these registers. The operations can be described in an RTL and may include complex arithmetic manipulations. The *RTL description* is not necessarily related to the final realisation.

At this level, we focus on the digital potentiometer as an example. The behaviour of this potentiometer can be described as:

$$Z = k \cdot A + (1 - k) \cdot B$$

When $k = 0$, Z will be equal to B and when $k = 1$, Z will be equal to A. The description does not yet give any information about the number of bits in which A, B and k will be realised. This is one thing that must be chosen at this level. The other choice to be made here is what kind of multiplier must perform the required multiplications. There are several alternatives for multiplier implementation, of which some are discussed as examples.

- *Serial-parallel multiplier:* The Ra input is bit-serial and the Rb input is bit-parallel, see figure 7.6.

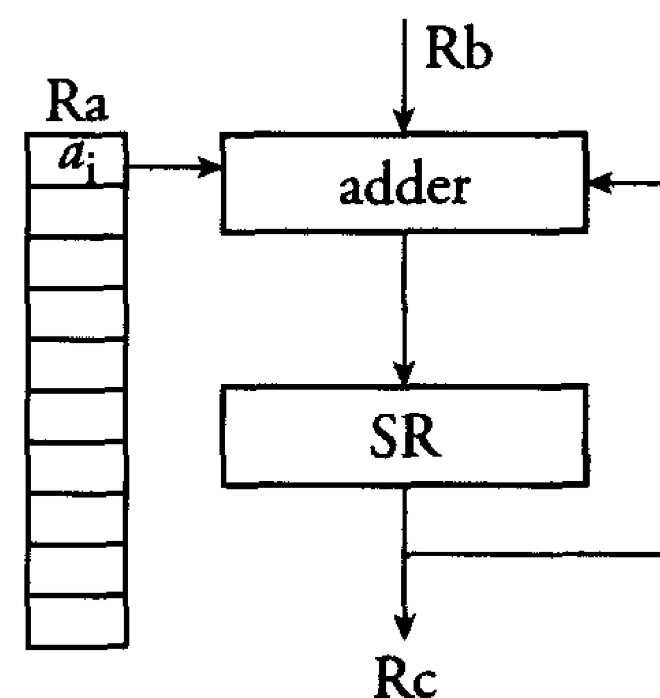


Figure 7.6: *Example of a bit-serial iterative multiplier*

During the execution of a multiplication, the *partial product* is present on the multiplier's parallel Rc output bits. These are initially zero.

If a_i=1, for instance, then the Rb bits must be shifted one place to the left and added to the existing partial product. This is a *'shift-and-add' operation*. When a_i=0, the Rb bits only have to be shifted one place to the left in a 'shift' operation and a zero LSB added to it.

- *Parallel multiplier:* The bits of both Ra and Rb are supplied and processed simultaneously. This *'bit-parallel' operation* requires a hardware realisation of the multiplier. Options include the *array or parallel multiplier*, schematically presented in figure 7.7.

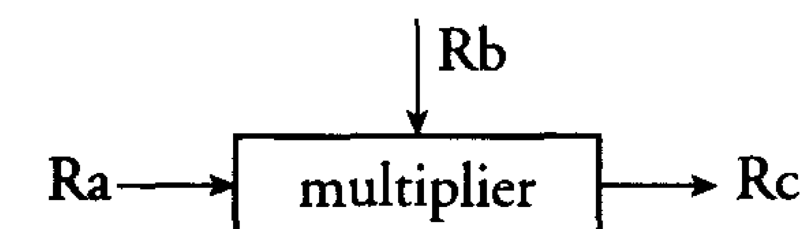


Figure 7.7: *A parallel multiplier*

The array multiplier necessitates the choice of a structure for the addition of the partial products. The possibilities include the following:

- *Wallace tree*: Here, bits with equal weights are added together in a tree-like structure, see figure 7.8.
- *Carry-save array*: Figure 7.9 illustrates the structure of this array, which consists of full adders that produce all the individual $x_i \cdot y_i$ product bits.

As an example, at this level, we choose the array multiplier with carry-save array. This would lead to a different behaviour from the serial multiplier, and thus to a different RTL description.

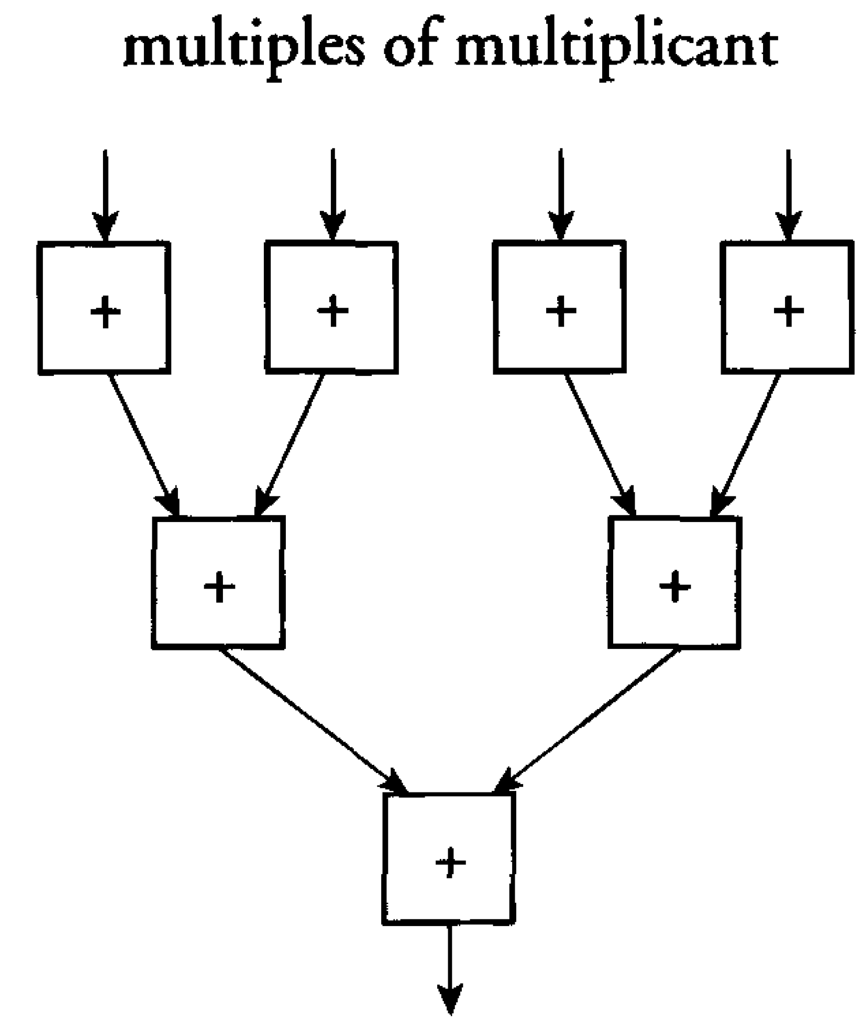


Figure 7.8: *Wallace tree addition*

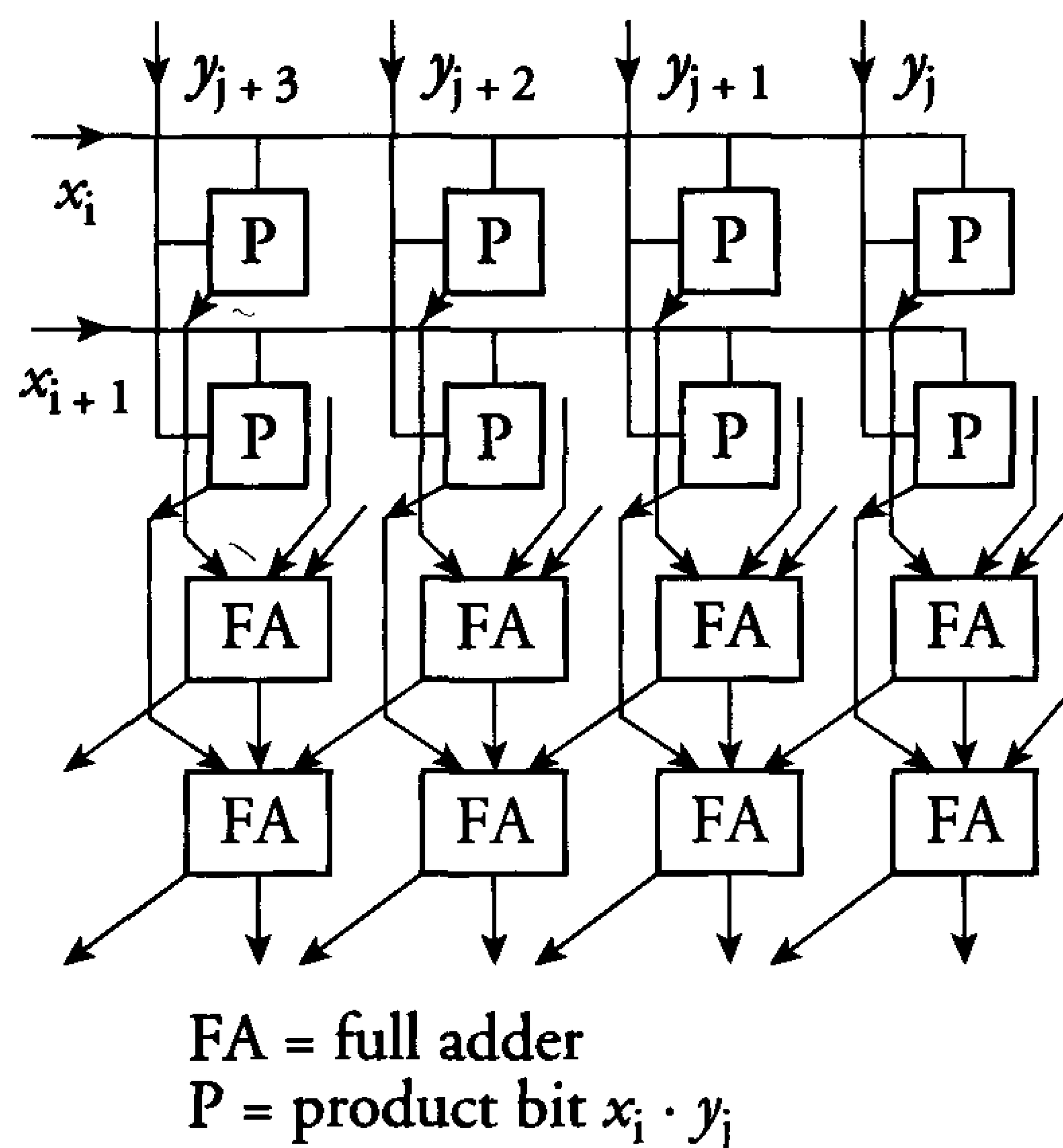


Figure 7.9: *Array multiplier with carry-save array*

7.3.5 Logic-gate level

As stated in section 7.4, the RTL description is usually mapped onto a library of cells (logic gates). This is done by an RTL synthesis tool, which transforms a VHDL code into a netlist. A *netlist* contains a list of the library cells used and how they are connected to each other. Examples of such library cells (logic gates) are: AND, NAND, flip-flop and full adder, etc. Suppose we choose the full adder, from which we will build the array multiplier. A full adder performs the binary addition of three input bits (x, y and z) and produces sum (S) and carry (C) outputs. Boolean functions that describe the operation of a *full adder* include the following:

(a) Generation of S and C directly from x, y and z:

$$C = x\,y + x\,z + y\,z$$
$$S = x\,\overline{y}\,\overline{z} + \overline{x}\,\overline{y}\,z + \overline{x}\,y\,\overline{z} + x\,y\,z$$

(b) Generation of S from C:

$$C = x\,y + x\,z + y\,z$$
$$S = \overline{C}(x + y + z) + x\,y\,z$$

(c) Generation of S and C with exclusive OR gates (EXORs).

The choice of either one of these implementations depends on what is required in terms of speed, area and power. Implementation (b) will contain fewer transistors than (a), but will be slower because the carry must first be generated before the sum can evaluate. The implementation in (c) is just to show another alternative. Suppose that area is the most dominant criterion, then, at this hierarchy level, we choose implementation (b) to realise our full adder. A logic-gate implementation is shown in figure 7.10.

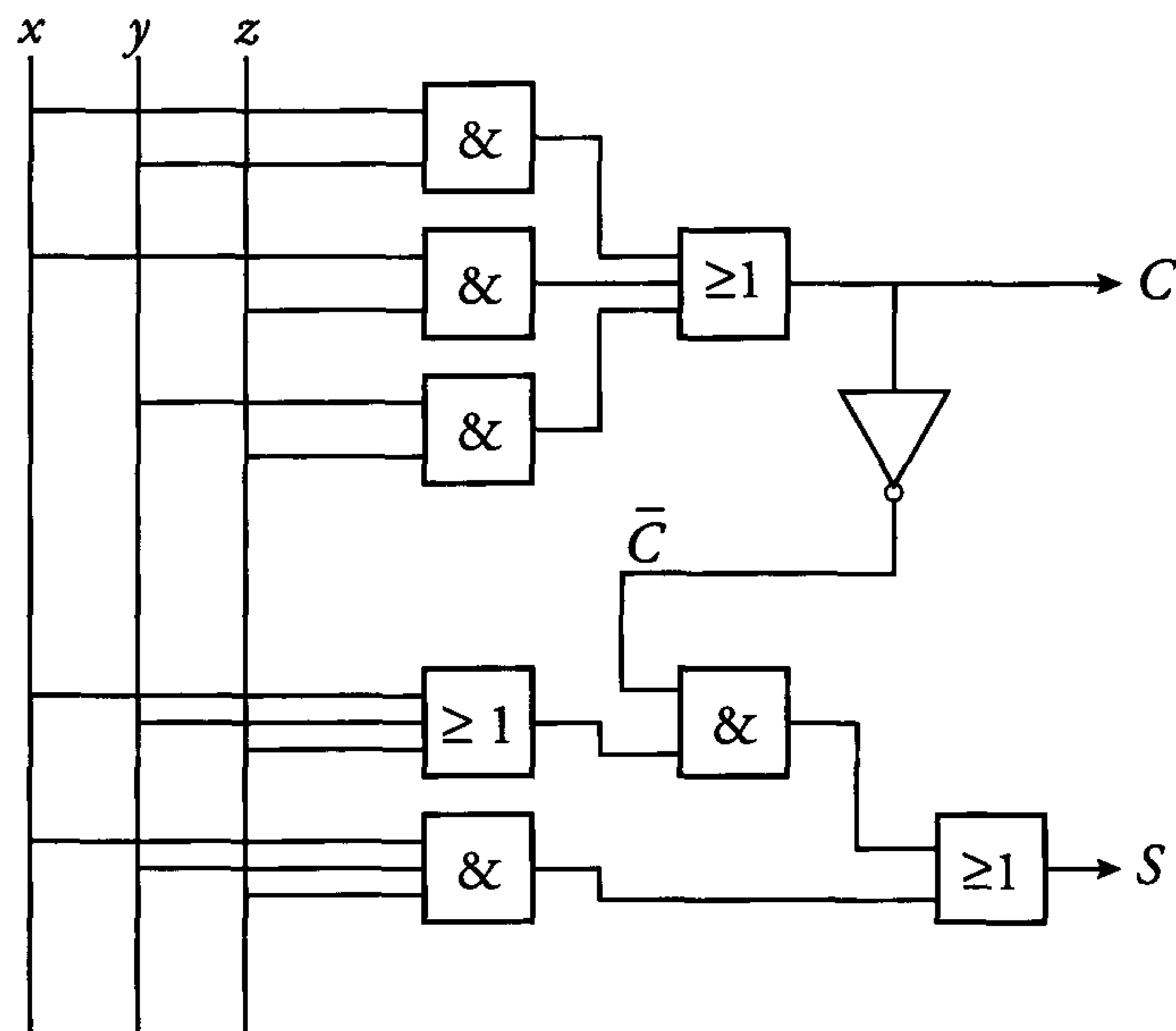


Figure 7.10: *Basic logic-gate implementation of a full adder*

7.3.6 Transistor level

At this level, the chosen full adder must be mapped onto a number of transistors. In some design environments, the logic-gate level is not explicitly present and the higher level code is directly synthesized and mapped onto a 'sea of transistors'. These are discussed in section 7.6. The transistor level description depends on the chosen technology (bipolar, nMOS, BICMOS, CMOS, etc.) and the chosen logic style, such as dynamic or static CMOS. For the realisation of our full adder, we choose a static CMOS implementation, as shown in figure 7.11.

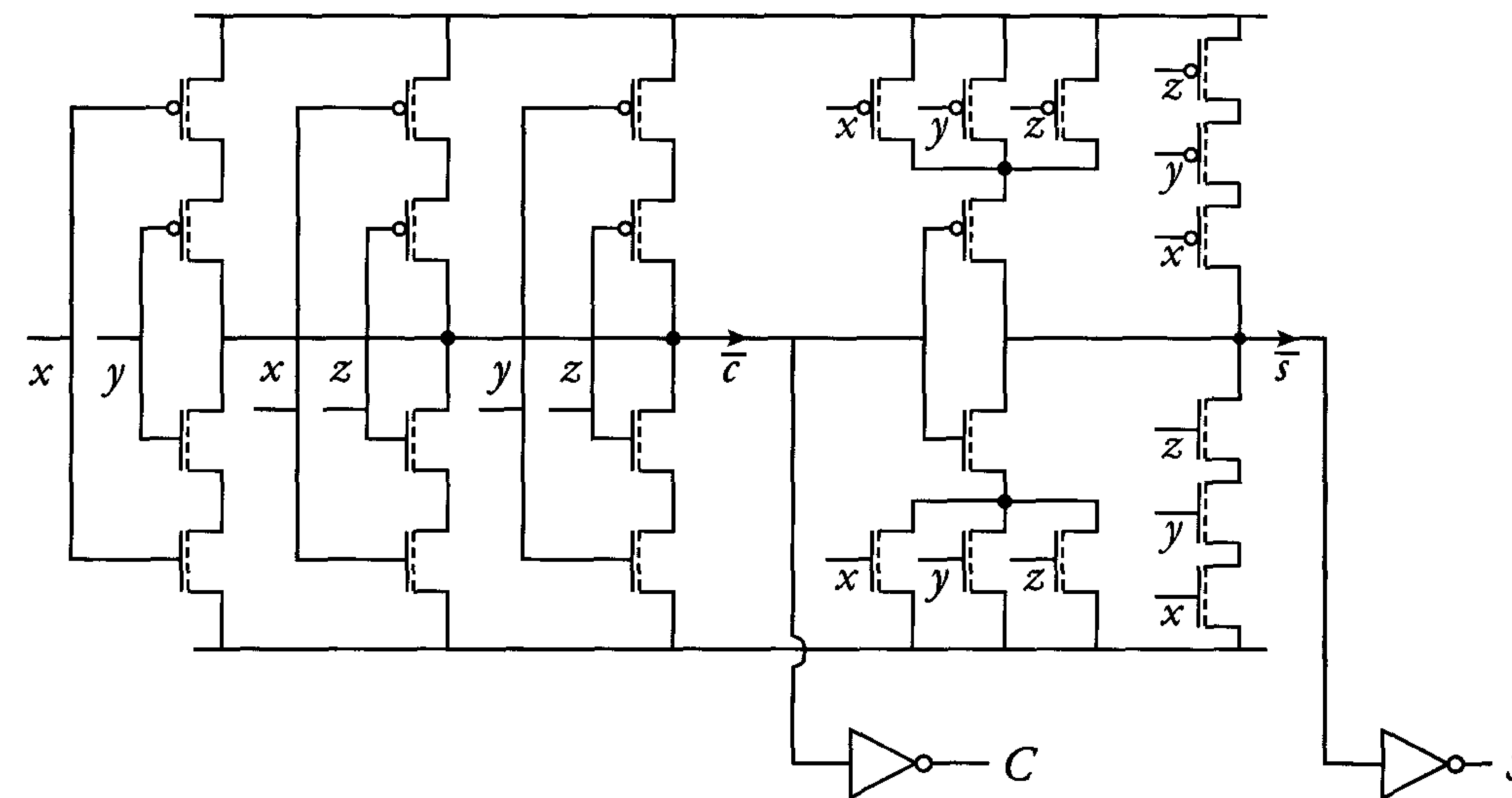


Figure 7.11: *Static CMOS realisation of the chosen full adder cell*

As this full adder consists of a relatively low number of transistors (30), it is efficient, both in terms of area and power dissipation, compared to the one realised with AND, OR and INVERT gates, see figure 7.10.

Thus, the transistor level implementation of the logic gate is determined by either speed, area or power demands, as is actually every IC implementation.

7.3.7 Layout level

The chosen transistor implementation must be translated to a layout level description at the lowest abstraction level of a design. Most of the time, these layouts are made by specialists, who develop a complete library of different cells in a certain technology. However, special requirements on high speed or low power may create the need for custom design, to optimise (part of) the chip for that requirement. In chapter 4, the layout process is explained in detail.

7.3.8 Conclusions

In the top-down design path, decisions have to be made at each level about different possible implementations. In this way, a *decision tree* arises. Figure 7.12 shows an example of a decision tree for a signal processor system.

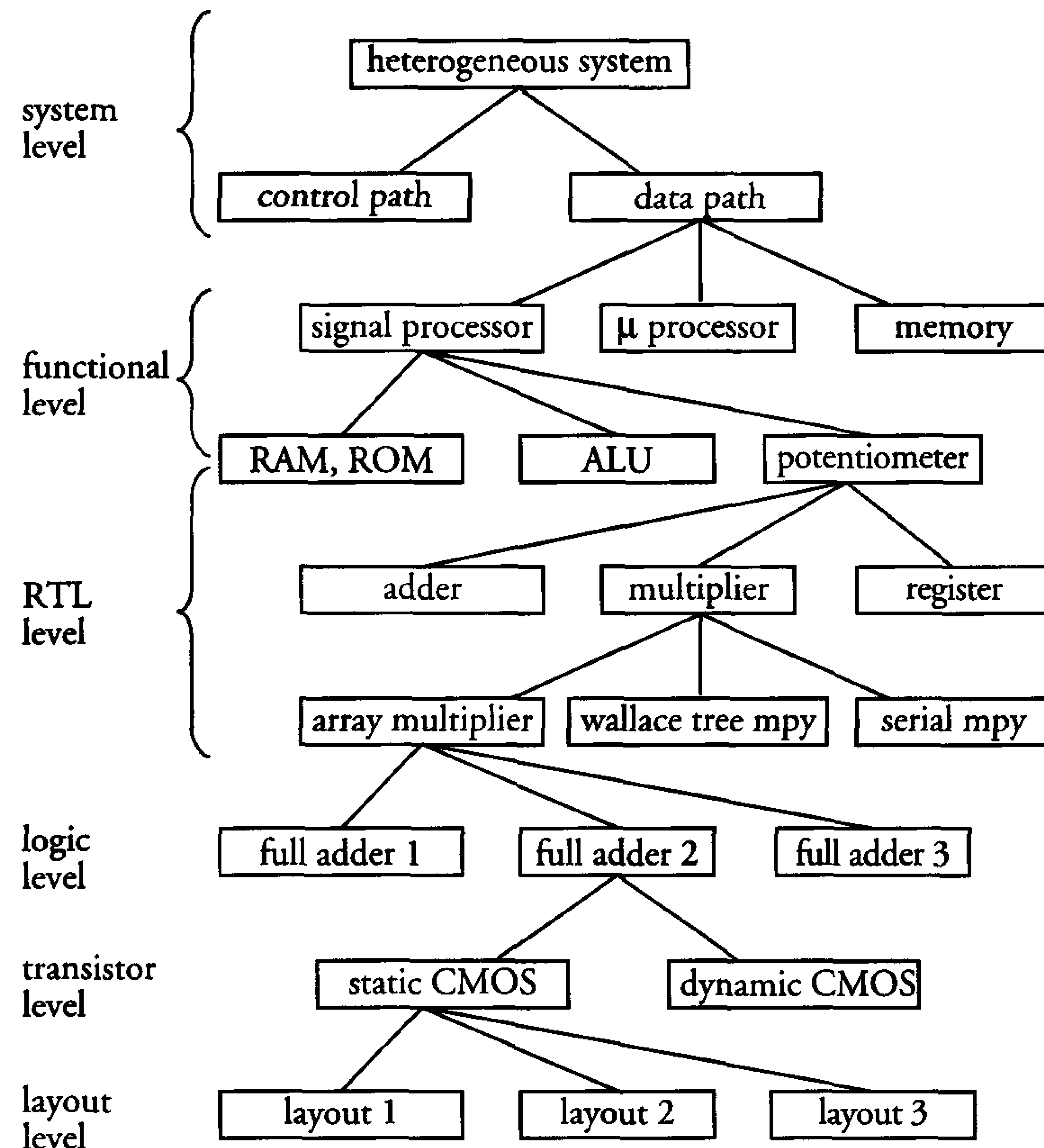


Figure 7.12: *Decision tree for a complex system on a chip*

The decision tree starts at the highest level, i.e. the system level. However, the decisions at each level can be strongly dependent on the possibilities available at a lower or at the lowest level. System designers who wish to achieve efficient area implementations therefore require a reasonable knowledge of IC techniques. For instance, the decision to implement a double data bus structure (figure 7.5) requires twice as many interconnections as a single bus implementation. This will be easier if multiple metal layers are available in the technology.

Decision trees and abstraction levels basically reduce the complexity of design tasks to acceptable levels. However, the abstraction levels are also accompanied by verification problems. More levels can clearly increase verification difficulties. Requirements at a certain level of abstraction depend on details at a lower level. Details such as propagation delays, for example, can influence higher level timing behaviour.

For example the final layout implementation of a full adder clearly influences its electrical behaviour. Delay times are also determined by factors such as parasitic wiring capacitances.

The bottom-up implementation and verification process begins at the layout level. Cell layouts are assembled to form modules, and these are combined to form the larger units that are indicated in the *floor plan* of the IC. The floor plan is a product of the top-down and bottom-up design process and is an accurate diagram which shows the relative sizes and positions of the envisaged IC components. Modules that are identified as *critical* during the design path are usually implemented first. These are modules which are expected to present problems for power dissipation, area or operating frequency. Verification of their layouts reveals whether they are adequate or whether an alternative must be sought. This may have far-reaching consequences for the chosen architecture.

The inter-dependence of various abstraction levels and implementations clearly prevents a purely top-down design followed by purely bottom-up implementation and verification. In practice, the design process generally consists of iterations between the top-down and bottom-up paths.

Abstraction level descriptions which contain sufficient information about lower-level implementations can limit the need for iterations in the design path and prevent wasted design effort. The maximum operating frequency, for example, of a module is determined by the path with the longest delay between two of its clocked latches. This *worst-case delay path* can be determined from suitable abstraction level descriptions and used to rapidly determine architecture feasibility. As an example, the multiplier in the previously-discussed signal processor is assumed to contain the worst-case delay path.

The dimensions of logic cells in a layout library, for example, could be used to generate floor plan information such as interconnection lengths. These lengths, combined with specified delays for the library cells (e.g. full adder, multiplexer, etc.) allow accurate prediction of performance. The worst-case delay path can eventually be extracted from the final multiplier layout and simulated to verify that performance specifications are met.

The aim of modern IC-design environments is to minimise the number of iterations required in the design, implementation and verification paths. This should ensure the efficient integration of systems on silicon.

7.4 Digital VLSI design

7.4.1 Introduction

The need for CAD tools in the design and verification paths grows with increasing chip complexity. The different abstraction levels, as discussed in the previous subsection, were created to be able to manage the design complexity at each level.

7.4.2 The design flow

The continuous growth in the number of transistors on a chip is a drive for a greater integration of synthesis and system level design. The increasing complexity of the system level behaviour, combined with an increasing dominance of physical effects of the interconnection (delay, cross-talk, etc.), is a drive for a greater integration of synthesis and physical design.

Figure 7.4 shows a *heterogeneous system on a chip (SOC)*. First, the entire design must be described in a complete specification. For several existing ICs, such a specification consists of several hundreds of textual pages. This design specification must be translated into a high-level behavioural description, which must be executable and/or emulatable.

In many cases, software simulation is too slow and inaccurate to completely verify current complex ICs. Also, the interaction with other system components is not modelled. Logic *emulation* is a way to let designers look before they leap. Emulation allows the creation of a hardware model of a chip. Here, proprietary emulation software is used, which is able to map a design on reprogrammable logic, and which mimics the functional behaviour of the chip. Emulation is usually done in an early stage of the design process and allows more effective *hardware/software co-design*. Once the high-level behavioural description is verified by simulation or emulation, all subsequent levels of design description must be verified against this top-level description. Figure 7.13 shows a general representation of a design flow.

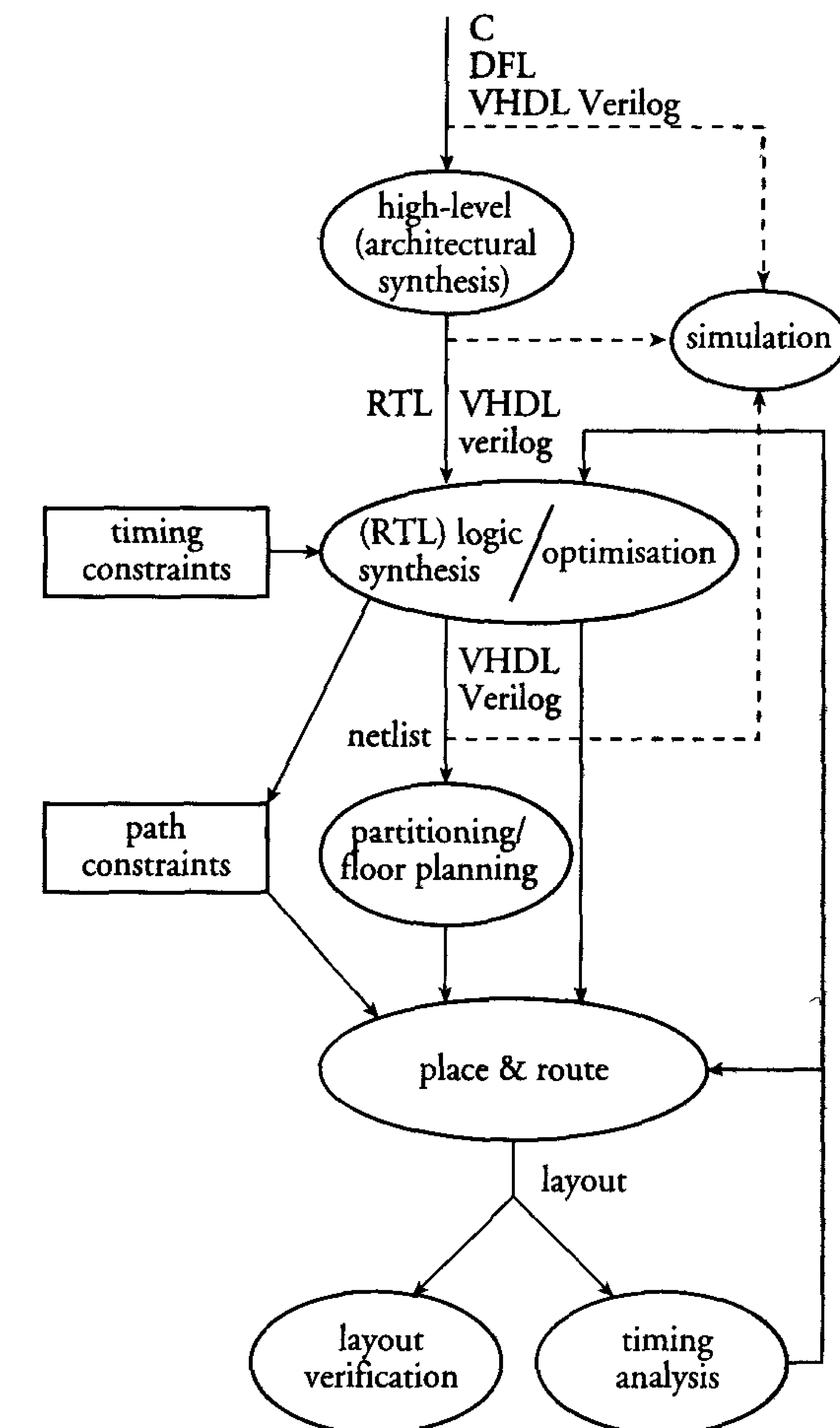


Figure 7.13: *General representation of a design flow*

Synthesis tools automatically translate a description at a higher hierarchy level into a lower level one. These tools are available at several levels of abstraction. High-level synthesis transforms a behavioural description into a sequence of possible parallel operations which must be performed on an IC. Such a behavioural description usually expresses functionality in a high-level computer programming language such as Pascal or C^{++}. The derivation of ordering operations in time is called *scheduling*.

The *allocation* (or *mapping*) process selects the required data-path components. These high-level components include complete signal processor and microprocessor cores, as well as co-processors, ALUs, RAMs and I/O blocks, etc. However, high-level synthesis at system level is

still in the R&D phase and its use is restricted to specific application areas, such as the design of digital signal processor ICs. For telecom and audio processor ICs, there are tools which are different from those that are created and used for the development of video signal processors. Such tools automatically generate hardware descriptions (in VHDL or Verilog) from the system specification.

Current and future *systems on silicon* (figure 7.4) are, and will be, designed by using a wide variety of pre-designed building blocks. This design *reuse* requires that these *Intellectual Property (IP)* parts, such as microcontrollers and microprocessors, can be easily ported from one chip design to another. Such a reuse must be supported by tools. Design reuse will be fuelled by the sharing of cores among companies. In many cases, a Reduced Instruction Set Computer (RISC) microprocessor core (ARM, MIPS, Sparc) is used. If we include the application in an on-chip ROM or other type of memory, this is called *embedded software*.

Synthesis tools must play a key role in integrating such pre-designed building blocks with synthesised glue logic onto one single chip. The most-used type of synthesis is from the RTL level to a netlist of standard cells. Each *system on a chip* can be considered to consist of many registers which store binary data. Data is operated on between these registers. The operations can be described in a *Register-Transfer Language* (RTL). Before the VHDL code is synthesised at this level, the code must be verified by simulation.

At higher functional levels, software (VHDL) simulators are often sufficiently fast. However, in many cases, RTL level simulation is a bottle-neck in the design flow. Besides an increase in the complexity of ICs, longer frame times (as in MPEG video and DAB) must also be simulated. Such simulations may run for several days, resulting in too long iteration times and allowing only limited functional validation of an RTL design.

A *hardware accelerator*, with accompanying software, is a VHDL simulator platform in which the hardware is often implemented as a large multiprocessor system, which is connected to the network or a host system. Gate level descriptions as well as memory modules can be downloaded into a hardware accelerator. However, most non-gate level parts (RTL and test bench) are kept in software. The accelerator hardware speeds up the execution of certain processes (i.e. gates and memory) and the corresponding events. In fact, the accelerator is an integrated part of the simulator and uses the same type of interface.

Generally, the raw performance of a hardware accelerator is less than with *emulation*.

When the RTL description is simulated and proven to be correct, RTL synthesis is used to transform the code (mostly VHDL or Verilog) into an optimised netlist. Actually, the described function or operation at RTL level is *mapped* onto a library of (standard) cells. Synthesis at this level is more mature than high-level synthesis and is widely used. CAD tools are also used for the validation in the IC-design verification path. This verification may include the comparison of design descriptions at two levels of abstraction. Simulation is the most commonly-used *design-verification* method.

As a result of the growing number of transistors on one chip and with the inclusion of analogue circuits or even sensors on the same chip, verification and analysis are becoming serious bottle-necks in achieving a reasonable design turn-around time. Extensive verification is required at each level in the design flow and, in addition, there is a strong need for cross-verification between the different levels. Verification often consumes 20 to 50 percent of the total design time. With increasing clock speed and performance, packaging can be a limiting factor in the overall system performance. Direct attachment of chip-on-board and flip-chip techniques continue to expand to support system performance improvements. Verification tools are therefore needed across the chip boundaries and must also include the total interconnect paths between chips.

7.4.3 Example of synthesis from VHDL description to layout

This paragraph discusses the design steps of the digital potentiometer (see section 7.3.4), starting at the RTL description level (in VHDL) and ending in a standard cell layout. Figure 7.14 shows the RTL-VHDL description of this potentiometer.

```
LIBRARY IEEE;

USE IEEE.std_logic_1164.ALL;
USE IEEE.std_logic_arith.ALL;
USE IEEE.std_logic_unsigned.ALL;

ENTITY potmeter IS

   GENERIC (par_width: natural := 4;
            operand_width : natural := 12);

   PORT (A, B: IN std_logic_vector(operand_width-1 DOWNTO 0);
         K: IN std_logic_vector(par_width-1 DOWNTO 0);
         Z: OUT std_logic_vector(par_width+operand_width-1 DOWNTO 0));

END potmeter;

ARCHITECTURE behaviour OF potmeter IS

BEGIN

   PROCESS (A, B, K)

      CONSTANT K_max: integer := 2**par_width-1;
      VARIABLE K_int: integer;

   BEGIN

      K_int := conv_integer(K);
      Z <= K*A + conv_std_logic_vector(K_max-K_int, par_width) * B;

   END PROCESS;

END behaviour;
```

Figure 7.14: *RTL-VHDL description of potentiometer*

Figure 7.15(a) shows a high abstraction level symbol of this potentiometer, while a behavioural level representation is shown in figure 7.15(b).

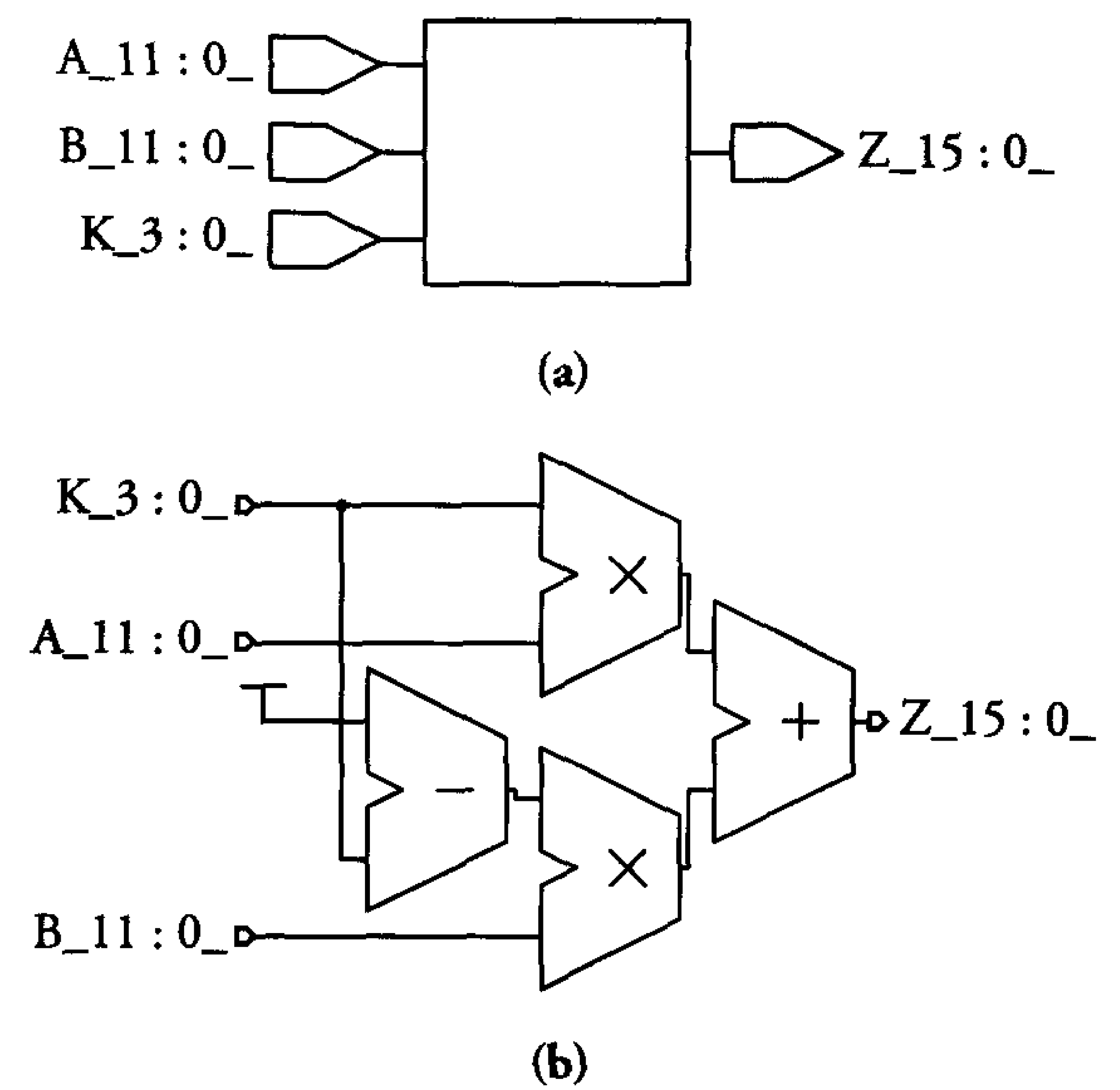


Figure 7.15: *(a) Abstraction level symbol and (b) behavioural level representation of the potentiometer*

After synthesis, without constraints, our potentiometer looks as shown in figure 7.16

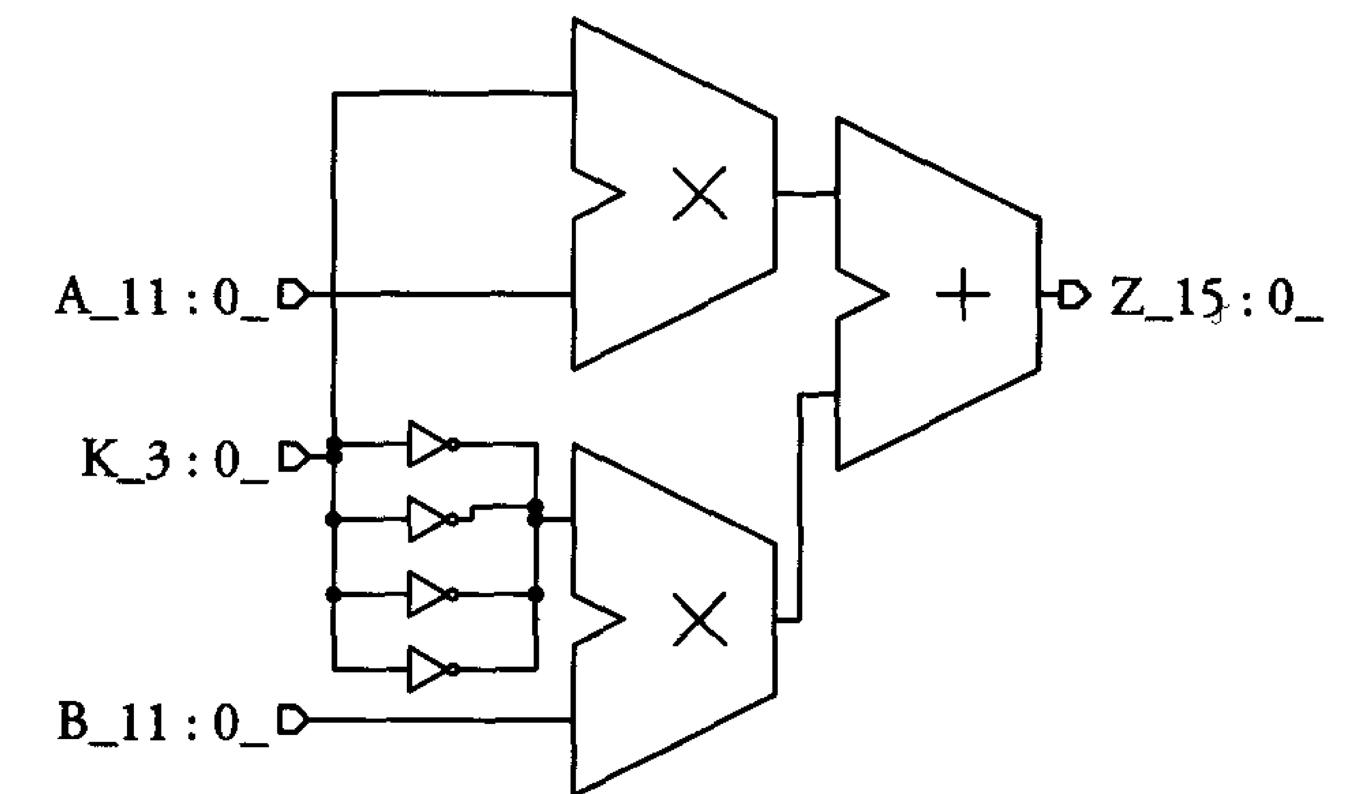


Figure 7.16: *Potentiometer schematic after synthesis with no constraints*

Figure 7.17 shows the multiplier and adder symbolic views after synthesis.

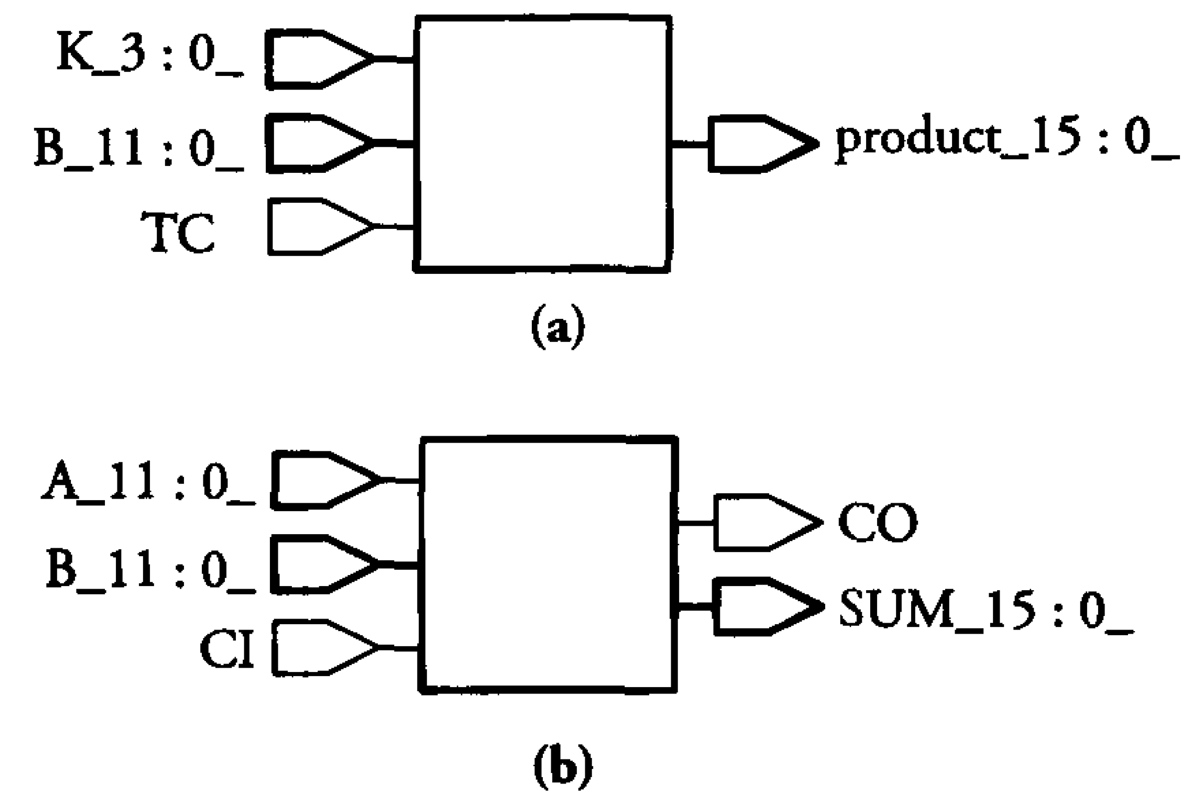


Figure 7.17: *Multiplier and adder symbolic views*

Figure 7.18 shows the schematics of the adder, after synthesis with no constraints.

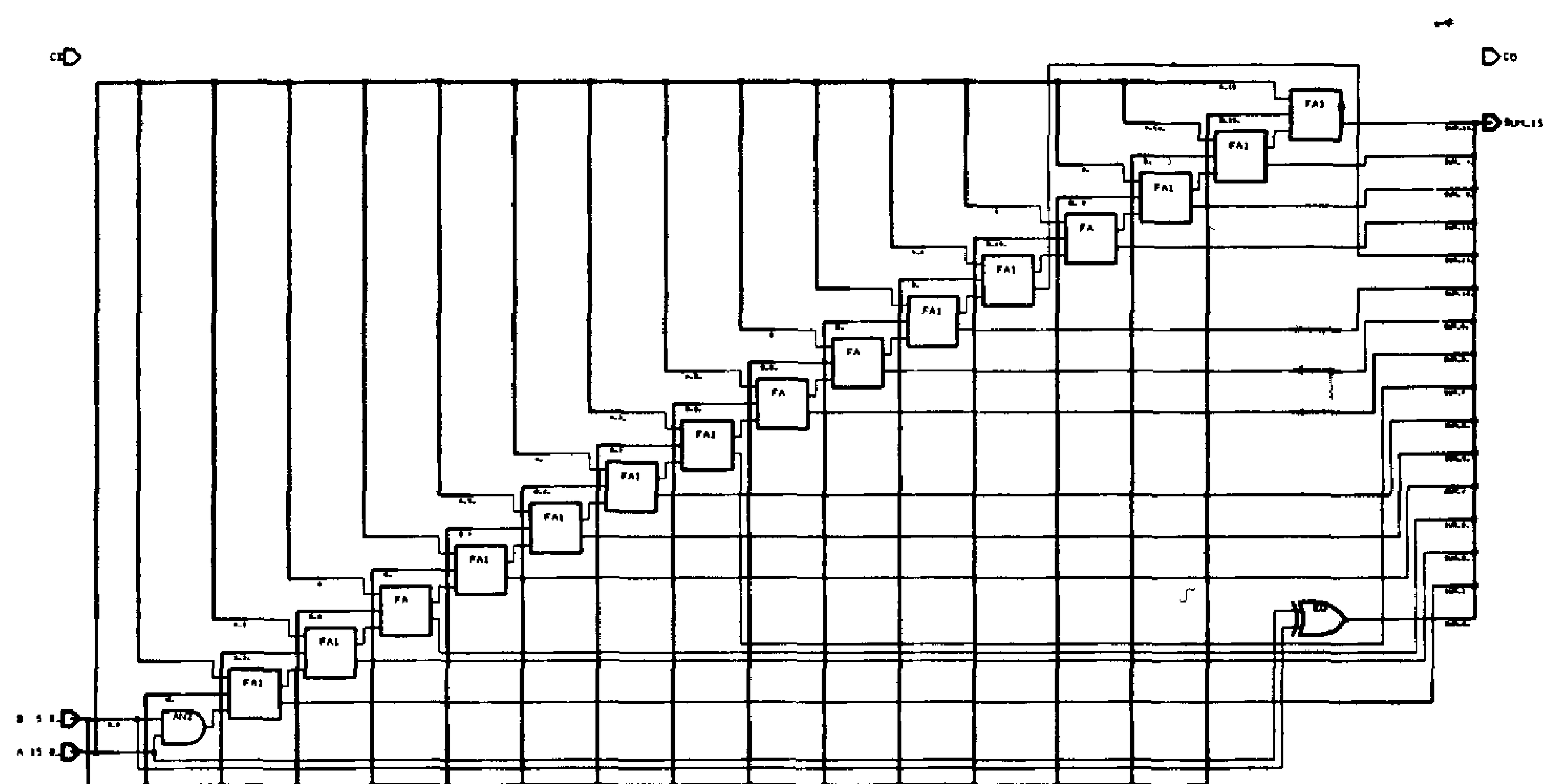


Figure 7.18: *Adder schematics after synthesis with no constraints*

Figure 7.19 shows the schematics of the adder, after synthesis with a timing constraint of 14 ns for the worst-case delay path.

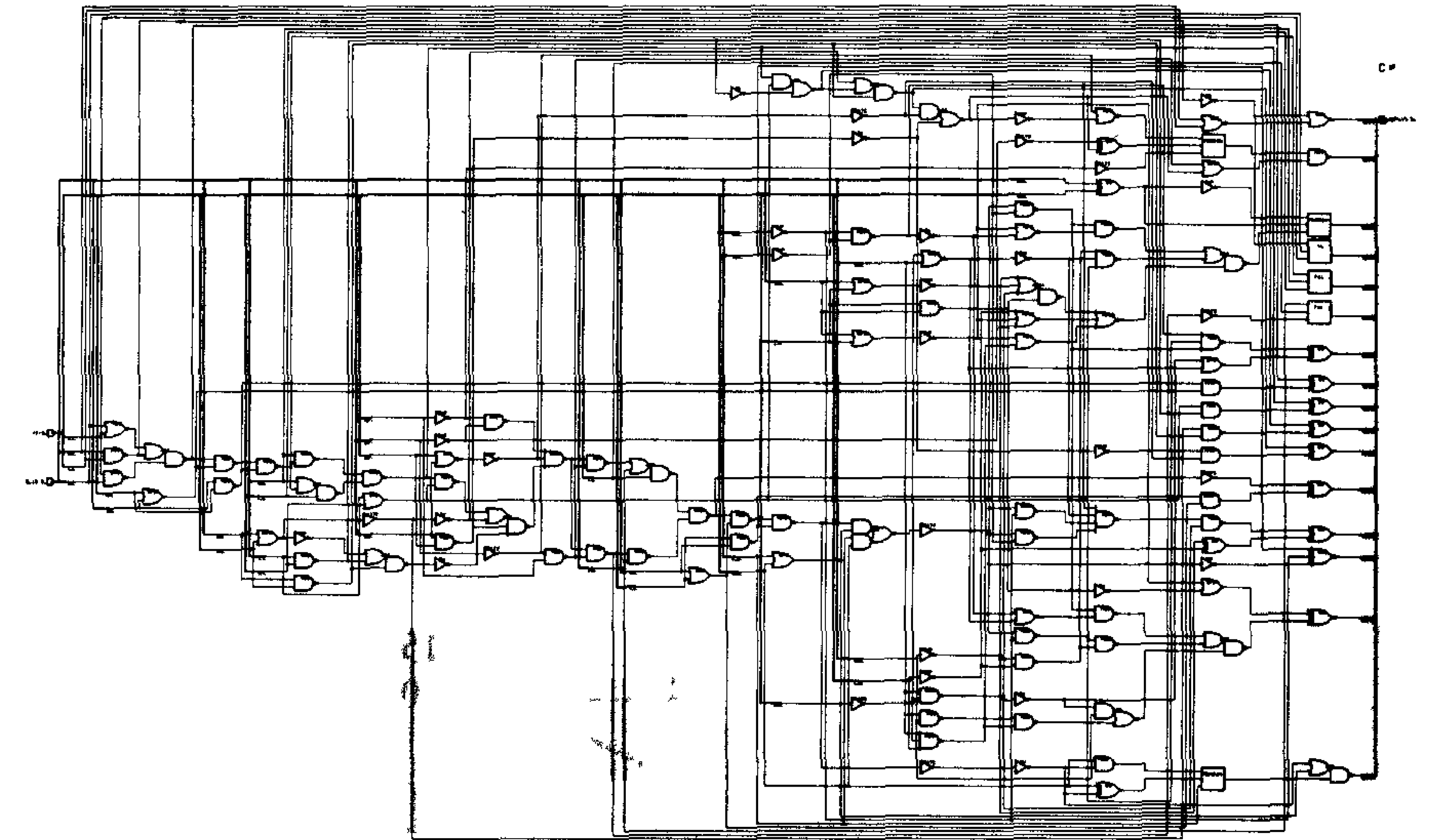


Figure 7.19: *Adder schematics after timing-constraint synthesis*

The additional hardware in figure 7.19 compared to that of figure 7.18 is used to speed up the carry ripple by means of carry look-ahead techniques. Figure 7.20 shows the relation between the delay and the area. The figure clearly shows that reducing the delay by timing constrainted synthesis can be achieved with relatively much additional hardware (area).

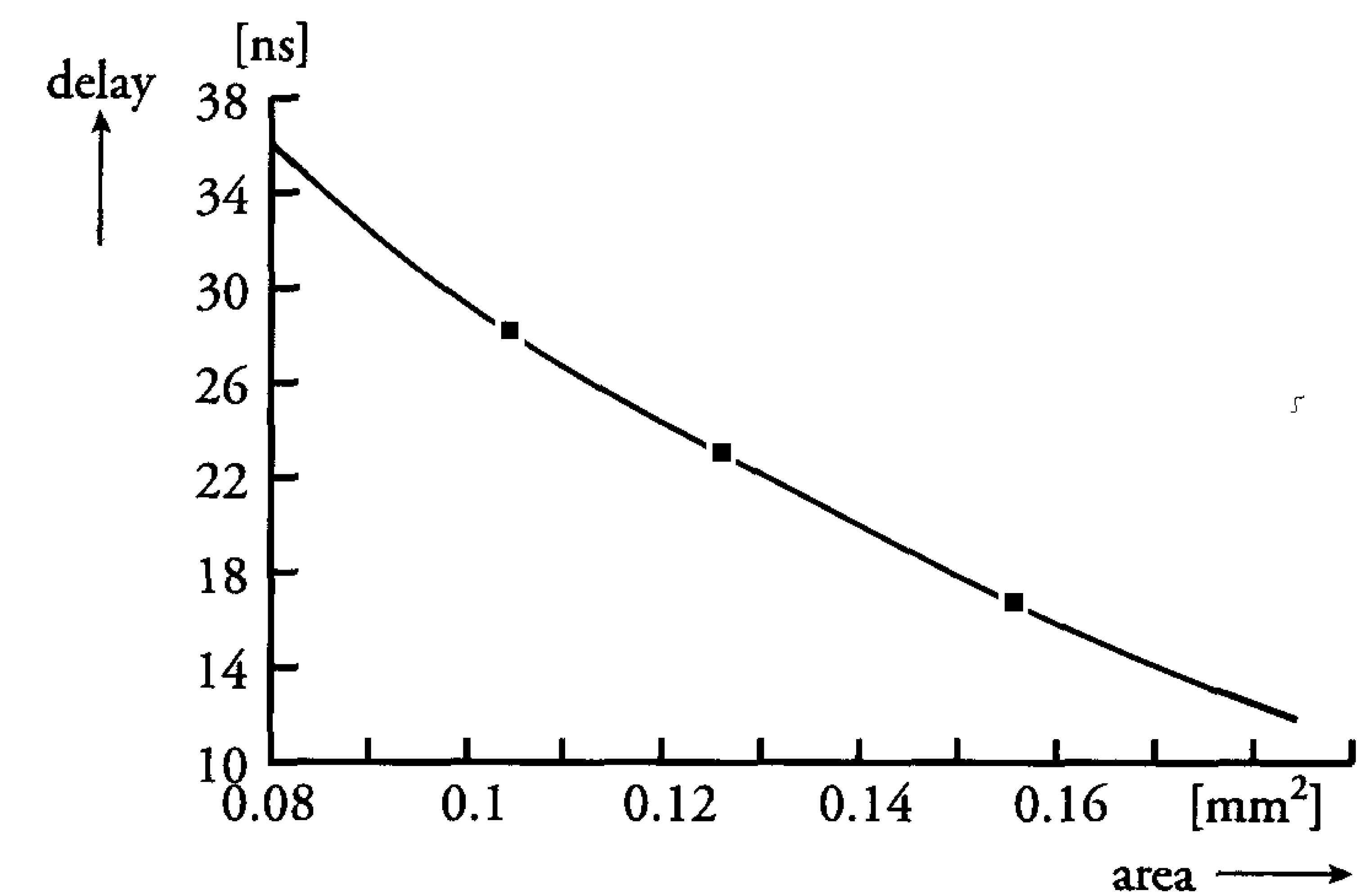


Figure 7.20: *Relation between maximum delay and the amount of hardware (area)*

Figure 7.21 shows a part of the netlist of library cells onto which the potentiometer function has been mapped. The figure shows the different library cells and the nodes to which their inputs and outputs are connected.

```
module potmeter_DW01_add_14_1 ( A, B, CI, SUM, CO );
input  [13:0] A;
input  [13:0] B;
output [13:0] SUM;
input  CI;
output CO;
    wire n52, n53, n54, n55, n56, n57, n58, n59, n60, n61, n62, n63, n64, n65,
        n66, n67, n68, n69, n70, n71, n72, n73, n74, n75, n76, n77, n78, n79,
        n80, n81, n82, n83, n84, n85, n86, n87, n88, n89, n90, n91, n92, n93,
        n94, n95, n96, n97, n98, n99, n100, n101, n102, n103, n104, n105;
    BF1T1 U5 ( .Z(SUM[2]), .A(A[2]) );
    BF1T1 U6 ( .Z(SUM[0]), .A(A[0]) );
    BF1T1 U7 ( .Z(SUM[1]), .A(A[1]) );
    AO6 U8 ( .Z(n52), .A(n53), .B(n54), .C(n55) );
    AO6 U9 ( .Z(SUM[3]), .A(n56), .B(n57), .C(n58) );
    AO32 U10 ( .Z(n59), .A(n60), .B(n61), .C(n62), .D(n63) );
    AO32 U11 ( .Z(n64), .A(n59), .B(n65), .C(n54), .D(n55) );
    NR2 U12 ( .Z(n66), .A(n67), .B(n68) );
    AO6 U13 ( .Z(n69), .A(A[7]), .B(B[7]), .C(n70) );
    NR2 U14 ( .Z(n71), .A(n65), .B(n63) );
    NR2 U15 ( .Z(n72), .A(n73), .B(n74) );
    AN2 U16 ( .Z(n75), .A(n76), .B(n77) );
    EO U17 ( .Z(SUM[9]), .A(n66), .B(n78) );
    EO U18 ( .Z(SUM[8]), .A(n79), .B(n80) );
    EO U19 ( .Z(SUM[6]), .A(n81), .B(n82) );
    EO U20 ( .Z(SUM[5]), .A(n71), .B(n83) );
    MUX21N U21 ( .Z(SUM[13]), .A(B[13]), .B(n84), .S(n85) );
    IV U69 ( .Z(n84), .A(B[13]) );
    IV U70 ( .Z(n105), .A(A[10]) );
    IV U71 ( .Z(n96), .A(B[7]) );
    IV U72 ( .Z(n79), .A(n95) );
endmodule
```

Figure 7.21: *Potentiometer netlist after synthesis with 14 ns timing constraints*

After the use of place and route tools, a standard cell design of the potentiometer is created, see figure 7.22 for the result.

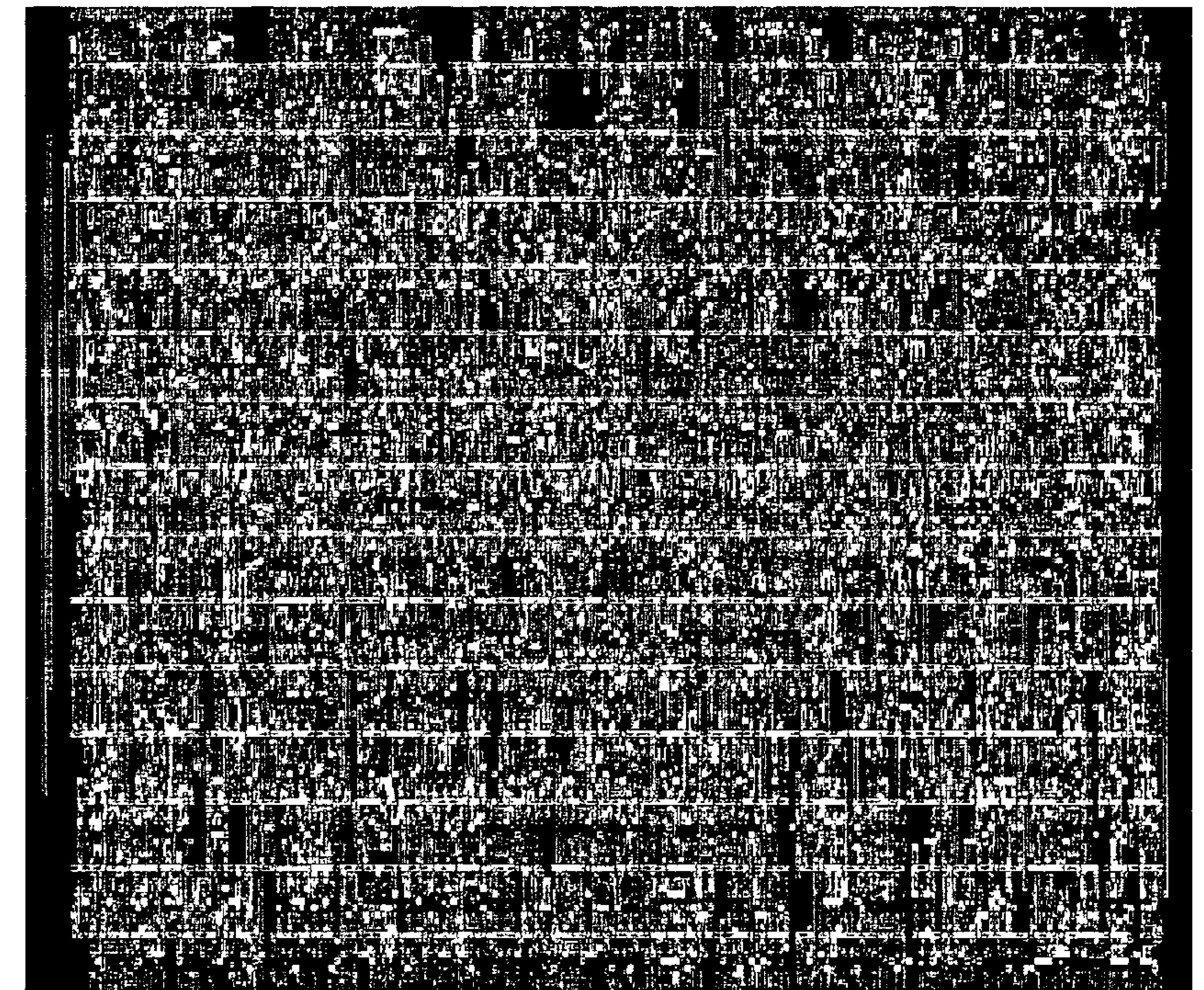


Figure 7.22: *Standard cell implementation of potentiometer*

Although the synthesis process uses tools which automatically generate a next level of description, this process is controlled by the designer. An excellent design is the result of the combination of an excellent tool and a designer with excellent skills in both control of the tools and knowledge of IC design.

7.5 The use of ASICs

The growth in the ASIC business is primarily the result of the increasing number of application areas and of the general increase in the use of ICs. ASICs often provide the only solution to problems attributed to speed and/or space requirements. Another incentive for the use of ASICs is the degree of concealment which they afford. This concealment poses extra difficulties to competitors interested in design duplication.

ASICs make it reasonably easy to add new functionality to an existing system without an extensive system redesign. In addition, the increased integration of system parts associated with the use of ASICs has the following advantages:

- Reduced physical size of the system
- Reduced system maintenance costs
- Reduced manufacturing costs
- Improved system reliability
- Increased system functionality
- Reduced power consumption.

The advantages afforded by ASICs can have a positive influence on the functionality/price ratio of products and have led to the replacement of standard ICs in many application areas. However, there are also disadvantages associated with the use of ASICs. These include the following:

- The costs of realising an ASIC are quite substantial and less predictable than those associated with standard ICs.
- Unlike standard products, ASICs are not readily available from a diverse number of suppliers. Inaccurate specifications or errors in the design process may cause delays in ASIC turn-around time and result in additional *non-recurring engineering* (NRE) costs. These are costs incurred prior to production. Typical NRE costs include the cost of:
  - Training and use of design facilities
  - Support during simulation
  - Placement and routing tools
  - Mask manufacturing (where applicable)
  - Test development
  - The delivery of samples.

  Furthermore, standard products are always well characterised and meet guaranteed quality levels. Moreover, small adjustments to a system comprising standard products can be implemented quickly and cheaply.

The advantages and disadvantages associated with the use of ASICs depend on the application area and on the required ASIC type and quantities. Improved design methods and production techniques combined with better relationships between ASIC customers and manufacturers will have a considerable influence on the transition from the use of standard products to ASICs.

7.6 Silicon realisation of VLSI and ASICs

7.6.1 Introduction

In addition to the need for computer programs for the synthesis and verification of complex ICs, CAD tools are also required for the automatic or semi-automatic generation of layouts. The design of INTEL's Pentium microprocessors, for example, took several hundreds of man-years. The same holds for the IBM PowerPC. Figure 7.23 shows a photograph of this processor. In fact, the increased use of CAD tools in recent years has very often merely facilitated the integration of increasingly complex systems without contributing to a significant man-year reduction in design time. This situation is only acceptable for very complex high-performance ICs such as a new generation of microprocessors. Less complex ICs, such as ASICs, require fast and effective design and layout tools. Clearly, the need for a fast design and layout process increases as the lifetimes of new ICs become shorter. The lifetime of a new generation of ICs for compact disc players, for instance, is about one to two years. This means that the design process may take only a couple of months. Each layout design must be preceded by a thorough floor plan study. This must ensure that the envisaged layout will not prove too large for a single chip implementation in the final design phase. A floor plan study can take considerable time and only leads to a definite floor plan after an iterative trial-and-error process. Layouts of some parts of the chip may be required during the floor plan study. Although we distinguish between the different ASIC categories of custom ICs, semi-custom ICs and PLDs in this book, the differences are rapidly diminishing as a result of the pace at which improvements in IC technologies are realised. PLDs are moving towards gate arrays, gate arrays are moving towards cell-based designs and cell-based designs may use sea-of-gates structures such as embedded arrays to implement the glue logic as well as for mapping of cores onto such arrays. Each category uses the best features of the others.

The choice of implementation is determined by the required development time, production volume and performance. Table 7.2 summarises the performance of various *layout implementation forms*. This table is only valid in general terms.

Table 7.2: Comparison of performance of different layout implementation forms

Implementation form	**Performance**	
	speed	area
handcrafted layout	++++	++++
bit slice	−+++	−+++
cell base	−−++	−−++
(sea-of-gates) gate array	−−−+	−−−+
PLD (FPGAs and CPLDs)	−−−+	−−−+

The different layout implementation forms are discussed separately in the next subsections.

7.6.2 Handcrafted layout implementation

A *handcrafted layout* is characterised by a manual definition of the logic and wiring. This definition must account for all relevant layout design rules for the envisaged technology. The design rules of modern technologies are far more numerous and complex than those used in the simple initial nMOS process. However, various CAD tools have emerged which ease the task of creating a handcrafted layout. These include interactive computer graphic editors (or *polygon pushers*), compactors and *design-rule-check* (DRC) programs.

An example of a handcrafted layout is illustrated in figure 7.24. Such an implementation yields considerable local optimisation. However, the required intensive design effort is only justified in MSI circuits and limited parts of VLSI circuits. The use of handcrafted layout is generally restricted to the design of basic cells. These may subsequently be used in standard-cell libraries, module generators and bit-slice layouts, etc.

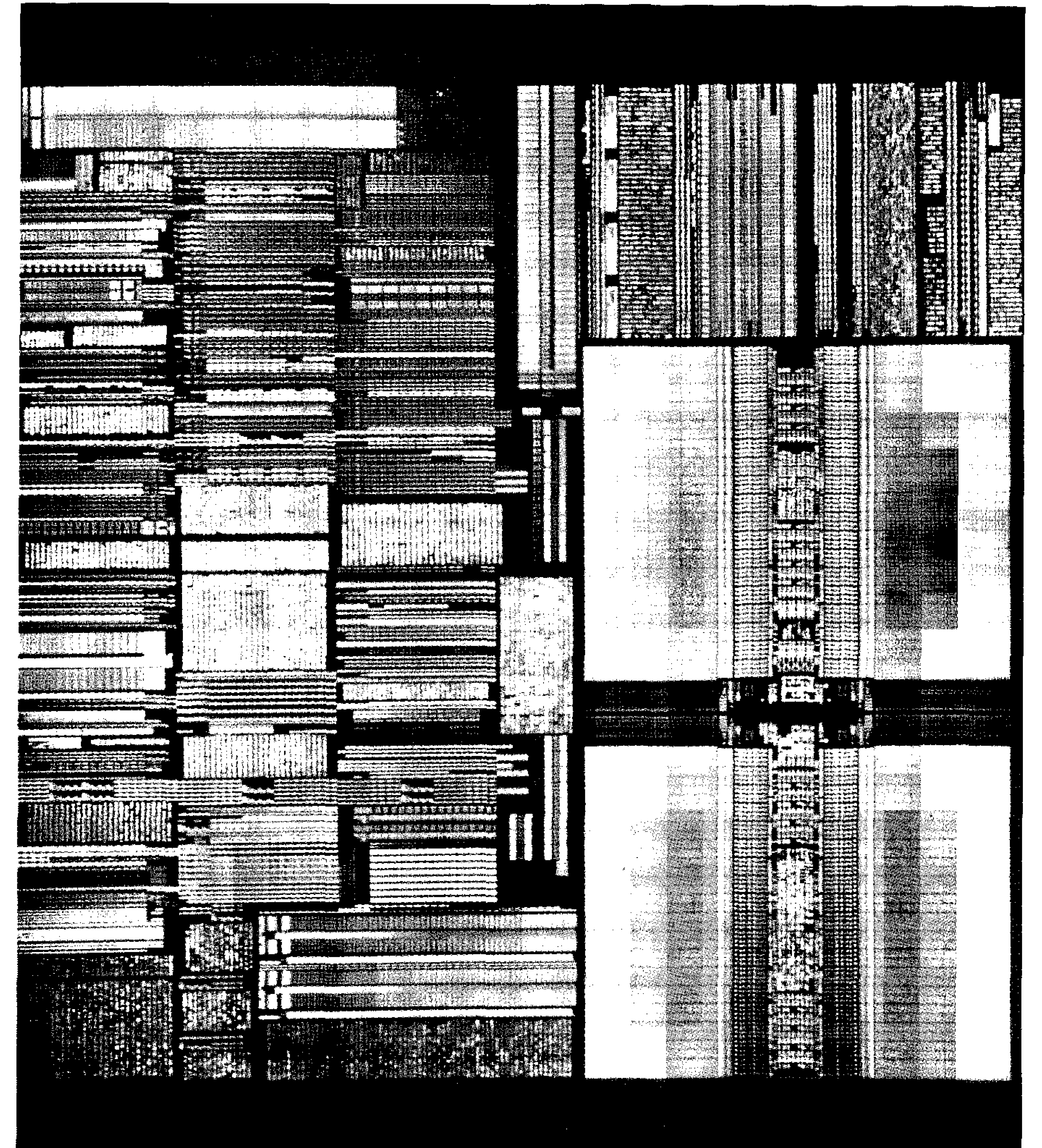


Figure 7.23: *The IBM power PC processor (Photo: IBM)*

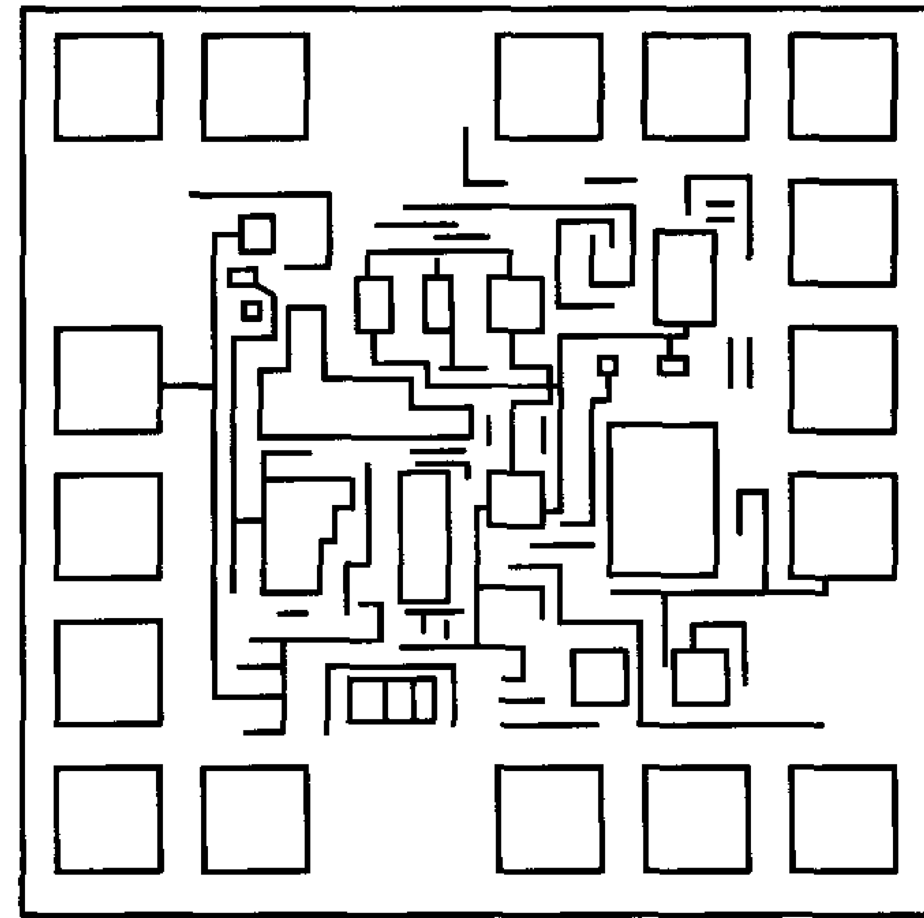


Figure 7.24: *Typical contours of a handcrafted layout*

7.6.3 Bit-slice layout implementation

A bit-slice layout is an assembly of parallel single-bit data paths. The implementation of a bit-slice layout of a signal processor, for example, requires the design of a circuit layout for just one bit. This bit slice is subsequently duplicated as many times as required by the word length of the processor. Each bit slice may comprise one or more vertically-arranged cells. The interconnection wires in a bit slice run over the cells with control lines perpendicular to data lines. Many available CAD tools facilitate the efficient assembly of bit-slice layout architectures. The bit-slice design style is characterised by an array-like structure which yields a reasonable packing density. Figure 7.25 illustrates an example of a bit-slice layout architecture. A bit-slice section is also indicated in the chip photograph in figure 7.45.

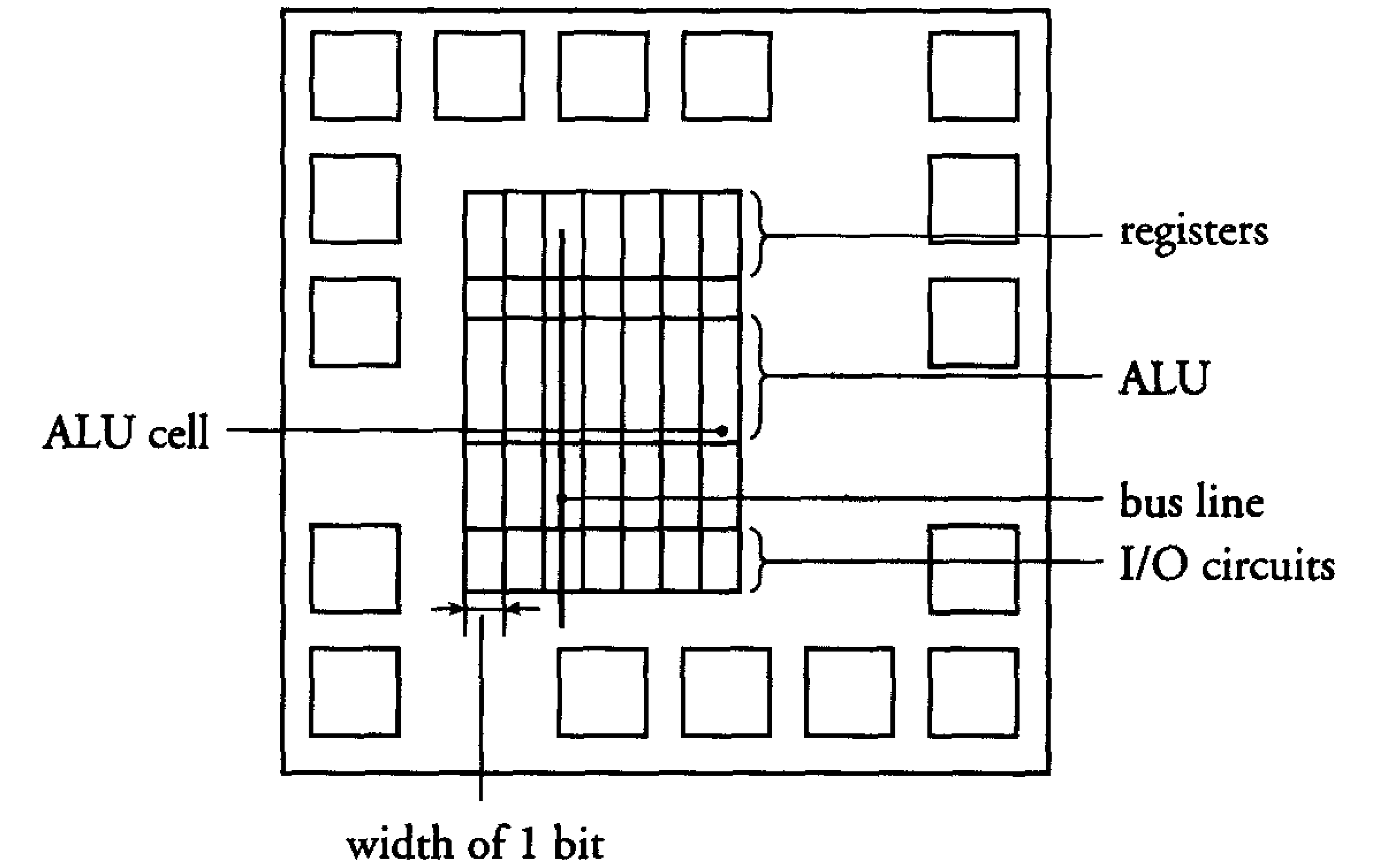


Figure 7.25: *Basic bit-slice layout*

7.6.4 ROM, PAL and PLA layout implementations

In addition to serving as a memory, a ROM can also be used to implement logic functions. An example is shown in figure 7.26.

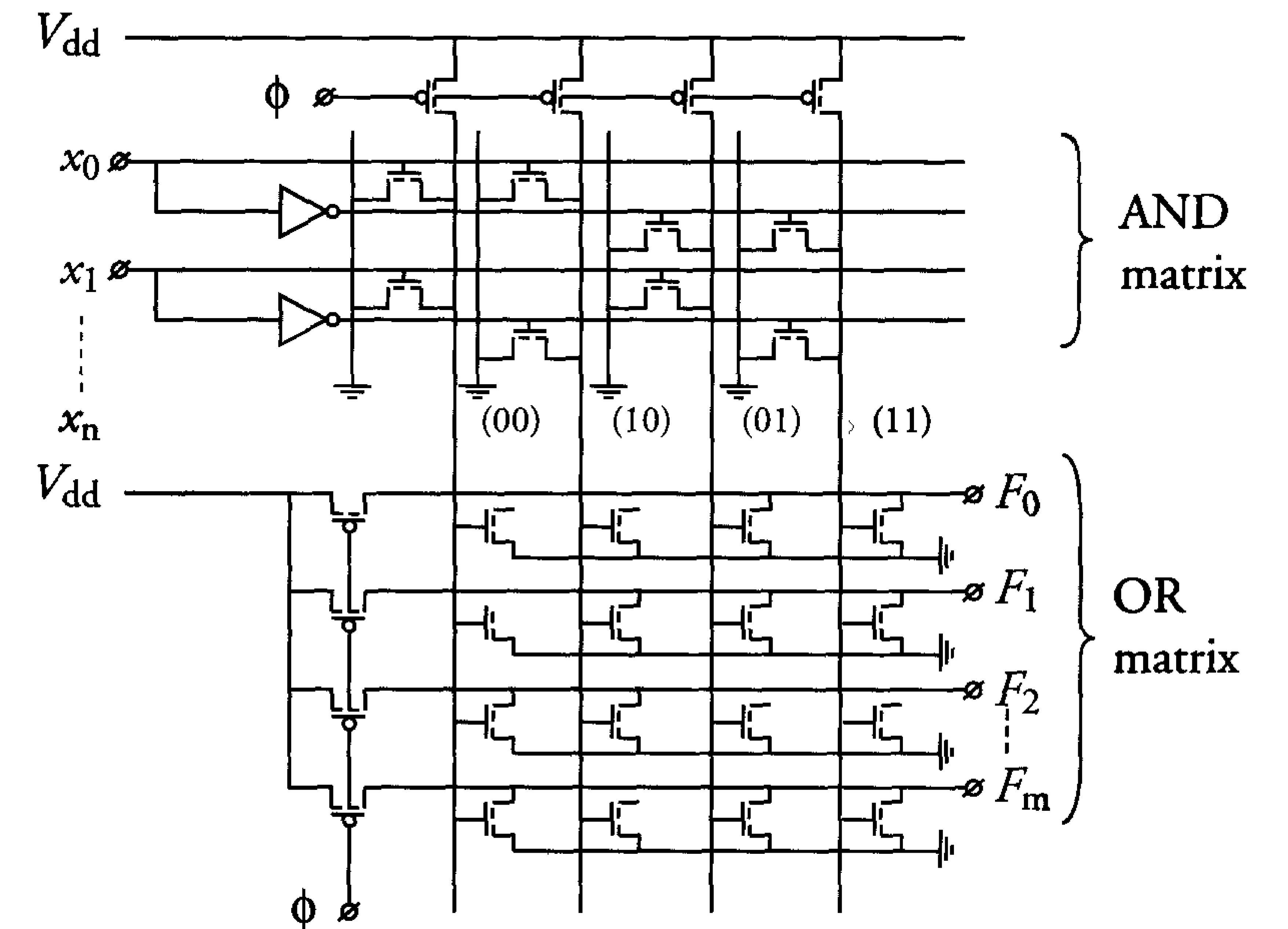


Figure 7.26: *Logic functions realised with a ROM*

Only one vertical line in this ROM will be 'high' for each combination of address inputs $x_n \cdots x_0$. This vertical line drives the gates of $m + 1$ transistors in the OR matrix. Outputs F_j, which are connected to the drains of these transistors, will be 'low'. If, for example, the address inputs are given by x_0x_1=01, then the second column line will be 'high'. A 'low' will then be present on outputs F_1 and F_2. The information stored in the ROM in figure 7.26 is determined by the presence or absence of connections between MOS transistor drains and the output lines. In this way, the structure of a ROM can easily be used to realise logic functions. Table 7.3 shows a possible truth table, which could be implemented with the ROM in figure 7.26.

Table 7.3: Example of a truth table implemented with the ROM in figure 7.26.

x_n	-	-	-	x_1	x_0		F_m	-	-	-	F_1	F_0
0	-	-	-	0	0		0	-	-	-	1	1
0	-	-	-	0	1		1	-	-	-	0	1
0	-	-	-	1	0		0	-	-	-	0	0
0	-	-	-	1	1		0	-	-	-	0	0

Clearly, the set of logic functions that can be realised in a ROM is merely limited by the number of output and address bits. The regular array structure of a ROM leads to a larger transistor density per unit of chip area than for random logic. A large number of logic functions could, however, require an excessively large ROM while the use of a ROM could prove inefficient for a small number of logic functions. In general, a ROM implementation is usually only cheaper than random logic when large volumes are involved.

Unfortunately, there are no easy systematic design procedures for the implementation of logic functions in ROM. Other disadvantages are as follows:

- Lower operating frequency for the circuit
- The information in a ROM can only be stored during manufacturing
- Increasing the number of input signals by one causes the width of the ROM to double

- A high transistor density does not necessarily imply an efficient use of the transistors.

It is clear from figure 7.26 that the vertical column lines in a ROM represent the *product terms* formed by the address inputs x_i. These product terms comprise all of the logic AND combinations of the address inputs and their inverses. Only the OR matrix of a ROM can be programmed.

Figure 7.27 illustrates the basic structure of a *programmable logic array* (PLA). Its structure is similar to that of a ROM and consists of an *AND matrix* and an *OR matrix*. In a PLA, however, both matrices can be programmed and only the required product terms in the logic functions are implemented. It is therefore more efficient in terms of area than a ROM. Area requirements are usually further reduced by minimising the number of product terms before generating the PLA layout pattern.

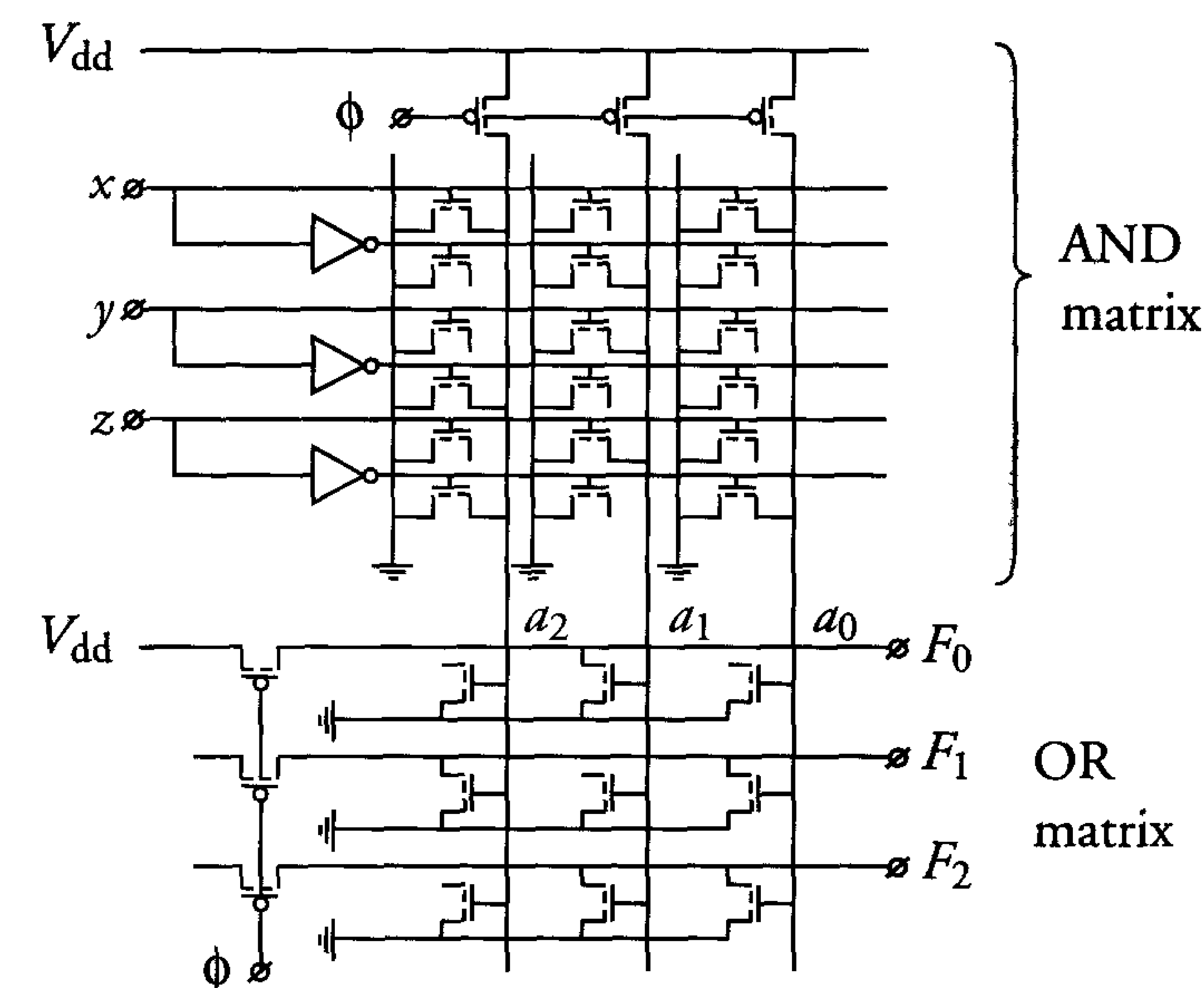


Figure 7.27: *Basic PLA structure*

The logic functions implemented in the PLA in figure 7.27 are determined as follows: a_0 is 'high' when x and z are low, i.e. $a_0 = \overline{x}\ z$. Similarly, $a_1 = x\ \overline{y}\ \overline{z}$ and $a_2 = \overline{x}\ y\ z$.

The outputs are therefore expressed as follows:

$$F_0 = \overline{a_1} = \overline{x\ \overline{y}\ \overline{z}}$$

$$F_1 = \overline{a_0 + a_2} = \overline{\overline{x}\ z + \overline{x}\ y\ z}$$

$$F_2 = \overline{a_0 + a_1} = \overline{\overline{x}\ z + x\ \overline{y}\ \overline{z}}$$

A PLA can be used to implement any combinatorial network comprising AND gates and OR gates. In general, the complexity of a PLA is characterised by $(A+C) \times B$, where A is the number of inputs, B is the total number of product terms, i.e. the number of inputs for each OR gate, and C is the number of outputs, i.e. the number of available logic functions.

Sequential networks can also be implemented with PLAs. This, of course, requires the addition of memory elements. A PLA can be a stand-alone chip or an integral part of another chip such as a microprocessor or a signal processor. PLAs are frequently used to realise the logic to decode *microcode instructions* for functional blocks such as memories, multipliers, registers and ALUs. Several available CAD tools enable a fast mapping of logic functions onto PLAs. As a result of the improvements in cell-based designs, ROM and PLA implementations are becoming less and less popular in VLSI designs. Another realisation form is the *Programmable Array Logic (PAL)*. In this concept, only the AND plane is programmable and the OR plane is fixed.

Table 7.4 summarises the programmability of planes (AND, OR) in the ROM, PAL and PLA devices. Programmable techniques include fuses (early and smaller devices), floating gate transistors ((E)EPROM) and flash devices. In some cases, a ROM (PLA) block is still used in a custom design; the programming is done by a mask. These are then called mask-programmable ROMs (PLAs).

Table 7.4: Programmability of AND and OR planes in ROM, PAL or PLA devices

Device	**Programmable**	
	AND-plane	OR-plane
ROM	no	yes
PAL	yes	no
PLA	yes	yes

7.6.5 Cell-based layout implementation

Figure 7.28 shows a basic layout diagram of a chip realised with *standard cells.* In this design style, an RTL description of the circuit is synthesized and mapped onto a number of standard cells which are available in a library. The resulting netlist normally contains no hierarchy. The standard-cell library usually consists of a number of different types of logic gates, which are all of equal height.

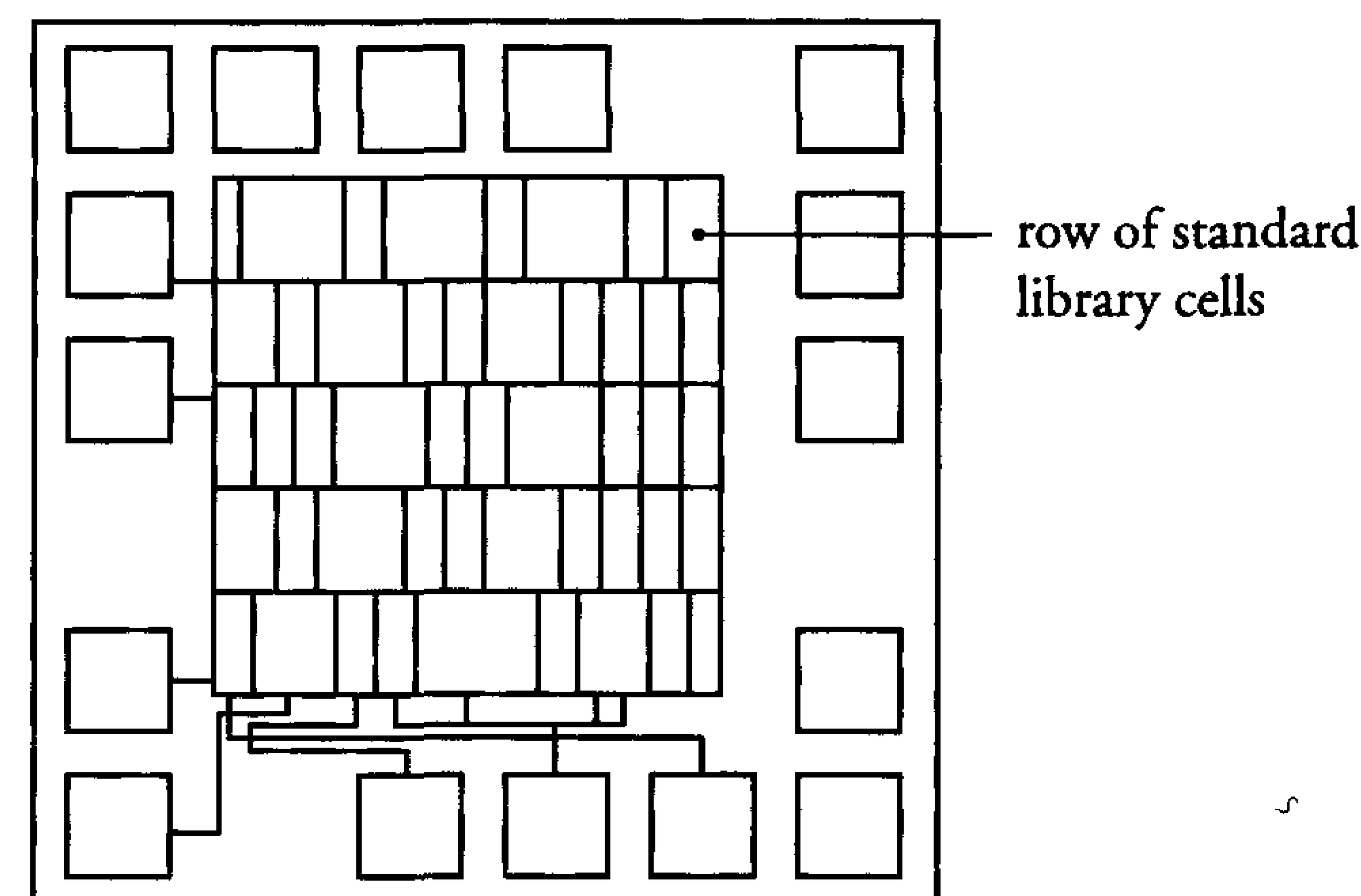


Figure 7.28: *Basic standard-cell layout*

The standard-cell layout method is supported by mature CAD tools for *placement* and *routing.* Routing is done at a fixed grid across the logic gates. The supply and sometimes also the clock wiring are specially structured to minimise their resistance and are usually an integral

part of the standard cell design approach. Modern standard-cell design environments facilitate the inclusion of larger user-defined cells in the library. These *blocks*, *macros* or *cores* may include multipliers, RAMs, signal processor cores, microprocessor cores, etc.

During the late eighties, extra attention was paid to advanced circuit test methods. These include *Boundary Scan test* and *self-test* techniques, see section 10.2. The Boundary Scan technique uses a sequential chain of memory elements to allow access to a large number of locations on an IC or on a printed circuit board. The self-test technique requires the addition of dedicated logic to an existing design. This logic generates the stimuli required to test the design and checks the responses. The result is a circuit or an IC which is effectively capable of testing itself.

The previously-discussed cell-based designs may include standard cells, macro cells, embedded memory blocks and IP cores, etc. A rather new development in cell-based designs is the inclusion of *embedded arrays*. In most cell-based designs that include an embedded array, all masks are customised, as in the cell-based designs. Embedded arrays combine a gate array-like structure and large cells such as microprocessor cores, memories and I/O functions. Cores can either be mapped onto the sea-of-gates array (see next subsection) or can be implemented as a separate block. Figure 7.29 shows the architecture of an embedded array ASIC.

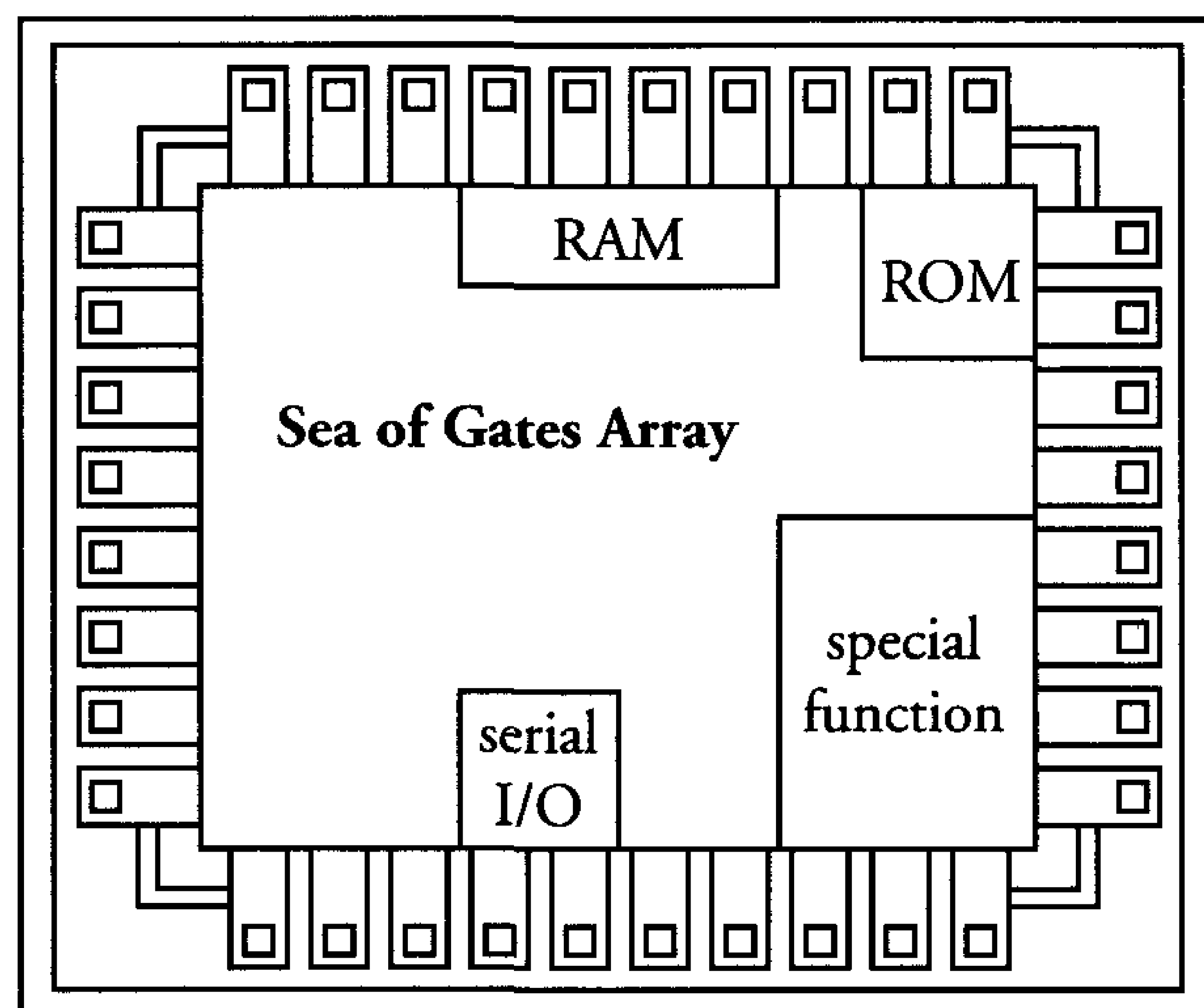


Figure 7.29: *Architecture of an embedded array ASIC (Source: ICE)*

Currently, the cores and other cells in the embedded array provide a higher level of integration than a pure gate array implementation. A faster turn-around time of an embedded-array ASIC and the relatively cheap possibility of making multiple customer variations of the same basic design are among its major advantages.

Quite early in the design process, it is known how much memory, which core(s), what other mega cells and which I/Os are needed to realise the ASIC. Although a design is not yet completely finished, the base layers (up to the interconnect layers) can already be processed in parallel with the completion of the design. Therefore, the embedded sea-of-gates array contains more transistors than are expected to realise the total chip, including the glue logic. Even small last-minute design changes are allowed, as long as they fit into the remaining sea-of-gates cells. When the design and verification is completed, the final interconnect masks are generated and processed. Thus, compared to a "conventional" cell-based version of the same design, the time to market is reduced. After the completion of the first samples, several wafers can remain unfinished to support fast redesigns, if necessary. Such a redesign allows complete resynthesis of a core again, as long as there are enough sea-of-gates cells available. Embedded arrays therefore allow rapid design iterations. Embedded arrays are gaining popularity in PLD design as well. Advanced PLDs include relatively small embedded arrays to improve flexibility and speed of dedicated parts of the design (system).

7.6.6 Gate array layout implementation

Gate arrays are also referred to as mask-programmable gate arrays. A conventional *gate array* contained thousands of logic gates, located at fixed positions. The layout could, for example, contain 10,000 3-input NAND gates. The implementation of a desired function on a gate array is called *customisation* and comprises the interconnection of the logic gates. The interconnections were located in dedicated *routing channels*, which were situated between rows of logic gates. In these conventional channelled gate arrays, the routing was often implemented in two metal layers.

This type of gate array is depicted in figure 7.30a. The channels are essential for interconnecting the cells when production processes with one or even two metal layers are involved.

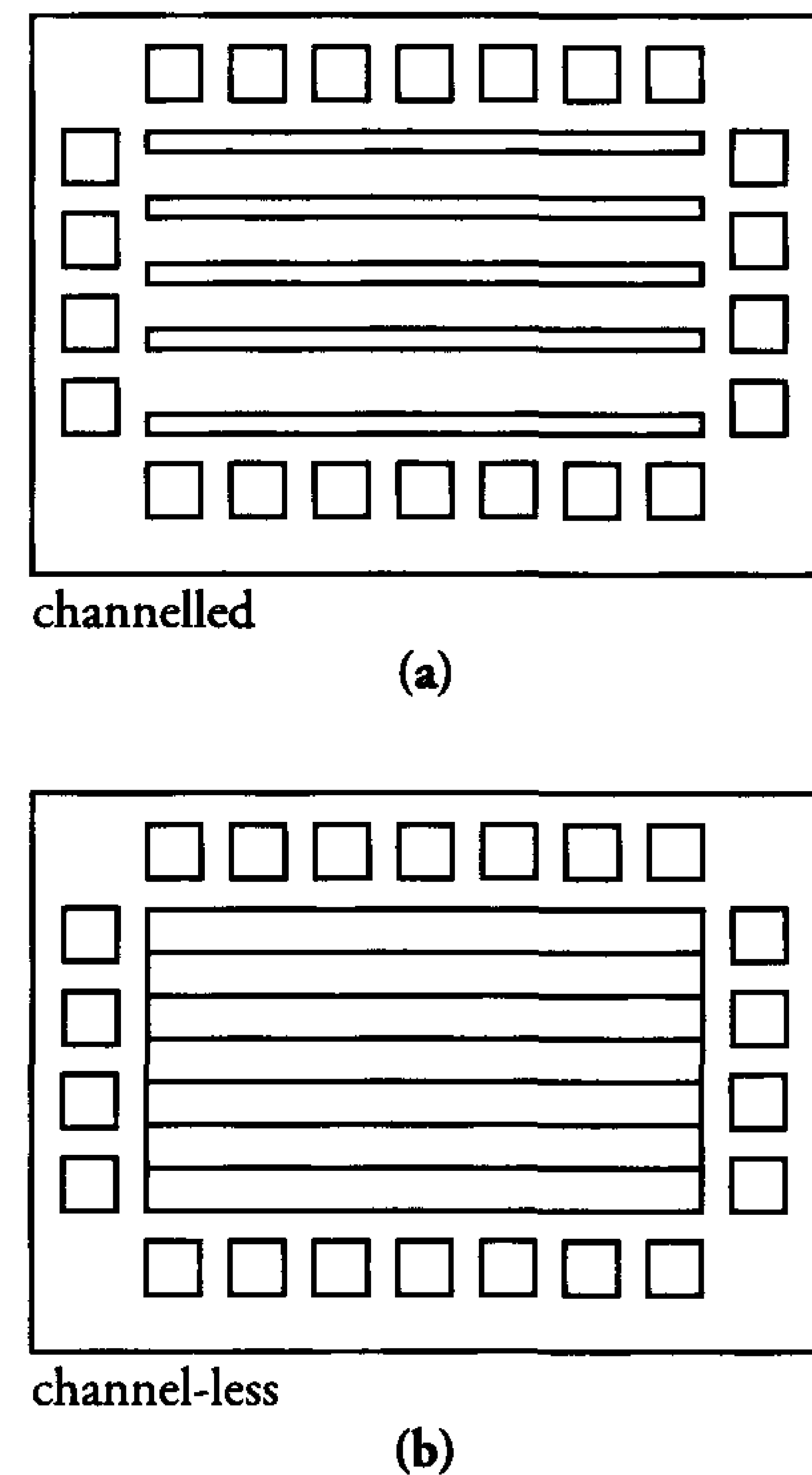


Figure 7.30: *Floor plan for (a) conventional and (b) channel-less gate arrays*

In a conventional gate array, the ratio between the available cell and routing channel areas was fixed. Obviously, the actual ratio between the areas used was dependent on the type of circuit. In practice, the available area is rarely optimally used. This feature is especially important for larger circuits. Furthermore, larger circuits require more complex interconnections and this increases the density in routing channels. The *channel-less gate array* architecture was therefore introduced. Other names encountered in literature for this architecture include: *high-density gate array (HDGA)*, *channel-free gate array*, *sea-of-gates*, *sea-of-transistors* and *gate forest*.

Figure 7.30b shows the floor plan for a channel-less gate array. It consists of an array of transistors or cells. It does not contain any specially reserved routing channels. Modern HDGAs comprise an array of *master cells*, which consist of between four and ten transistors. In some cases, the master cells are designed to accommodate optimum implementations of static RAMs, ROMs or other special circuits. A given

memory or logic function is implemented by creating suitable contact and interconnection patterns in three or more metal layers. The master cells in an HDGA can be separated by *field oxide isolation*, which is created by using the LOCOS or STI techniques described in chapter 3. An example of such an HDGA master-cell structure is shown in figure 7.31, which also shows an example of an HDGA floor plan.

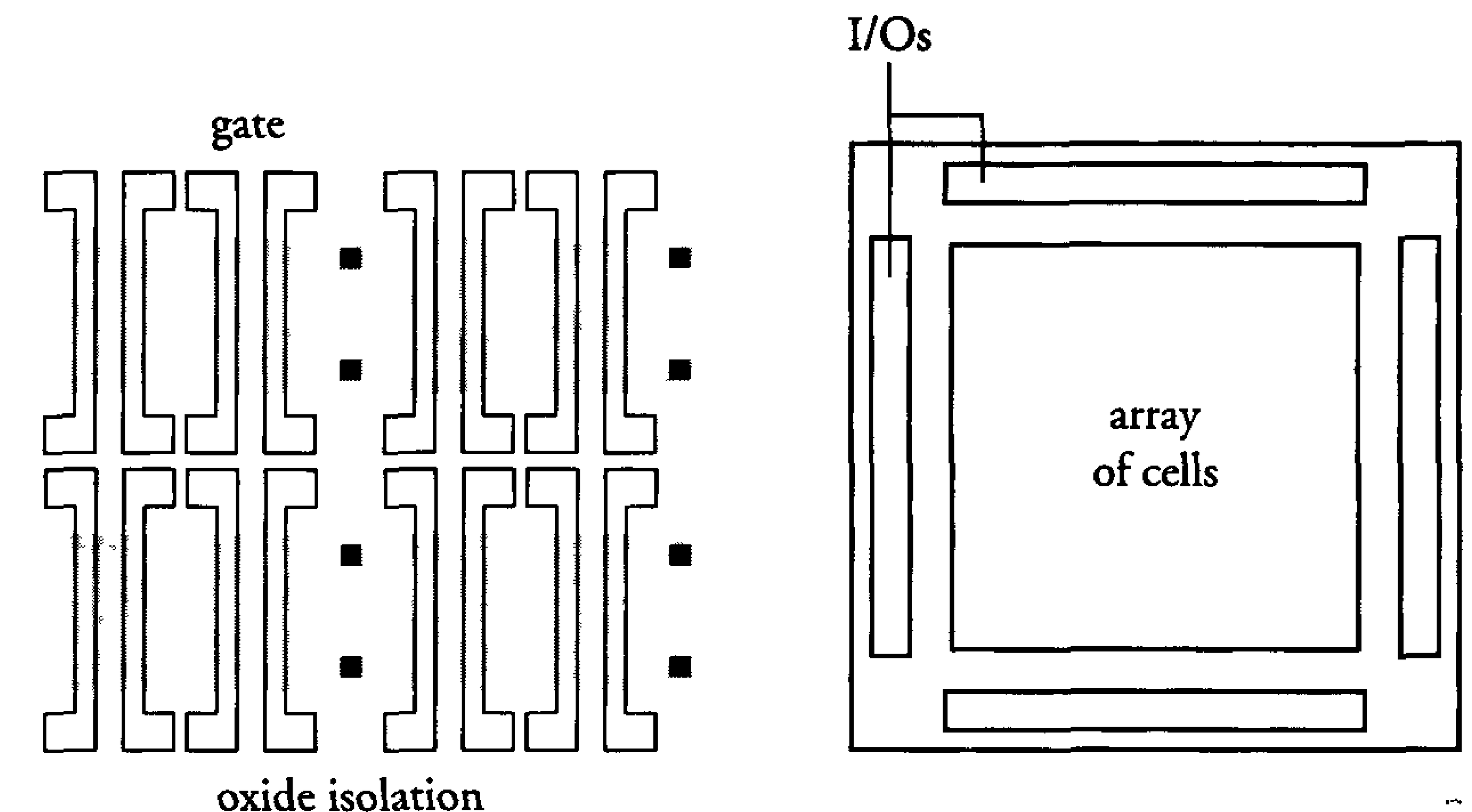


Figure 7.31: *An example of an HDGA master-cell structure and floor plan*

Figure 7.32 shows a section of a sea-of-transistors array, which comprises a row of pMOS and nMOS transistors. The complete array is created by copying the section several times in the horizontal and vertical directions. These HDGAs are also often called *continuous arrays* or uncommitted arrays. The rows are not separated by routing channels and the floor plan is therefore the same as shown in figure 7.30b. These HDGA architectures facilitate the implementation of large VLSI circuits on a single gate array when three or more metal layers are used. The logic and memory functions are again implemented by applying interconnection and contact hole patterns.

The various logic gates and memory cells in a sea-of-transistors HDGA are separated by using the *gate-isolation technique* illustrated in figure 7.32.

The layout in the figure is a D-type flip-flop, based on the logic diagram shown. The gate-isolation technique uses pMOS and nMOS isolation transistors, which are permanently switched off by connecting them

to supply and ground, respectively. This technique obviously requires both an nMOS and a pMOS isolation transistor between neighbouring logic gates [12].

The NRE costs of these devices depend on circuit complexity and are in the order of 25 k$-100 k$. Small transistors placed in parallel with larger transistors facilitate the integration of logic cells with RAMs, ROMs and PLAs in some of these HDGA architectures [13].

The design methods used for gate arrays are becoming increasingly similar to those used for cell-based design. This trend facilitates the integration of scan-test techniques in gate array design. As a result of the increasing number of available cells, the software for gate array programming resembles that of cell-based designs. Also, the availability of complete cores that allow reuse (IP) are becoming available to gate array implementation.

Off-the-shelf families of gate arrays are available and include the source and drain implants. Customisation therefore only requires the processing of several contact and metal masks. This facilitates a short *turn-around time* in processing and renders gate arrays suitable for fast prototyping.

Modern gate array publications include advanced low-power schemes and technologies (SIMOX). For high speed gate arrays, gate delays (3-input NOR with a fan-out of two) below 50 ps have currently been reported. The complexity of advanced gate arrays has exceeded several million gates.

7.6.7 Programmable Logic Devices (PLDs)

A PLD is a Programmable Logic Device, which can be programmed by fuses, anti-fuses or memory-based circuits. Another name currently also used for these devices is Field Programmable Device (FPD). The first user-programmable device that could implement logic was the programmable read-only memory (PROM), in which address lines serve as logic inputs and data lines as output (see also sections 6.4.4 and 7.6.4). PLD technology has moved from purely bipolar technology, with a simple fuse-blowing mechanism, to complex architectures using antifuse, (E)EPROM, flash or SRAM programmability.

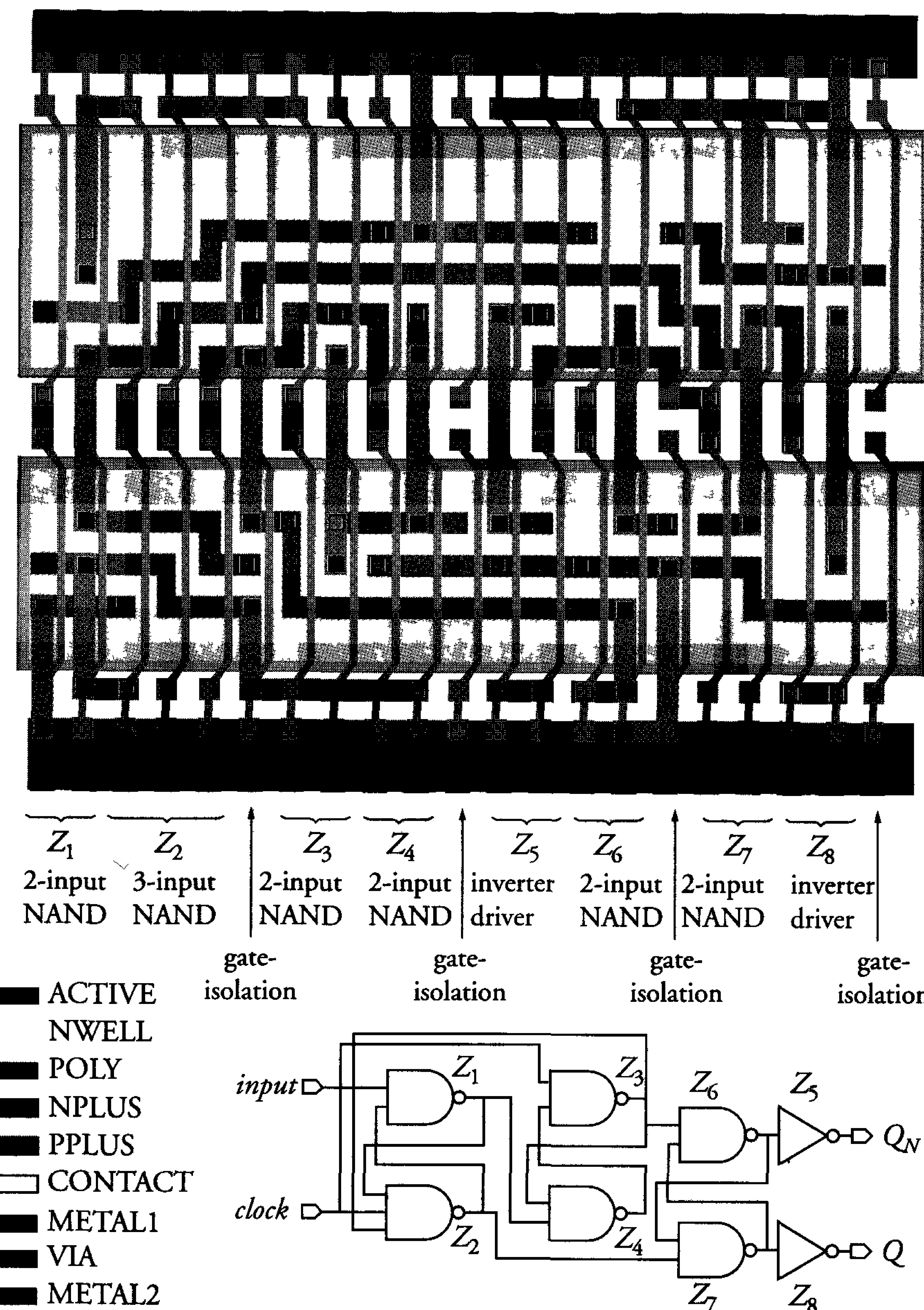


Figure 7.32: *Sea-of-transistors array with gate isolation*

As a result of the continuous drive for increased density and performance, simple PLDs are losing their market share in favour of the high-density flexible PLD architectures. In this way, PLDs are moving closer and closer towards a gate array or cell-based design and are becoming a real option for implementing *systems on silicon*. Another piece of evidence for this trend is the fact that several vendors are offering libraries of embedded cores and megacells. During the nineties, the PLD market has shown to be the most innovative one in terms of new product introductions. In the following, several architectures are presented to show the trend in PLDs.

Field Programmable Gate Arrays (FPGAs)

FPGAs combine the initial PLD architecture with the flexibility of an *In-System Programmability* (ISP) feature. Many vendors currently offer very high-density *FPGA* architectures to facilitate system-level integration (SLI). Current FPGAs are mostly SRAM-based and combine memory and Look-Up Tables (LUTs) to implement the logic blocks. Vendors offering LUT-based FPGAs include Xilinx (XC3000-XC5000 and Virtex families), Lucent Technologies (ORCA families) and ALTERA (FLEX families).

Initially, FPGAs were used to integrate the glue logic in a system. However, the rapid increase in their complexity and flexibility make them potential candidates for the integration of high-performance, high-density (sub)systems, previously implemented in gate arrays [14]. The Xilinx Virtex Series are used as an example of current state-of-the-art FPGA technology. Using a 0.25 μm, five-layer metal CMOS process technology, it offers one million gate devices and addresses the emerging telecommunication, networking and computer market segments. Features of this FPGA family include [15]:

- High density; 1,000,000 gates (1998)
- High performance > 100 MHz
- ASIC design flow (uses synthesis and HDL simulation)
- System features such as multiple sizes of RAM and PLLs
- Support for intellectual properties (cores)
- Support for interface standards (GTL+, LVTTL, SSTL, LVC-MOS....).

Figure 7.33 shows a functional block diagram of the Virtex family architecture. It is built up from different configurable structures (logic, I/O and routing, etc.).

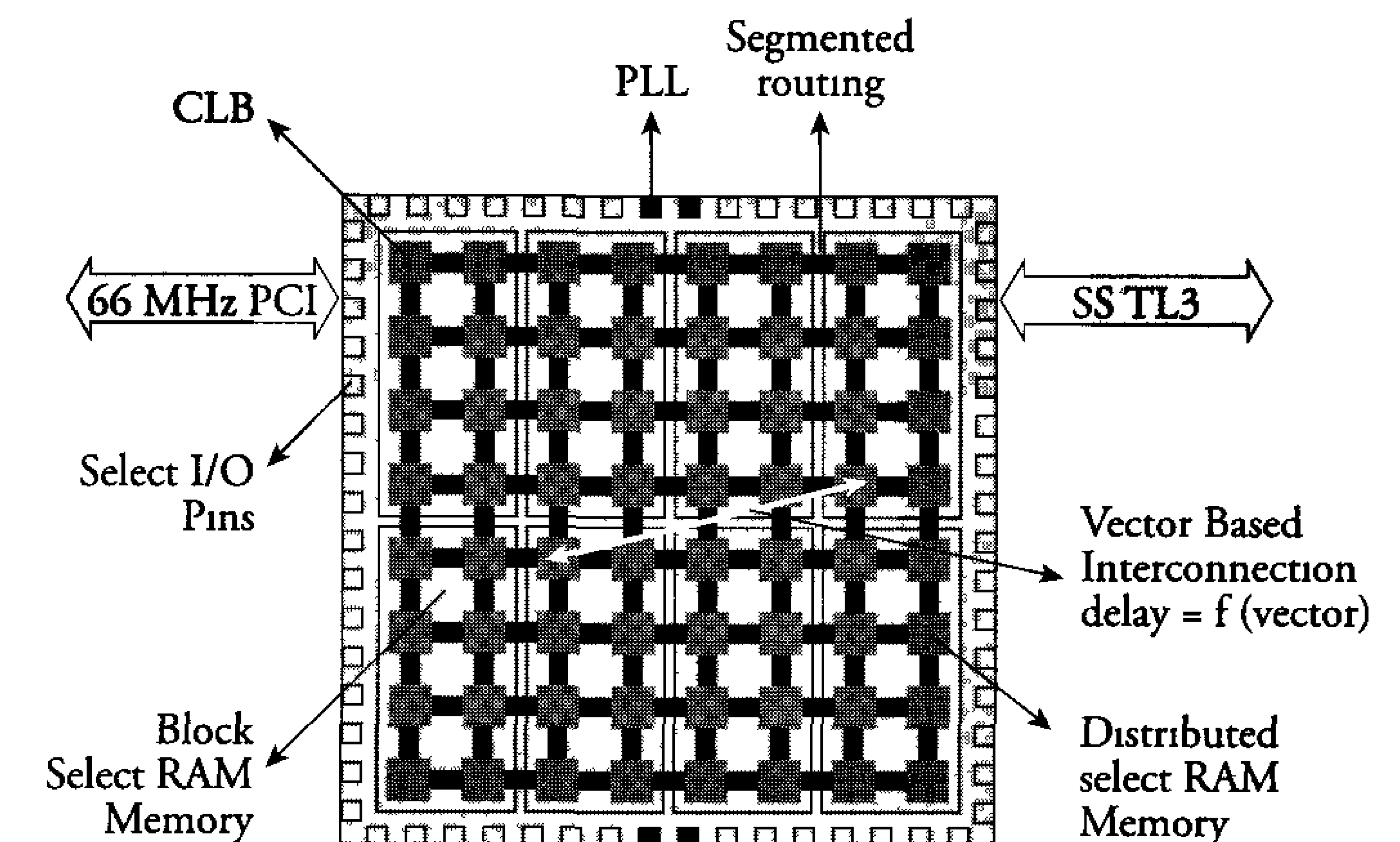


Figure 7.33: *Block diagram of the Virtex family architecture (source: Xilinx)*

The internal structure of this FPGA family consists of an array of VersaBlocks. Each of these blocks contains a Configurable Logic Block (CLB) and a general routing matrix [16]. Figure 7.34 shows such a VersaBlock.

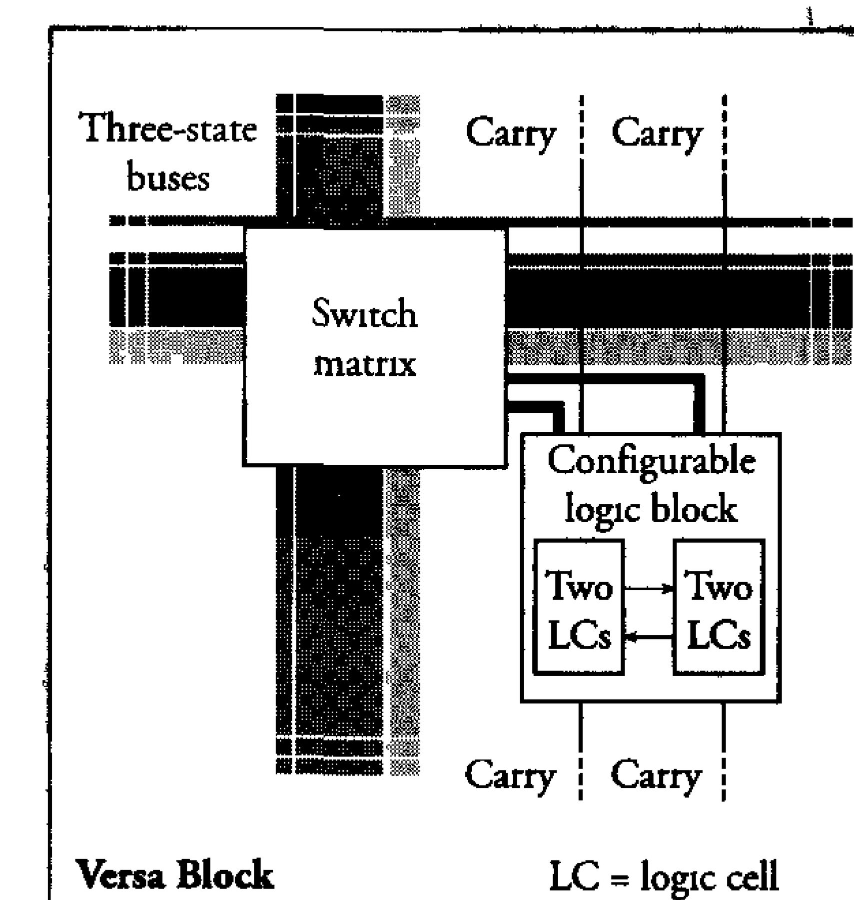


Figure 7.34: *The Virtex VersaBlock containing a CLB and associated routing (source: EDN, Nov. 20, 1997)*

Each CLB contains two logic slices. As shown in figure 7.35, each logic slice contains two logic cells implemented as 4-input LUTs, two registers and dedicated arithmetic carry logic.

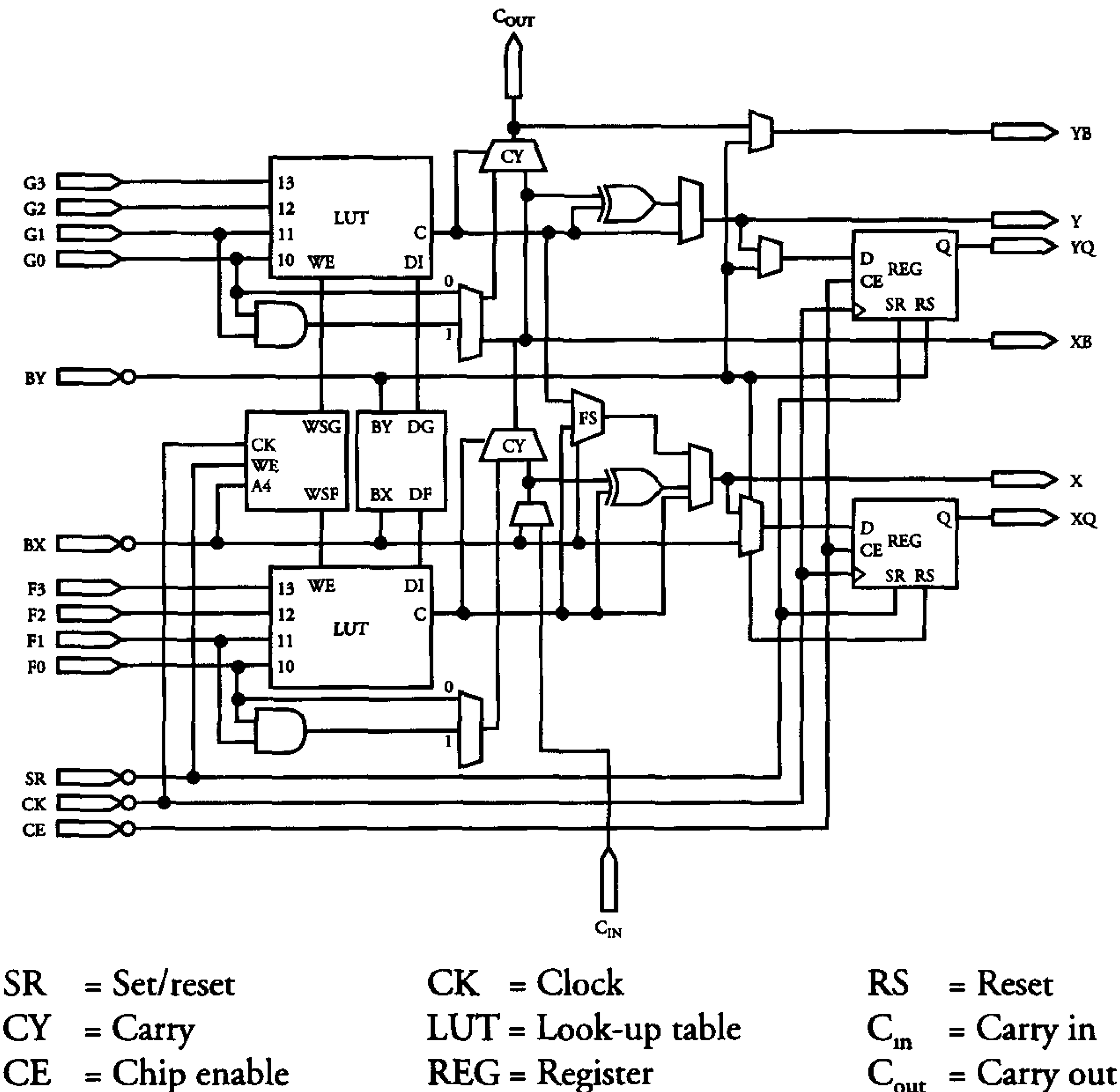


Figure 7.35: *Block diagram of one logic slice of a CLB (source: EDN, Nov. 20, 1997)*

A LUT can generate any function of its inputs. For example, it can generate a 4-input EXOR (9 gates) or, using the built-in carry logic, a 2-bit full adder (9 gates). The register in the CLB accounts for six to twelve gates, depending on the application. Assuming that all three LUTs and both flip-flops are used, a single CLB may contain from 15 to 48 gates of logic. Both memory and logic functions can be realised on this FPGA family. The low-end of this family typically uses all the CLBs for logic with a utilisation of about 18 gates per CLB. The high end of this family assumes that between 20% to 30% of the CLBs are used as memory, while the remaining CLBs are used as logic, with 128

gates per CLB for memory functions and 26 gates per CLB for logic functions.

Logic designs of the Virtex family are supported by software, which includes complete support for libraries, macro-placement and routing. The use of Boundary Scan allows in-system testing of these FPGAs.

Complex Programmable Logic Devices (CPLDs)

The structure of a *CPLD* has evolved from the original PAL^{TM} devices. Vendors offering CPLDs include ALTERA (MAX9000 family), Xilinx (XPLA and CoolRunner^{TM} families), AMD (Mach 4), Lattice (ISPLSI series). As a result of their large densities and high speed, CPLDs can be used in many applications, from implementing glue logic to prototyping gate arrays.

The growth of the CPLD market is driven by the conversion of complex designs, built-up with multiple simple PLDs (SPLDs), into a smaller number of CPLDs [17]. As an example of the current state-of-the-art CPLD complexity, the Xilinx CoolRunner^{TM} PZ3960C/N CPLD is discussed in some detail. The architecture of this CoolRunner^{TM} family is based on eXtended Programmable Logic Array (XPLA), see figure 7.36(a), which combines PAL and PLA structures to offer high density and high performance [18].

The XPLA consists of "Fast Modules" that are interconnected by a Global Zero-power Interconnect Array (GZIA), which is a virtual cross-point switch. Each Fast Module consists of four Logic Blocks of 20 macro cells each, which are connected together by the Local ZIA (LZIA), see figure 7.36(b). Each logic block is a device with 36 inputs from the ZIA and 20 macro cells. It also provides 28 ZIA feedback paths from the macro cells and I/O pins.

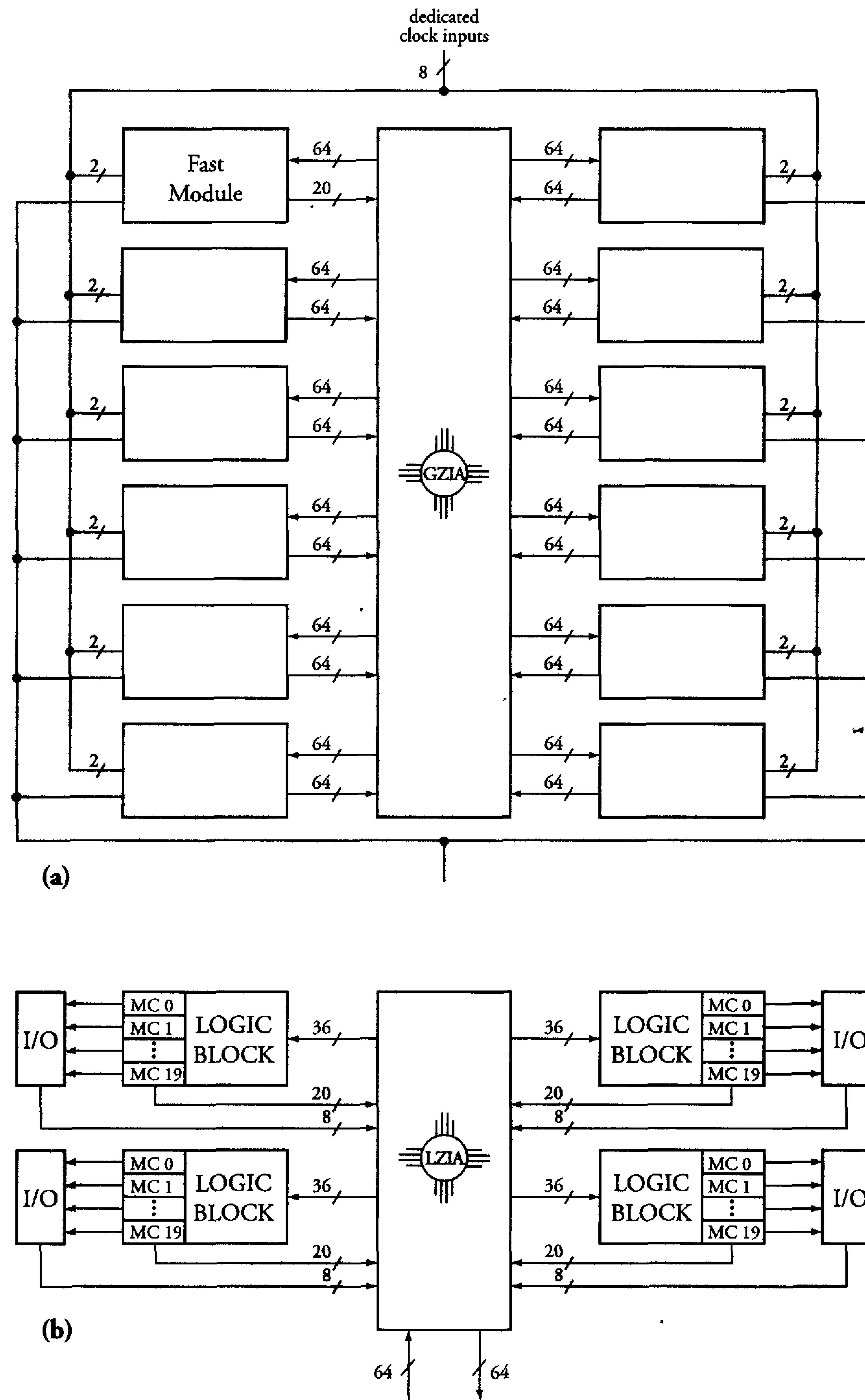


Figure 7.36: *(a) Xilinx XPLA2 CPLD architecture and (b) the XPLA2 Fast Module (source: PHILIPS)*

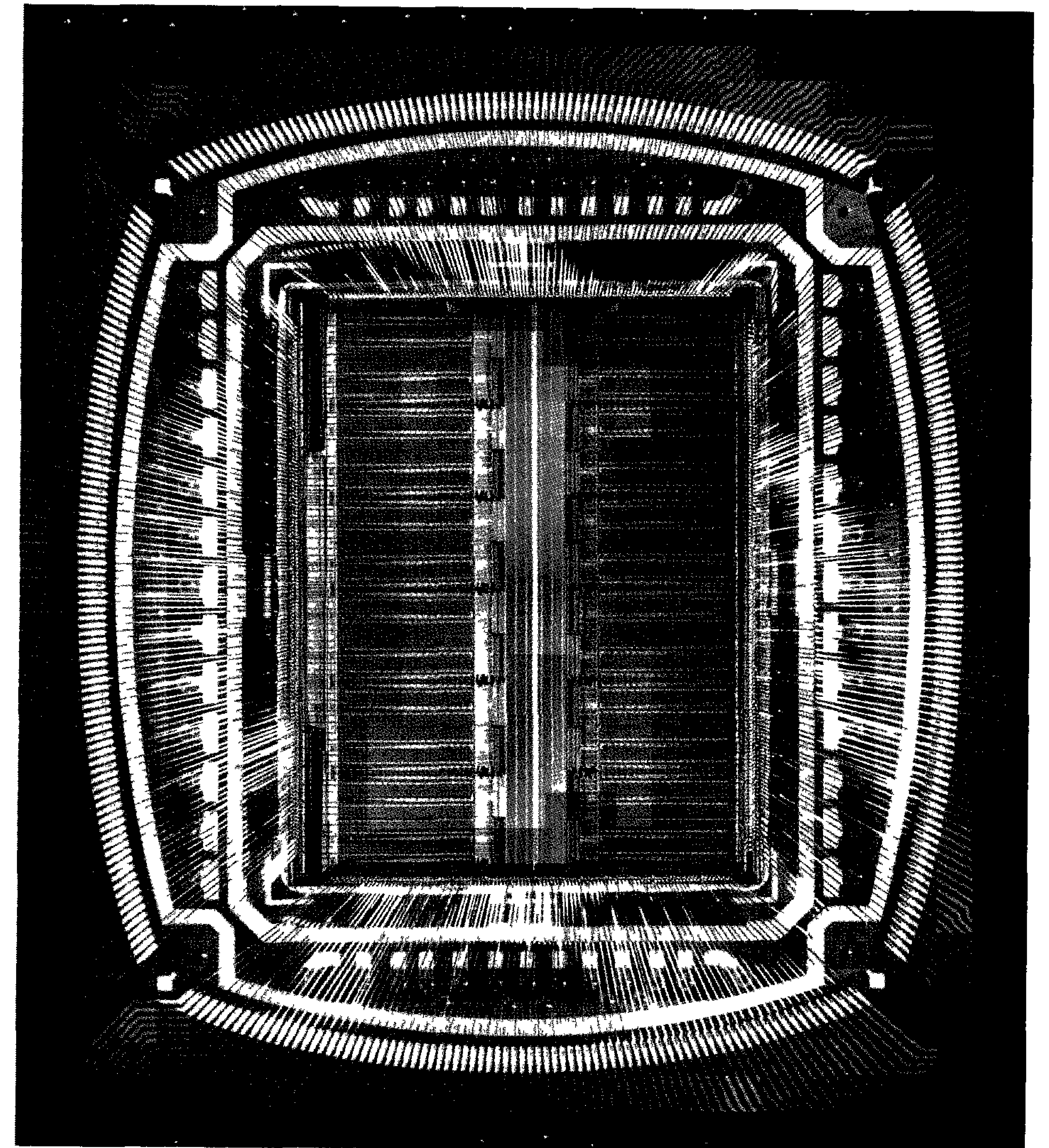


Figure 7.37: *Example of state-of-the-art CPLD chip (photo: PHILIPS)*

The architecture of the Logic Block is shown in figure 7.38.

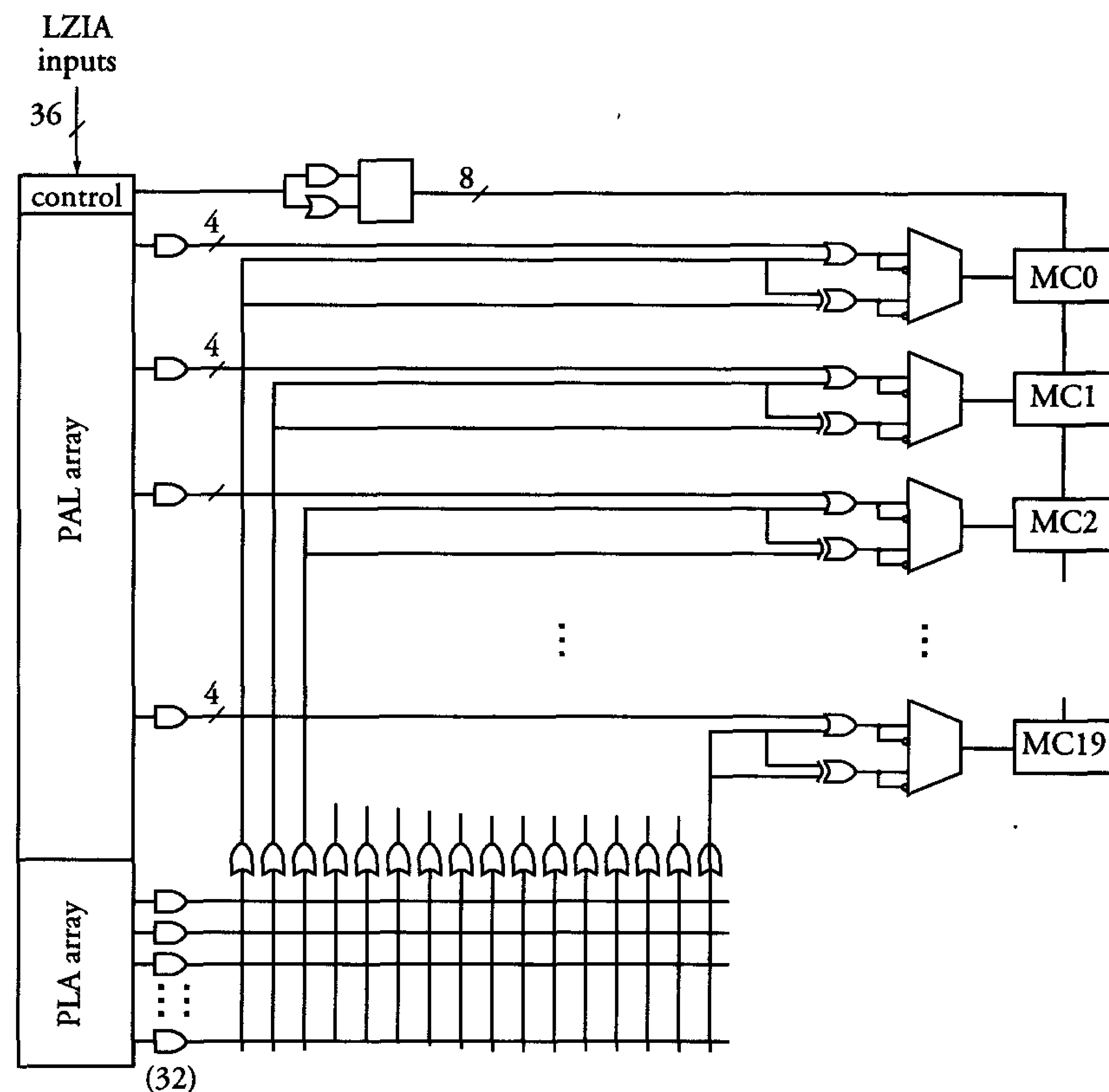


Figure 7.38: *The Logic Block architecture (source: PHILIPS)*

Each logic block contains eight control terms, a PAL array, a PLA array and 20 macro cells. The eight control terms can each be configured as SUM or PRODUCT terms, and are used to control preset/reset and output enable signals of the 20 macro cells' flip-flops. Each macro cell has four dedicated product terms from the PAL array. The PLA array consists of 32 product terms, which are all available for use by each of the 20 macro cells. Figure 7.39 shows the macro cell architecture.

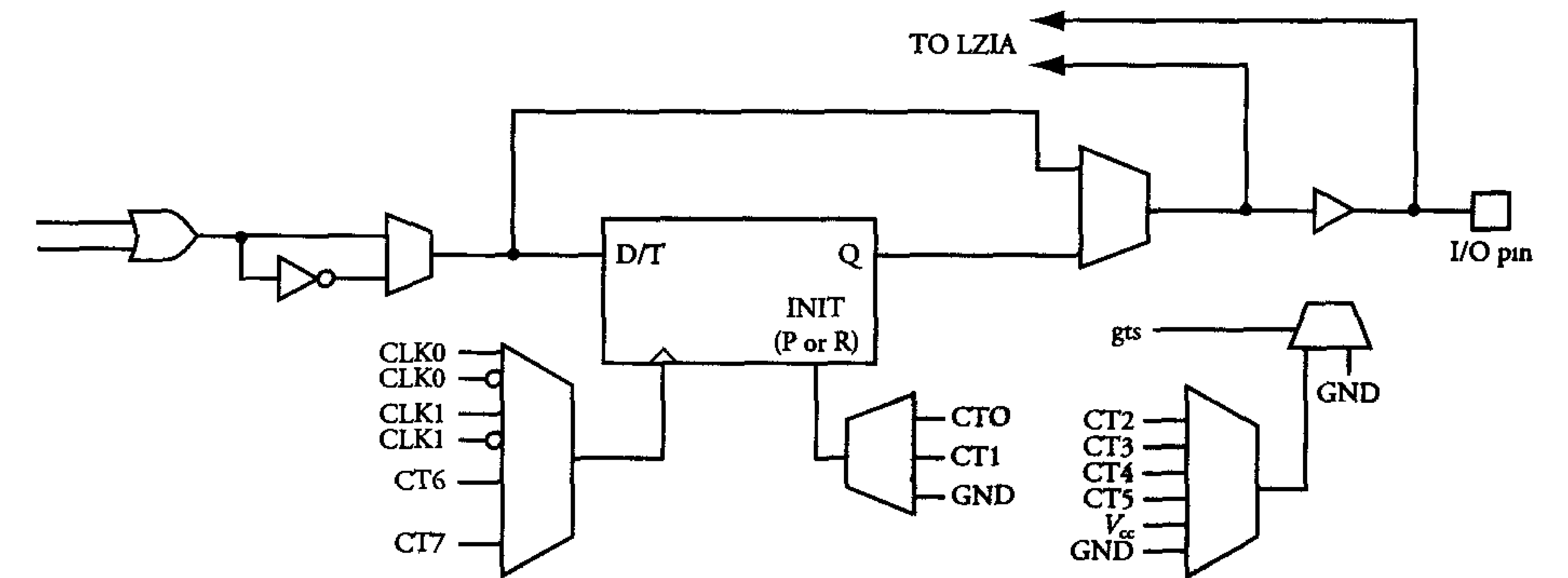


Figure 7.39: *Macro cell architecture (source: PHILIPS)*

The macro cell consists of a flip-flop that can be configured as a D or T type, for use in state machines and data buffering or in implementing counters, respectively. The macro cell also includes several multiplexers for the control of clocking, reset and output enabling functions. A detailed discussion about the operation of this device goes beyond the scope of this book. Some of the features of this device include:

- Low power and high speed
- 3.3 V, ISP
- > 1000 erase/program cycles
- > 20 years retention time
- 0.35 μm EECMOS process
- 492-pin BGA package.

Figure 7.37 shows a photograph of the Coolrunner^{TM}960 CPLD The previous examples of FPGA and CPLD are not discussed in detail: they are only meant to let the reader taste the flavour of the pace of development of these devices and the features that they offer.

Programmability of FPGAs and CPLDs

The most important switch-programming techniques currently applied in FPGAs are SRAM and anti-fuse. Figure 7.40 shows an example of an SRAM-controlled pass transistor to configure the routing of signals through available interconnect patterns. SRAM cells are also used to

configure logic functions. In the figure, the select lines of a multiplexer are controlled by SRAM cells [17].

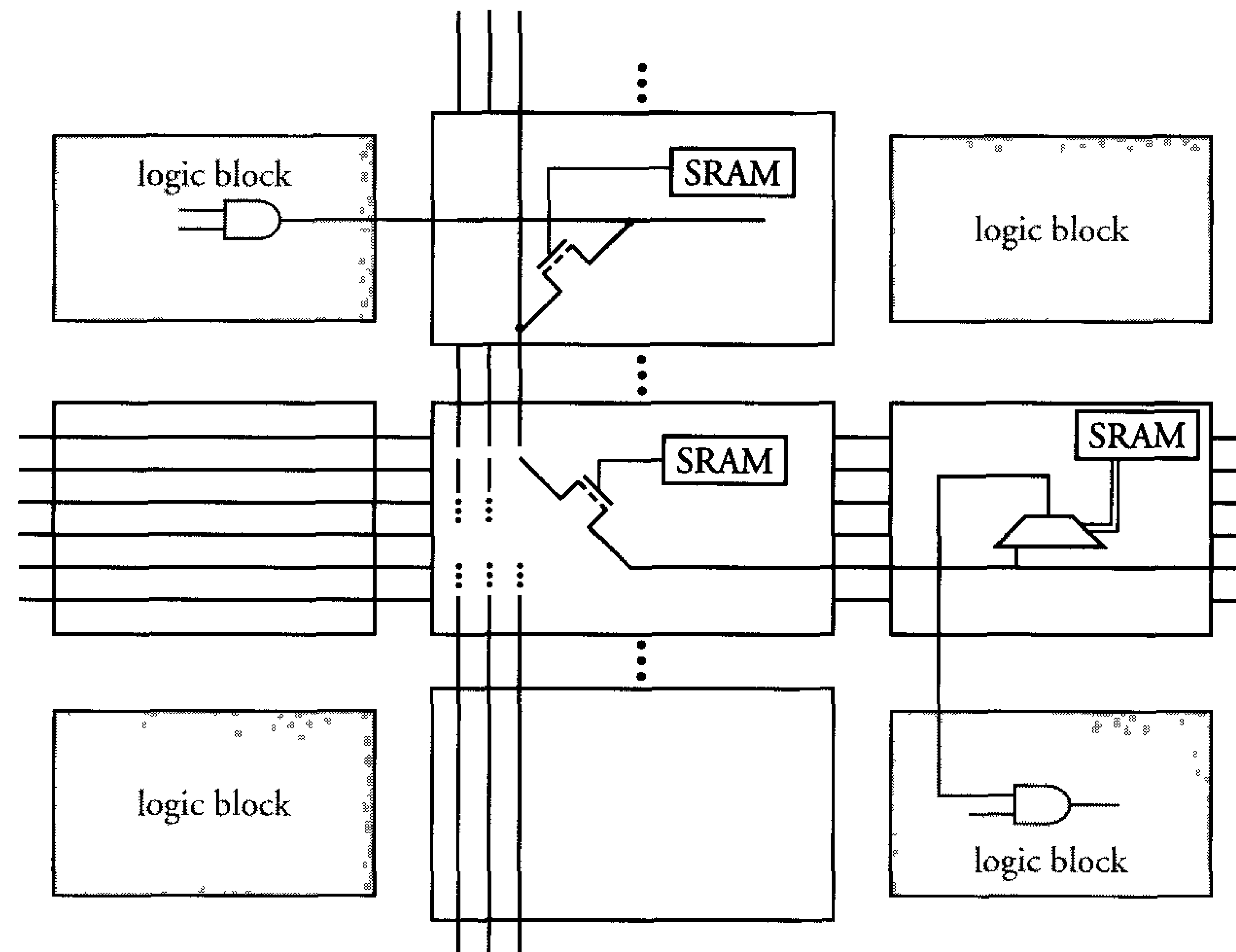


Figure 7.40: *Use of SRAM programmability in FPGAs (source: IEEE Design & Test of Computers, Summer 1996)*

In the majority of current commercially-available CPLDs, the switches are implemented as floating-gate devices, like those in (E)EPROM and flash technologies [17]. However, CPLDs with SRAM programmability are also starting to appear on the market. Here, the switches are used to program PAL and/or PLA arrays, see figure 7.41. In 90% of the CPLDs, the connections are made through programmable multiplexers or full cross-point switches. If an input is not used in a product term (minterm) in an AND plane on a CPLD, the corresponding EPROM gate transistor is programmed to be in the off-state.

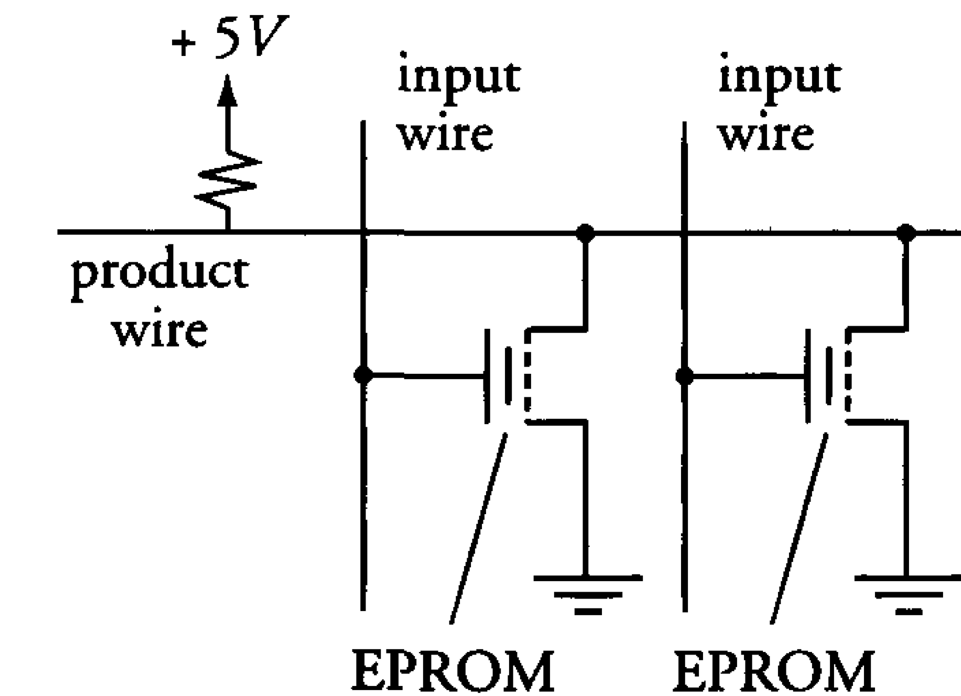


Figure 7.41: *Floating gate device used to program a CPLD (source: IEEE Design & Test of Computers, Summer 1996)*

Similar architectures can be built with EEPROM transistors.

Because SPLDs (which are usually programmed by blowing fuses) are quickly losing their market share, their programming technology is not discussed here. Large complexity PLDs, based on anti-fuse programmability, are also not discussed here, as it is expected that the memory-based programmable PLDs/FPGAs will soon dominate the PLD market.

7.6.8 Hierarchical design approach

The *hierarchical layout* design style is characterised by a modular structure, see figure 7.42 for an example. The different modules are identified during the design path. With a complex system on chip, for example, the various functional modules required emerge from the specification. These modules may include microprocessor core, ROM, RAM and signal processors, etc.

A top-down design strategy generally leads to a satisfactory implementation of a hierarchical layout. The hierarchical division allows various designers to simultaneously produce layouts of the identified modules. Reasonable gate or bit densities are combined with a reasonable speed. The afforded performance renders the hierarchical layout design style suitable for most VLSI and ASIC designs. The design time for hierarchical layouts can be drastically reduced with good CAD tools. Available libraries may contain parameterised *module generators*. Also, IP cores (which are available from different vendors) can be "plugged in", see section 7.2 (definitions: IP) and section 7.4.2.

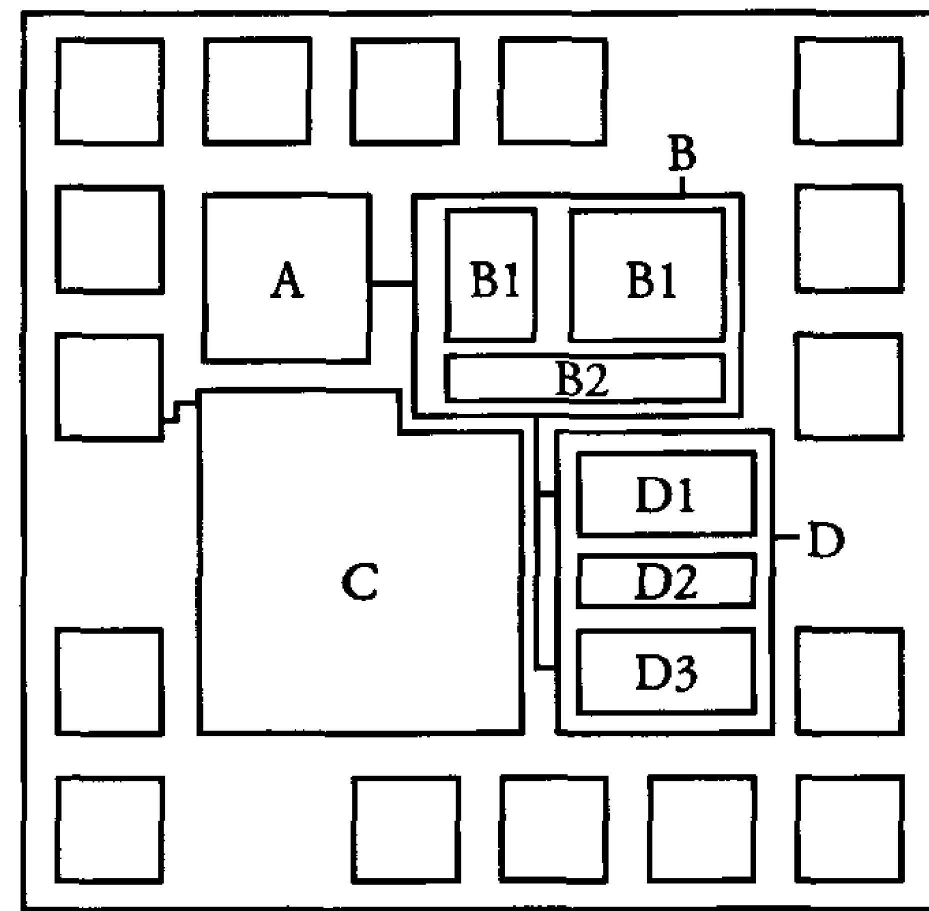


Figure 7.42: *Basic hierarchical layout of a chip*

These (mostly) software descriptions can be used to create layouts of required modules. Assembly of the resulting instances and bond pads leads to the creation of a complete chip layout. Even the assembly and interconnection is automated in *placement and routing* programs.

The hierarchical design style can, of course, include modules which are created by using different layout design styles, e.g. standard-cell or handcrafted module layouts. The hierarchical style was disadvantaged by the relatively large routing areas that could be necessary. However, with the present availability of five to seven metal layers, interconnections and buses can be routed across the logic blocks. In some cases, however, the chip area may not be optimum as a result of the *Manhattan skyline* effect, which results from different block shapes.

Figure 7.43 shows the *meet-in-the-middle strategy* used in the hierarchical design approach. Here, the high-level system description is used to synthesise a design description comprising macro blocks at the implementation level. This implementation level lies roughly in the middle of the top-down design path. The choice of implementation form is still open at this level and possibilities may include a gate array or a cell-based layout. It must be possible to generate these macros from existing design descriptions. Sometimes, module generators are also used to generate a core. The (re)use of IP cores allows a fast "plug-in" of different functional blocks, which are standardised to a certain extent. Clearly, the results of design and layout syntheses meet at the implementation level.

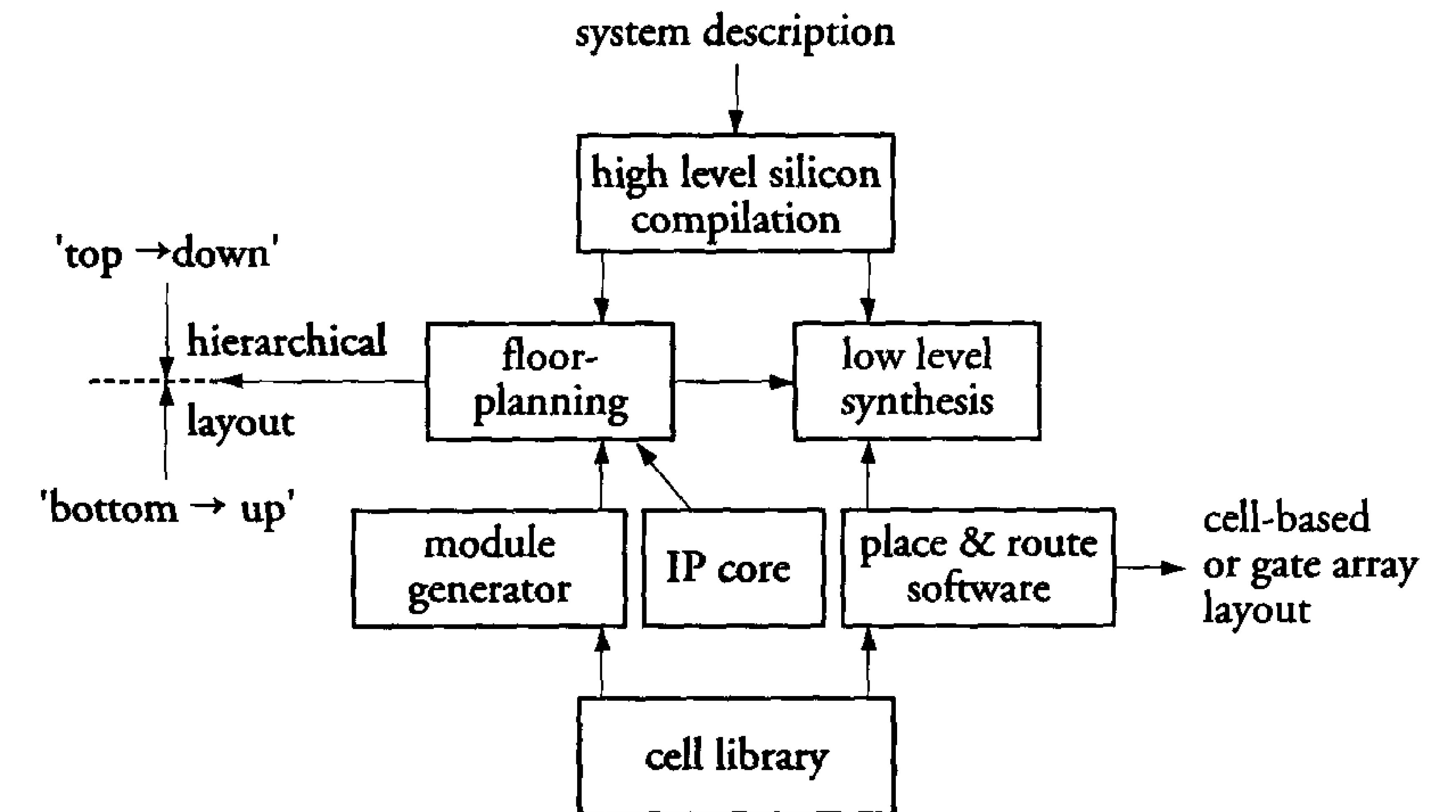


Figure 7.43: *Meet-in-the-middle strategy*

7.6.9 The choice of a layout implementation form

The unique characteristics of each form of layout implementation determine its applicability. The choice of implementation form is determined by chip performance requirements, initial design costs, required volumes and time-to-market requirements. Figure 7.44 shows a cost comparison of the different forms of layout implementation.

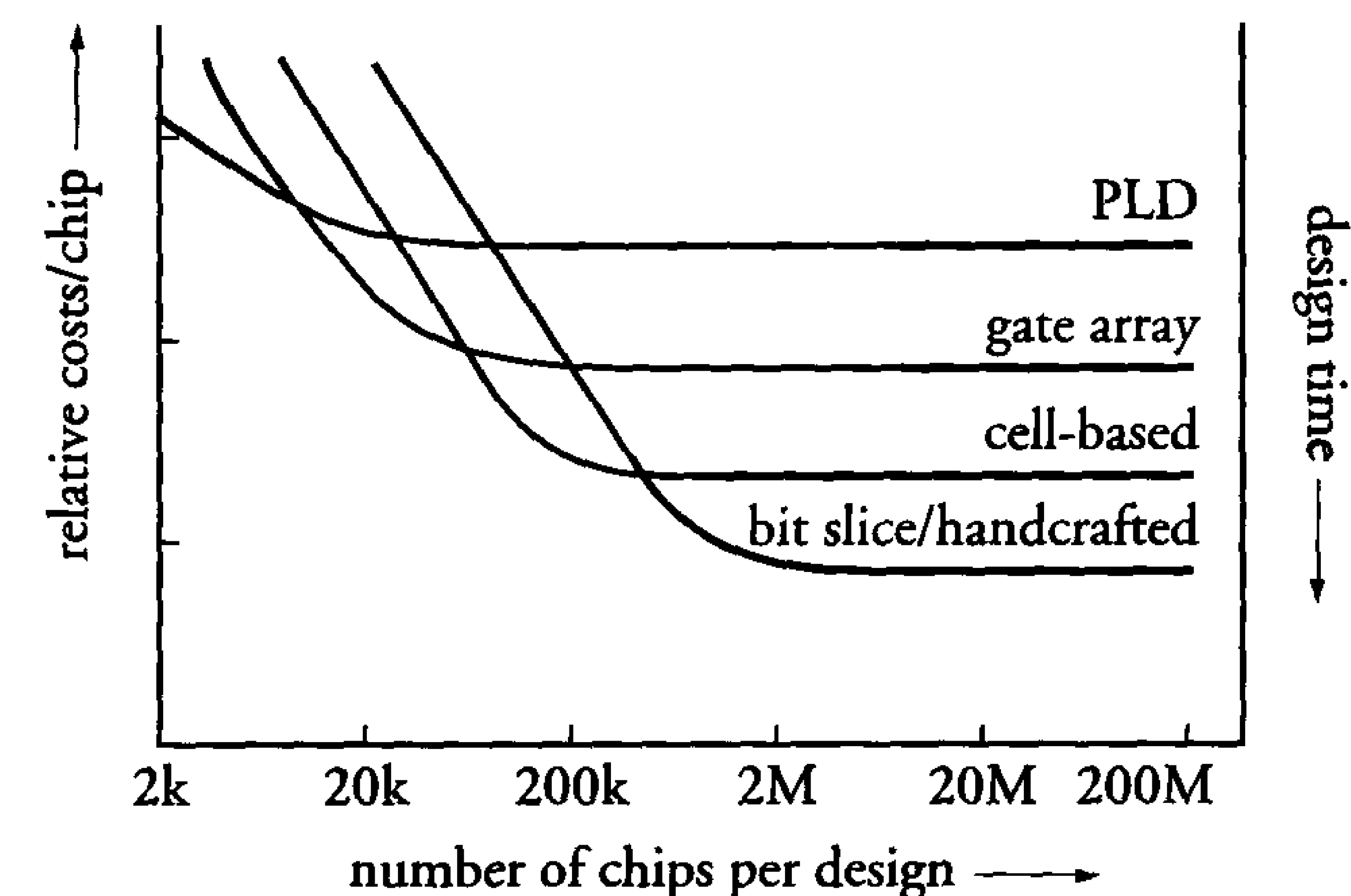


Figure 7.44: *Cost comparison of the different layout implementation forms*

A single chip may combine different implementation forms. Figure 7.45, for example, shows a photograph of a conventional microprocessor in which handcrafted, bit-slice and memory layout styles are combined.

In some cases, a symbolic layout and compaction is used. A *symbolic layout* is a technology-independent design, which can be used for every layout implementation form. In a symbolic layout, transistors and contacts are represented by symbols whose exact dimensions are unspecified while wires are represented by lines whose widths are also unspecified. The abstract symbolic layout is transformed to an actual layout by a compaction program, which accounts for all of the design rules of the envisaged manufacturing process.

The symbolic-layout technique allows a short design time and relieves designers of the need to know specific layout and technology details. The technique is, however, disadvantaged by the associated relatively low gate density and low switching speed. These compare unfavourably with handcrafted layout results. Furthermore, the abstract nature of a symbolic layout only loosely reflects technological aspects. This may result in fatal design errors, such as the implementation of the main clock line in polysilicon instead of in metal. Currently, symbolic layout and compaction are becoming increasingly less popular.

The dimensions of all circuit components and wiring in an IC layout are scaled versions of the actual on-chip dimensions. This *geometric layout representation* is generally described in a *geometric layout description language* (GLDL). GDSII is an example of a GLDL. Such languages are common to many CAD tools and usually serve as the data-interchange format between IC design and manufacturing environments. A GLDL has the following typical features:

- It facilitates the declaration of important layout description parameters, e.g. masks, resolution, dimensions
- It facilitates the definition of geometrical forms, e.g. rectangles and polygons
- It facilitates the definition of macros, e.g. patterns or symbols
- It enables transformations, e.g. mirroring and rotation
- It contains statements for the creation of matrixes.

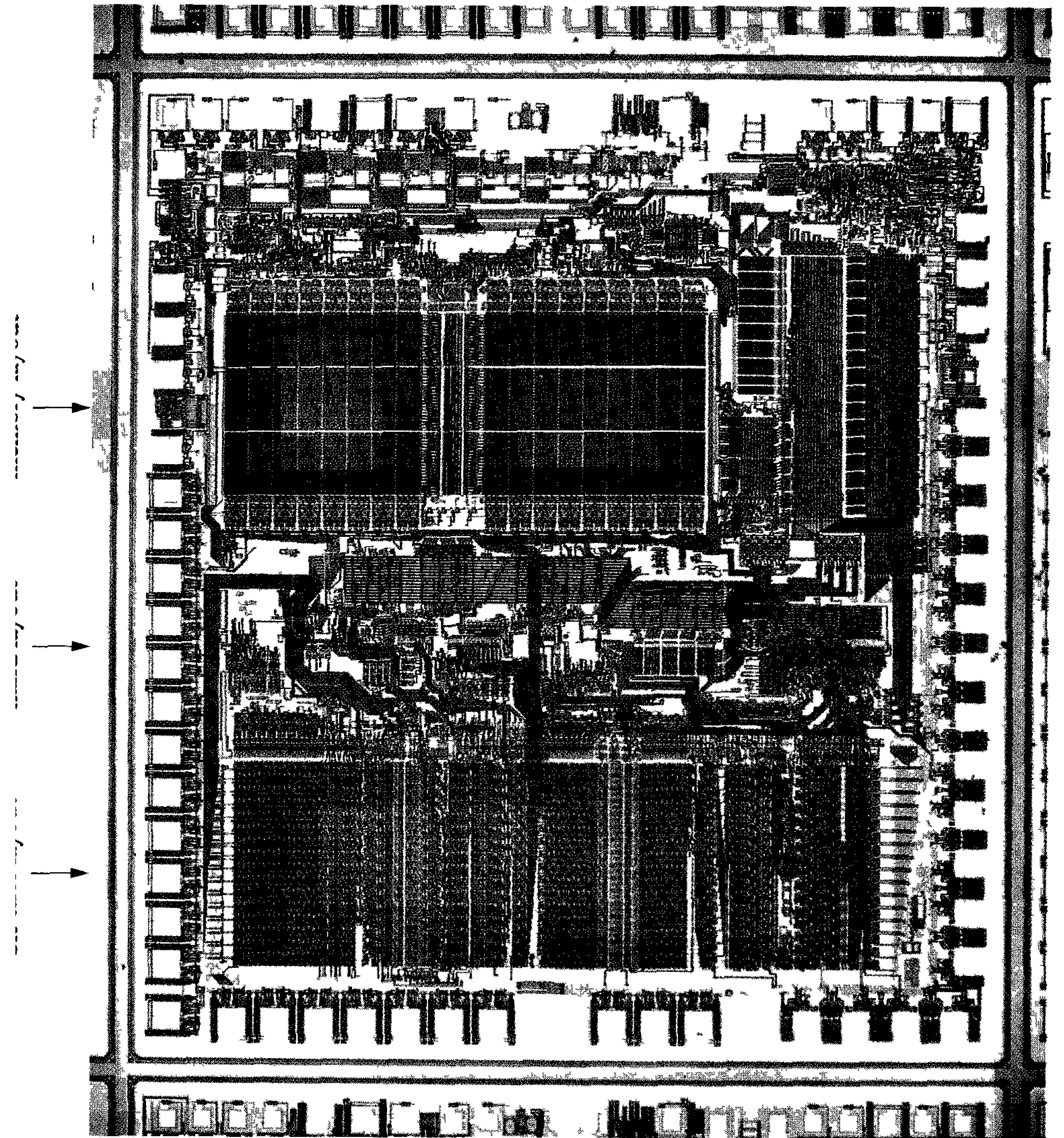


Figure 7.45: *A microprocessor chip which combines different layout implementation forms (photo: PHILIPS)*

7.7 Conclusions

This chapter introduces various design and layout realisations and their characteristic properties. A top-down design approach, combined with a bottom-up implementation and verification through a hierarchical layout style appears suitable for most VLSI circuits. In practice, the design process consists of a number of iterations between the top-down and bottom-up paths, the aim being to minimise the number of iterations.

The use of IP cores that are available from different vendors is fuelling the reuse of existing functionality, such as microprocessor and signal processing cores and memories, etc. This reuse increases the problems with timing and communication between cores from different origins. Chapter 9 discusses these problems in detail.

A good IC design must be accompanied by a good test strategy. Testability is discussed in section 10.2 and requires considerable attention during the design phase. The use of an extra 10 % of chip area to support testability might, for instance, lead to a 50 % reduction in test costs.

7.8 References

General

[1] C. Mead, L. Conway,
'Introduction to VLSI Systems',
Addison-Wesley, 1980

[2] N. Weste, K. Eshraghian,
'CMOS VLSI Design, a Systems Perspective',
Addison-Wesley, 1985

[3] L.A. Glasser, D.W. Dobberpuhl,
'The Design and Analysis of VLSI circuits',
Addison-Wesley, 1985

[4] J. Mavor, et al.,
'Introduction to MOS LSI Design',
Addison-Wesley, 1980

Logic synthesis

[5] R.K. Brayton, R. Camposano, G. DeMichelli, R.H.J.M. Otten and J. van Eijndhoven,
'the Yorktown Silicon Compiler', in silicon Compilation

[6] D.D. Gajski (editor), G. DeMichelli, A. Sangiovanni-Vicentelli, P. Antognetti,
'Designs Systems for VLSI Circuits: Logic Synthesis and Silicon Compilation',
Martinus Nijhof, Dordrecht, the Netherlands, 1987

Abstraction levels

[8] S.G. Shiva,
'Automatic Hardware Synthesis',
Proceedings of the IEEE; Vol 71, No 1, Jan 1983, pp 76-87

[9] B.A. Walker, D.E. Thomas,
'A Model of Design Representation and Synthesis',
Proceedings of the 22nd Design Automation Conference, Las Vegas, 1985, pp 453-459

[10] D. Gajski, R. Kuhn,
'Guest Editors' Introduction: New VLSI Tools',
IEEE Computer; Vol 16, No. 12, December 1983, pp 11-14

Gate arrays

[11] H. Takahasi, et al.,
'A 240 k Transistor CMOS Array with Flexible Allocation of Memory and Channels',
IEEE Journal of Solid-State Circuits, SC-20, No. 5, October 1985

[12] I. Okhura, et al.,
'A novel basic cell configuration for CMOS gate-array',
CICC 1982, pp 307-310, May 1982

[13] H.J.M. Veendrick, et al.,
'An efficient and flexible Architecture for High-Density Gate Arrays',
ISSCC 1990, Digest of Technical Papers, San Francisco

PLDs

[14] Xilinx,
'The future of FPGAs',
Xilinx Web-site, 1998

[15] Xilinx,
'The Virtex Series of FPGAs',
Xilinx Web-site, 1998

[16] Dave Bursky,
'High-Density FPGA Family delivers Megagate capacity',
Electronic Design, 20 November 1997

[17] S. Brown, J. Rose,
'FPGA and CPLD Architectures: A tutorial',
IEEE Design&Test of Computers, Summer 1996

[18] Philips Semiconductors,
'Datasheet of PZ3960C/N',
May 1998

Related publications

[19] 'ASIC outlook 1999' and 'Status 1999, A report on the IC Industry',
Integrated Circuit engineering

[20] V.K. Madisetti,
'Digital Signal Processors, An Intro to Rapid Prototyping and Design Synthesis',
Butterworth-Heineman, 1995

[21] Jan M. Rabaey,
'Digital Integrated Circuits: A Design Perspective',
Prentice Hall, 1995

[22] Kerry Bernstein, et al.
'HIGH SPEED CMOS DESIGN STYLES',
Kluwer Academic Publishers, 1999

7.9 Exercises

1. Why are abstraction levels used for complex IC designs?

2. What is meant by floor planning?

3. Explain what is meant by logic synthesis.

4. What does the term 'Manhattan skyline' describe in relation to a VLSI layout?

5. Assume that a standard-cell and a gate array library are designed in a two-metal layer CMOS technology. The libraries consist of logic cells with identical logic functions. Describe the main differences between the two libraries in terms of:
   a) Cell design
   b) Chip area
   c) Production time and cost
   d) Applications

6. Random logic functions can, for instance, be implemented using a ROM or a standard-cell realisation. Explain when each of these possibilities is preferred.

7. Draw a schematic diagram of a PLA which implements the following logic functions:

   $$F_0 = \overline{\bar{x}\,\bar{y} + xyz}$$

   $$F_1 = \overline{x\,\bar{y} + \bar{x}y + x\bar{z}}$$

   $$F_2 = \overline{xyz + \bar{x}\,\bar{y}\,\bar{z}}$$

8. Explain what is meant by mixed-level simulation.

9. Explain in your own words what is meant by IP. What is the cause of its existence? How can it affect design efficiency and what are the potential problems involved with it?

10. Explain the differences between an FPGA and a CPLD.

Chapter 8

Low power, a hot topic in IC design

8.1 Introduction

Although already used in the seventies, it took until the mid-eighties before CMOS became the leading edge technology for VLSI circuits. Prior to that time, only a few designs were implemented in CMOS. At that time, only those applications that really required the low-power features of CMOS were designed in it. Most examples, then, were battery supplied applications, such as wristwatches (tens of millions per year), pocket calculators, portable medical devices (hearing aids and implantable heart controls) and remote controls.

Through the eighties into the nineties, however, the number of transistors per chip increased from hundreds of thousands to millions, while chip frequencies increased from several Megahertz to hundreds of Megahertz (alpha-DEC; Pentium-Intel). The power consumption of many CMOS ICs therefore reached the level of 1 W, which is the maximum allowed power consumption of a cheap plastic package. This is one of the main driving forces for low-power CMOS. It was also the reason for switching from nMOS to CMOS technology in the early eighties.

Currently, the requirement to also have access to powerful computation at any location is another driving force to reduce CMOS power dissipation.

The increasing number of portable applications is a third driving force for low-power CMOS. In the consumer market, we can find examples such as games, walkmans, cameras, CD players and TVs. In

the PC market, an increasing percentage of computers is sold as notebook or laptop computers. Digital cellular telephone networks, which use complex speech compression algorithms, form a low-power CMOS application in the telecommunication field.

Finally, the emerging multimedia market will also show many new products in the near future. At the time of going to print, we see the portable full motion video as an example of a low-power application. The personal digital assistant (PDA) is already available to the consumer. The development of these portable and hand-held devices has advanced battery technology significantly in the last couple of years. Longer battery life requires higher energy efficiency, while smaller weight and size requires a reduced number of stacked battery cells. The performance of cells in a series is substantially worse than that of individual cells. A single-cell battery with both high cell voltage and high energy efficiency is a real need. Nickel-cadmium batteries have dominated the battery market, but they suffer from low cell voltage and low energy efficiency (120 Wh per litre). Now, Lithium-ion batteries have emerged as the more favoured choice, because they offer higher cell voltage and higher energy density (up to 400 Wh per litre). As the world becomes more mobile, the demand for better battery technology will continue to increase. More information on battery technologies can be found in [16].

Another important driving force for low power is the future system requirement. In a 0.1 μm CMOS technology (2003), $4 \cdot 10^9$ to $40 \cdot 10^9$ transistors will be packed on a board of 20 by 20 cm with very high-density packaging techniques (multi-chip modules MCM).

Current power levels are not acceptable for these systems. In general, low power also leads to simpler power distribution, less supply and ground bounce and a reduction of electromigration and electromagnetic radiation levels. A low-power design attitude should therefore be common in every IC design trajectory, because it is beneficial for power consumption, robustness and reliability of current and future ICs and systems.

8.2 Sources of CMOS power consumption

During the operation of CMOS circuits, there are four different sources that contribute to the total power consumption:

$$P_{\text{total}} = P_{\text{dyn}} + P_{\text{stat}} + P_{\text{short}} + P_{\text{leak}}$$

where P_{dyn} represents the dynamic dissipation.

This is the power dissipated as a result of charging and discharging (switching) of the nodes, and can be represented by the following equation:

$$P_{\text{dyn}} = C \cdot V^2 \cdot a \cdot f$$

where C is the total capacitance,
V is the voltage swing,
f is the switching frequency and
a is the activity factor.

The activity factor represents the average fraction of gates that switch during one clock period. This number can be as low as 0.1 (low activity), for example, but it can also be as high as 4 (very high activity) because of hazards, see paragraph 8.4.3. P_{stat} represents the static dissipation. This is the power dissipated as a result of static (temporary or continuous DC) current. In section 8.4, the basic causes of the different contributions are explained in detail.

The contribution of the short-circuit dissipation is represented by P_{short}. This is the power dissipated in logic gates as a result of short-circuit currents between supply and ground during transients.

Finally, the last contribution to the total power dissipation is made by the leakage dissipation P_{leak}. This is the power dissipated as a result of substrate currents and sub-threshold currents. Both technology and design can affect several of these power dissipation contributors, see table 8.1.

Table 8.1: Power dissipation contributors

Contributor	Technology dependent	Design dependent
P_{dyn}	x	x
P_{stat}		x
P_{short}		x
P_{leak}	x	

The following sections discuss the technological and design measures that can be taken to reduce the different power consumptions.

8.3 Technology options for low power

As can be seen from in table 8.1, technology can affect both the dynamic and the leakage power dissipations.

8.3.1 Reduction of P_{leak} by technological measures

As a result of scaling the channel length over generations of technologies, we have arrived at a point (with channel lengths of 0.35 μm and less), where we also have to reduce the supply voltage to limit the electrical fields inside a MOS transistor, see chapter 2. Reducing the supply voltage means that the circuits become relatively slower. Therefore, the threshold voltage also has to be reduced.

This can have consequences for the leakage currents as well as for the noise margin within digital circuits. Because of the sub-threshold (weak-inversion) currents, as discussed in chapter 2, we will have a leakage current through an nMOS transistor when its gate voltage is at zero volt. The higher the threshold voltage, the less leakage current will flow at $V_{gs} = 0\,\text{V}$.

Let us define the sub-threshold slope $S_{\text{subthr.}}$ to be the change in threshold voltage, causing a ten-fold increase of the sub-threshold current at $V_{gs} = 0\,\text{V}$. In current technologies, $S_{\text{subthr.}}$ is between:

$$50\,\text{mV/decade(I)} < S_{\text{subthr.}} < 100\,\text{mV/decade(I)}$$

This means that a reduction of the threshold voltage of 100 mV leads to an increase of leakage current (at $V_{gs} = 0\,\text{V}$) of a factor between 10 to 16. It should be clear that, for power and speed reasons, an optimum has to be found for the threshold voltages of both nMOS and pMOS transistors.

Example:
Assume a reference transistor with an aspect ratio of: $W/L = 10/0.25$. If $V_T = 0.5\,\text{V}$, then its leakage current might be 100 fA. Suppose the threshold voltage shifts to 0.3 V, now the current will increase by a factor of about $15 \cdot 15 \approx 225$ to 2.25 pA. Present standby currents in large RAMs can be in the order of several tens of nano amperes.

With decreasing channel lengths, the threshold voltage also decreases as a result of the small channel effects (threshold voltage roll-off; chapter 2). Consequently, the threshold voltage can be as low as 0.3 V for a 0.18 μm CMOS process. Also, for real low-voltage applications, the threshold voltage should be low to allow for a certain speed. However,

at these low threshold voltages, the circuits suffer from a relatively large loss of power caused by leakage currents, especially in the standby mode.

There are several solutions to this problem. One is to vary the threshold voltage by applying a back-bias voltage during standby mode [1]. Depending on the K factor in the equation for the threshold voltage (equation (1.16)), the threshold voltage can be increased by several hundreds of millivolts by applying a substrate (back-bias) voltage of several volts. These additional back-bias voltages (both nMOS and pMOS need back-bias in the standby mode) can either be supplied by additional supply pads, or generated on the chip. The back-bias voltage can be offered to the complete chip, or only to the core and memories, excluding the I/O circuits. If core and I/O are to be treated differently (e.g. for protection reasons), a triple-well technology is required [2].

Figure 8.1 shows a cross-section of a triple-well device. In this technology, all substrate and well areas in the core are connected to separate back-bias voltages. The nMOS transistors are isolated from the substrate. In this way, the I/O noise can hardly affect the core circuit.

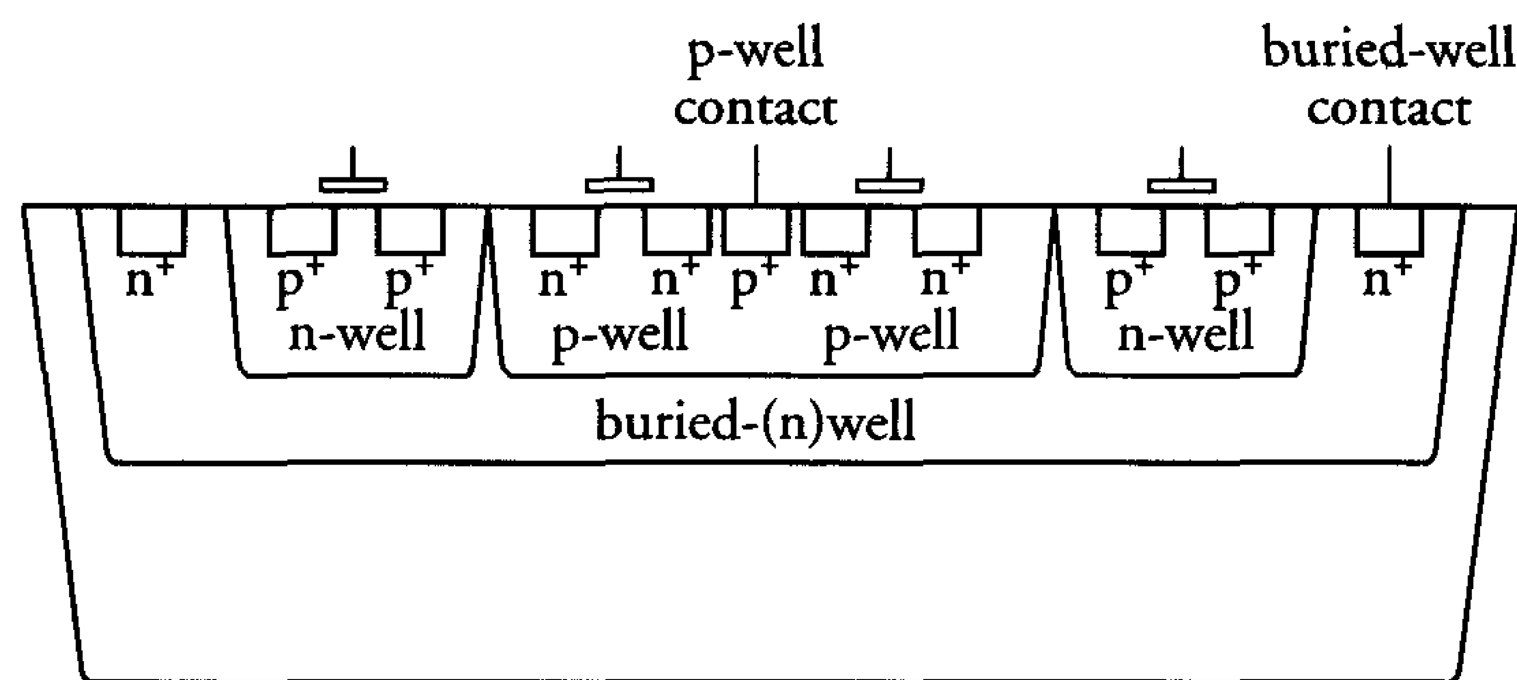


Figure 8.1: *Cross-section of a triple-well device*

A second approach to reduce standby (leakage) currents is to use multiple thresholds [3]. Now, the power supply of the core (with low V_T circuits) is switched by a very large transistor with high V_T, see figure 8.2.

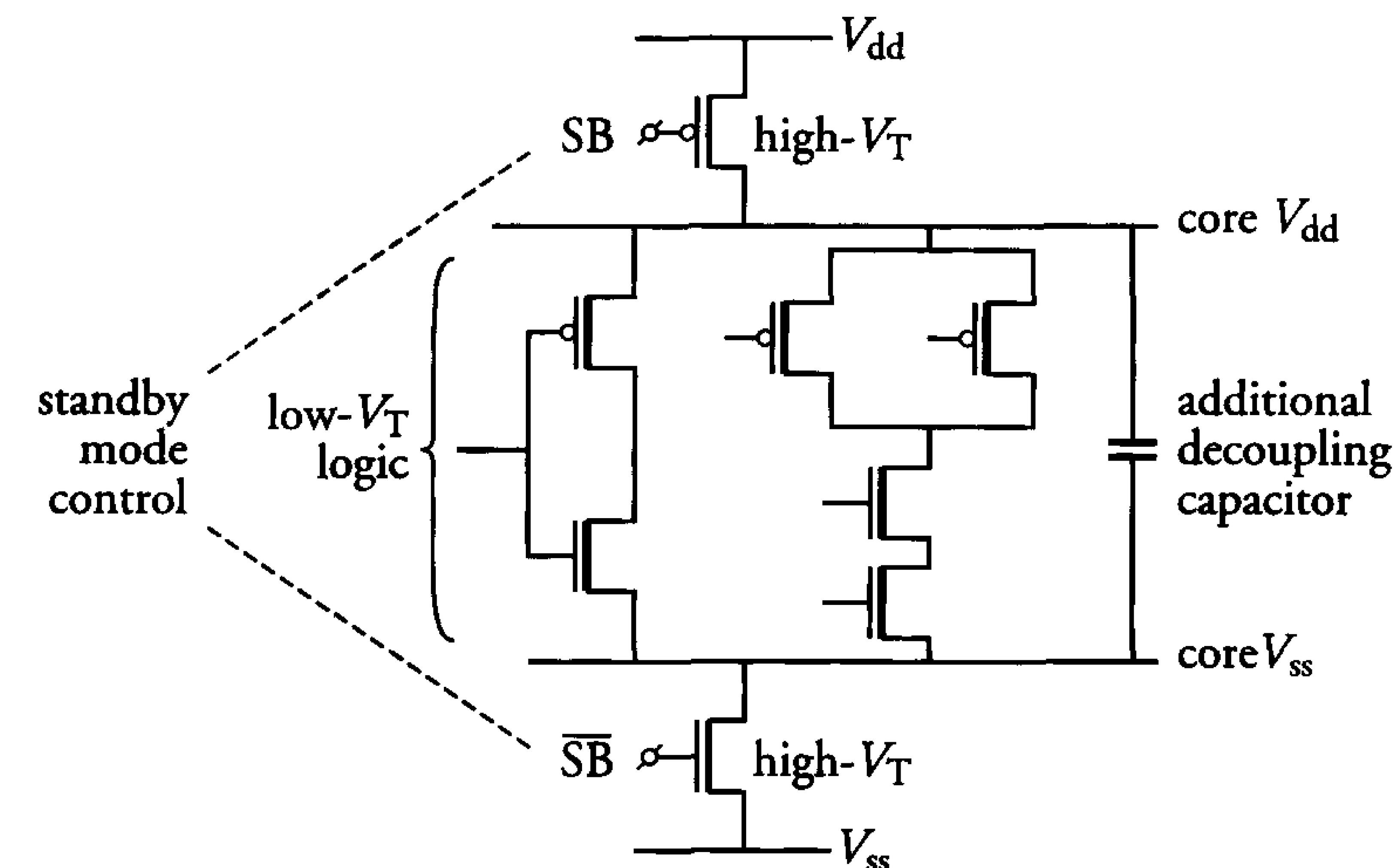


Figure 8.2: *Power supply switch in a multiple V_T environment*

The width of this transistor is such that there is only a marginal voltage drop across it. However, all storage cells and memories must be connected to the permanent power supply and have a high V_T, in order not to lose the cell data.

Although low voltage analogue circuits show only marginal performance, they can still be included with digital circuits on a mixed analogue/digital low-voltage IC.

Usually, the analogue supply voltage is then increased by a voltage-up (DC-DC) converter. In all low-voltage cases, the analogue circuits must be put within a well (triple-well technology). If back-bias is applied globally, the nMOS transistor in the analogue part would show quite a different performance. The analogue circuits would no longer operate properly. Therefore, the analogue nMOS transistors must also be isolated from the digital part, if a back-bias is used to power down the digital part in the standby mode.

To reduce substrate currents, protection must be given from hot carrier effects. This means that LDD techniques are included from 0.8 μm up to 0.25 μm CMOS processes, to reduce the peak of electric field in the

pinch-off region close to the drain, see section 2. Also, optimised dopes for source/drain and n-well implants must be chosen to limit junction leakage currents.

8.3.2 Reduction of P_{dyn} by technology measures

In the following formula for the dynamic dissipation, both capacitance C and voltage V are partly determined by the technology:

$$P_{\text{dyn}} = C \cdot V^2 \cdot a \cdot f$$

Generally, the load (capacitance) of a logic gate is formed by the interconnection capacitance, the gate capacitance (fan-in of the connected logic) and the parasitic junction capacitances in the logic gate itself.

A reduction of the gate capacitance means a thicker gate oxide, which also affects the β and thus the speed of a MOS transistor dramatically. There is no alternative. The reduction of the interconnect capacitances depends on the thickness and the dielectric constant of the oxide and on the track thickness, see figure 8.3.

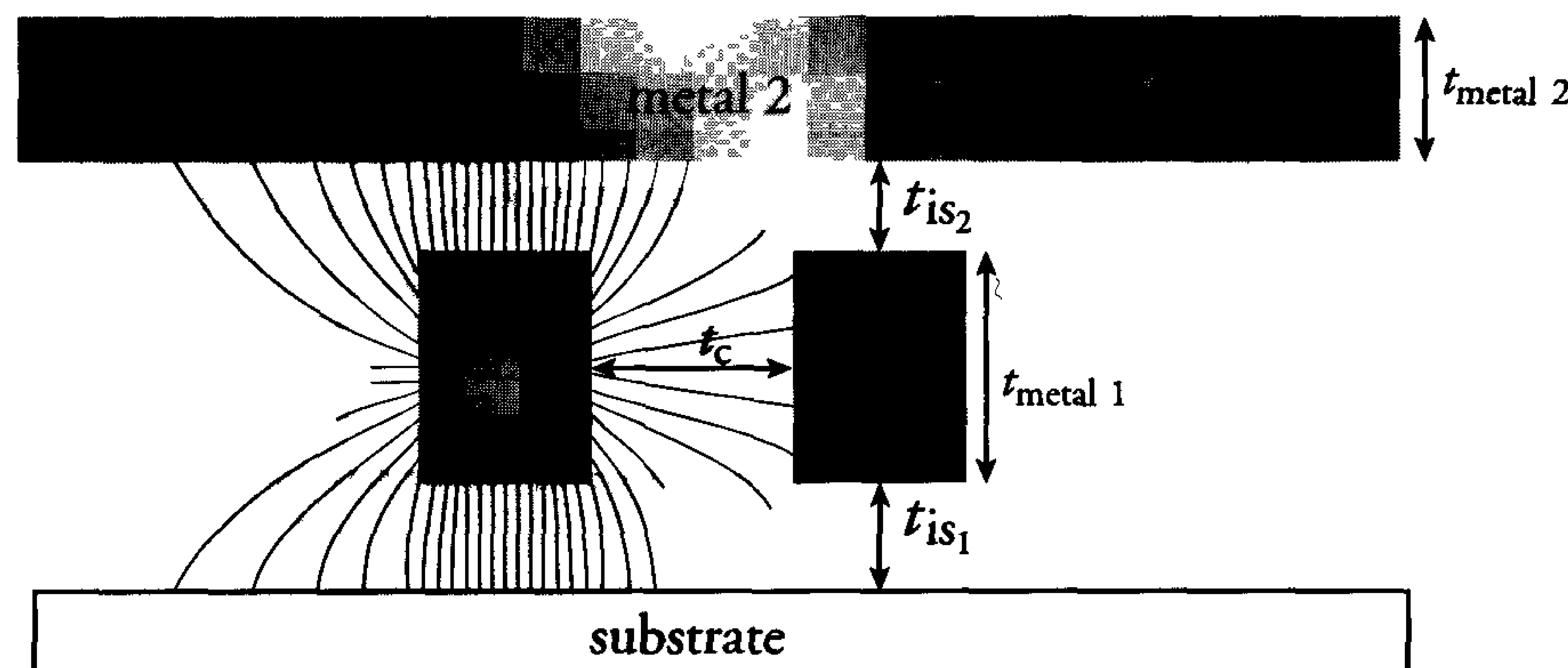


Figure 8.3: *Cross-section of multilevel metal to show capacitance*

As a result of resistive and electromigration effects, the thickness of the metal layers can only be reduced if other metals are used for routing. In this respect, copper is found to be a good candidate, allowing 25 to 30% reduction of the metal thickness. This leads to less mutual capacitance. A reduction of power dissipation and cross-talk (chapter 9) is the result.

Thicker oxides require more advanced planarisation steps. The space between two metal tracks in the same layer cannot be increased much,

as the chip area will increase as well. It thus hardly affects the power dissipation.

One way to decrease the dielectric capacitance is to find materials with a lower ϵ. The ϵ of SiO_2 is around 4, the ϵ of air is 1. A factor of two must be achievable in the near future.

Junction capacitances are formed by the depletion regions of the source and drain junctions of both nMOS and pMOS transistors. The thicknesses of the depletion regions and, therefore, the values of their capacitances, are determined by the dope of the n^+ and p^+ regions. A reduction of the junction capacitances can thus be achieved by optimised dope of n^+ and p^+ regions. An alternative to the main stream current CMOS processes for low power might be Silicon on Insulator CMOS and SIMOX. These technologies are discussed in section 3.8.4.

8.3.3 Reduction of P_{dyn} by reduced-voltage processes

The decrease of the channel length over generations of technologies has increased the peak of the electrical field in the pinch-off region near the drain to unacceptable values. For a 0.7 μm technology, LDD structures (section 2.7) brought a satisfactory reduction of this electrical field. However, from about 0.6 μm technologies and beyond, these LDD structures are no longer sufficient. The only way to reduce the peak electrical field is to lower the supply voltage, see figure 8.4.

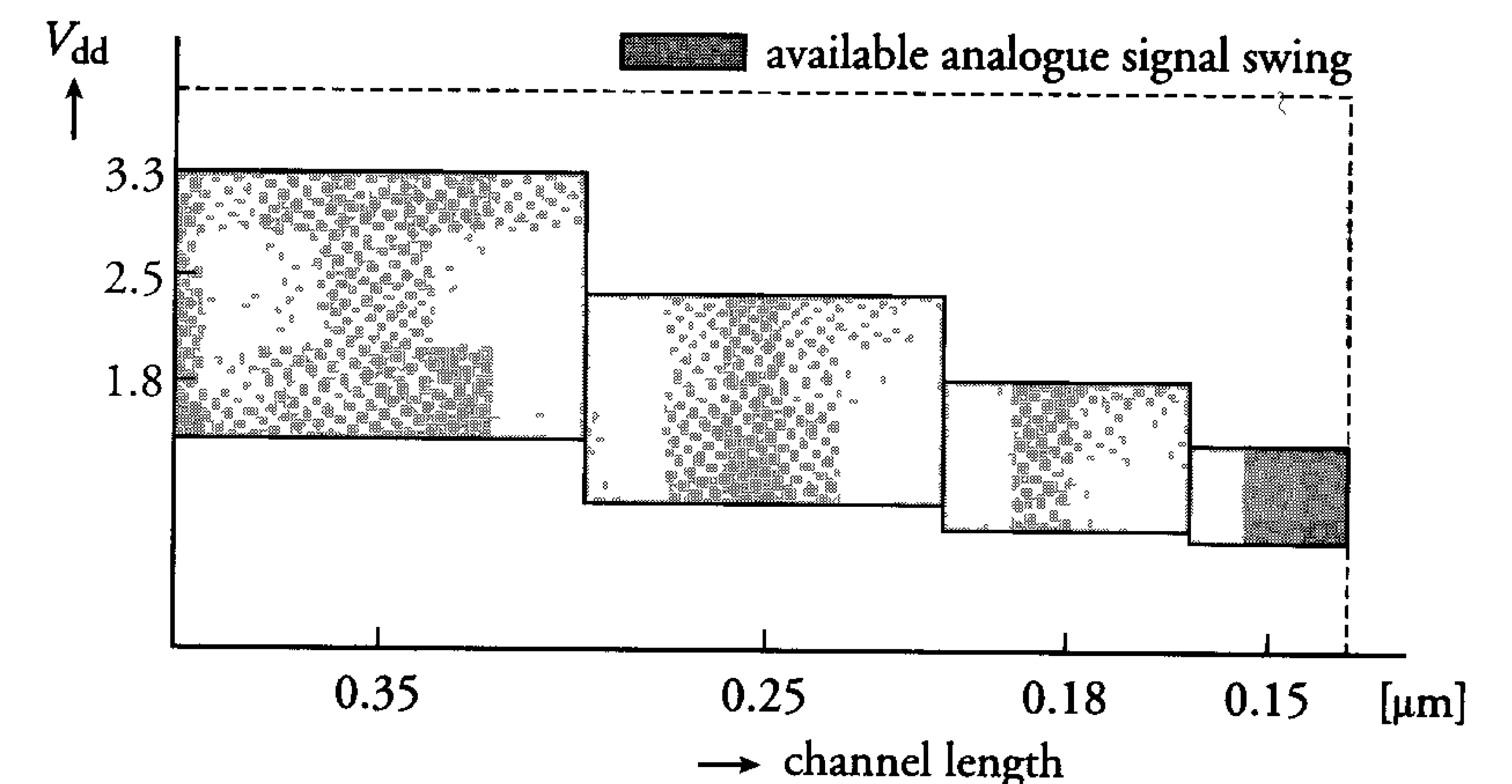


Figure 8.4: *Reduction of supply voltage as a function of the channel length*

In the future, shorter channel lengths will require lower V_{dd} voltages. For performance reasons, the threshold voltage V_T should also be reduced. However, this leads to an increase of the sub-threshold (leakage) currents, see section 8.3.1. A very important parameter that reflects the performance of a technology is the power-delay product ($\tau \cdot D$). Minimisation of this product means that the circuit is very efficient with respect to its power consumption:

$$\text{power}(D) = C \cdot V^2 \cdot a \cdot f$$

$$\text{delay}(\tau) = \frac{2C \cdot V}{\beta(V - V_T)^2} \text{ from}: \left\{ \begin{array}{rcl} Q = I \cdot t & = & C \cdot V \\ I & = & \frac{C \cdot V}{t} \\ \beta(V_{gs} - V_T)^2 & = & \frac{2C \cdot V}{t} \end{array} \right\}$$

To reduce both the power and the delay, capacitance C must be reduced. From the previous two equations, the $\tau \cdot D$ product will be equal to:

$$\tau \cdot D = b \cdot \frac{V^3}{(V - V_T)^2}$$

where b is a constant.

The minimum will exist for $\frac{\delta \tau D}{\delta V} = 0$, which results in: $V = 3 \cdot V_T$. Thus, when a ratio of three is used between the supply voltage and the threshold voltage, the process should allow for optimum performance.

In the above derivation, however, the simple equations do not keep track with second- and third-order effects (e.g. small channel effects). Therefore, and because of some other practical reasons, a factor of five is usually used.

8.4 Design options for low power

As shown in table 8.1, we can reduce the dynamic, the static and the short-circuit dissipation by taking measures in the design. Because the measures for the latter two are clear and compact, we start with these two first.

8.4.1 Reduction of P_{short} by design measures

During an input transition at a CMOS logic gate, there may be a temporary current path from supply to ground. The resulting short-circuit

power dissipation can be relatively high if no attention has been paid to this [4]. Consider the example of figure 8.5.

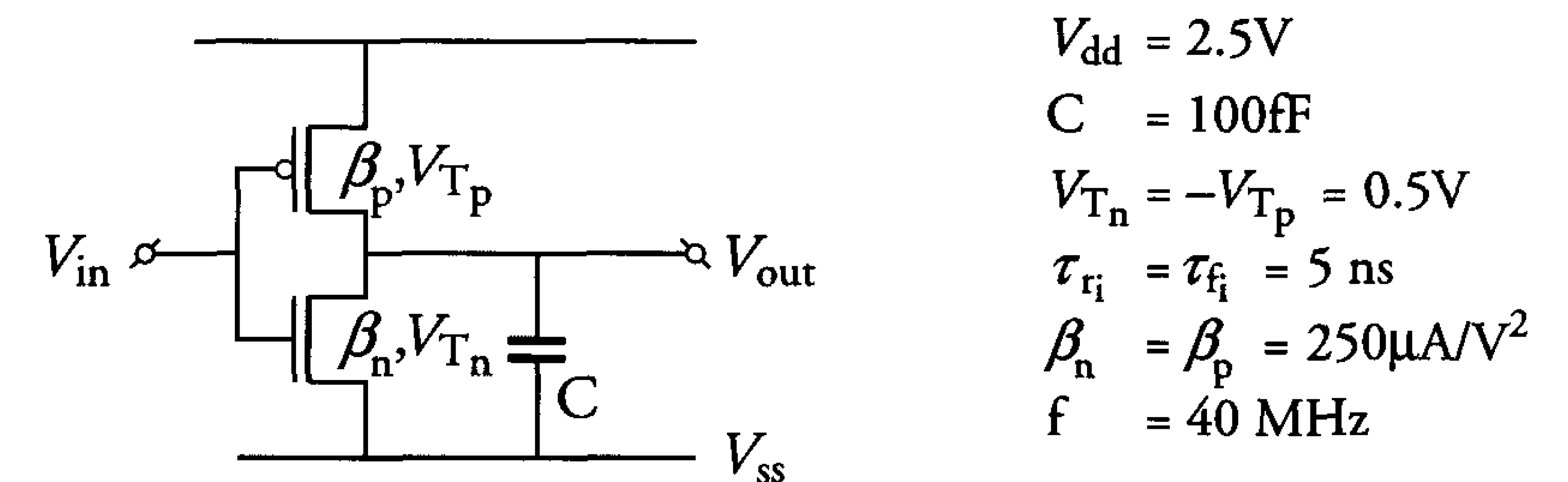


Figure 8.5: *Inverter example to illustrate the level of short-ciruit power dissipation*

With these numbers, the dynamic power dissipation becomes:

$$P_{dyn} = C \cdot V^2 \cdot a \cdot f = 25\,\mu\text{W}$$

and the short-circuit power dissipation becomes [4]:

$$P_{short} = \frac{\beta}{12} \cdot (V_{dd} - 2V_T)^3 \cdot \frac{\tau}{T} = 13.3\,\mu\text{W}$$

Conclusion: either τ_f and τ_r on the inputs are much too large or the β of the pMOS and nMOS transistors must be reduced. For CMOS drivers (internal, clock and output drivers), this short-circuit power can be minimised when τ_f and τ_r are equal on all nodes. This requires tapering of the inverters in such a driver, see figure 8.6.

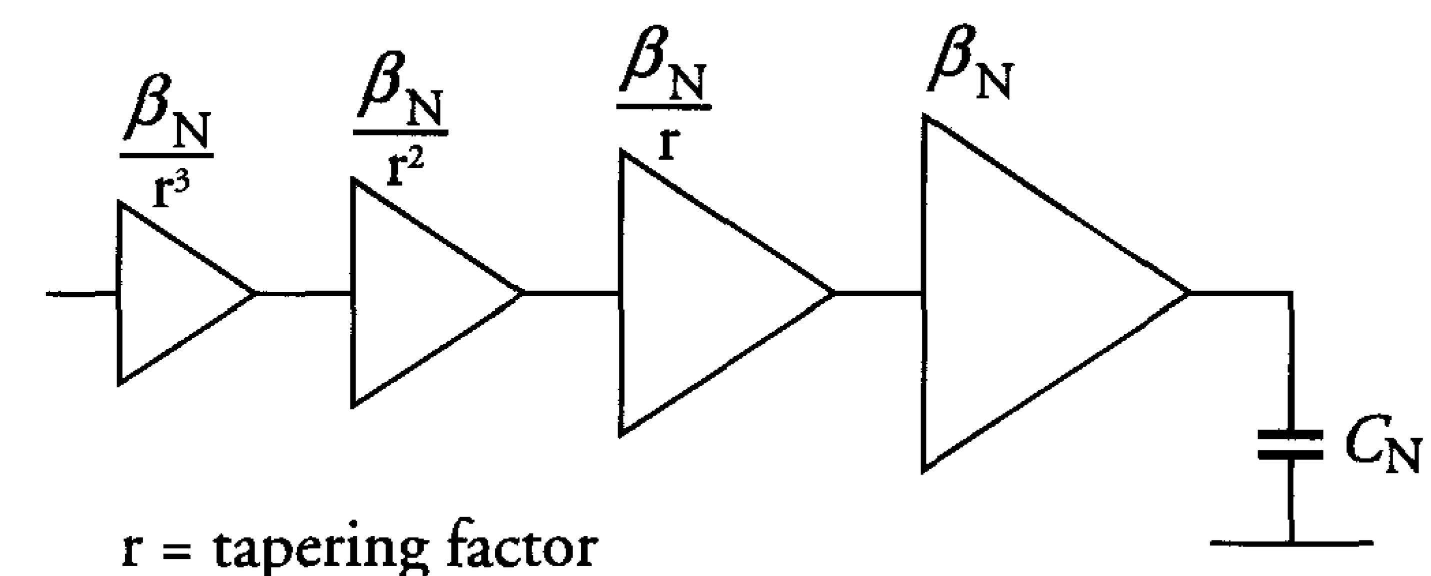


Figure 8.6: *Inverter chain with tapering factor*

A tapering factor between 8 to 16 (section 4.3.2) will usually result in a minimum short-circuit dissipation, which will then be less than 20% of the total dissipation (10% in most cases) [4].

An important remark to be made here is that the pMOS and the nMOS transistors can never conduct simultaneously during a transient when $V_{dd} < V_{T_n} + |V_{T_p}|$, thereby eliminating the short-circuit dissipation completely.

8.4.2 Reduction/elimination of P_{stat} by design measures

In complex logic gates which require many pMOS transistors in series (five or more input NOR gates, address decoder in memories, etc.), pseudo-nMOS solutions are sometimes applied, see figure 8.7. When the output of such a gate is low, there is a continuous static current from V_{dd} to ground.

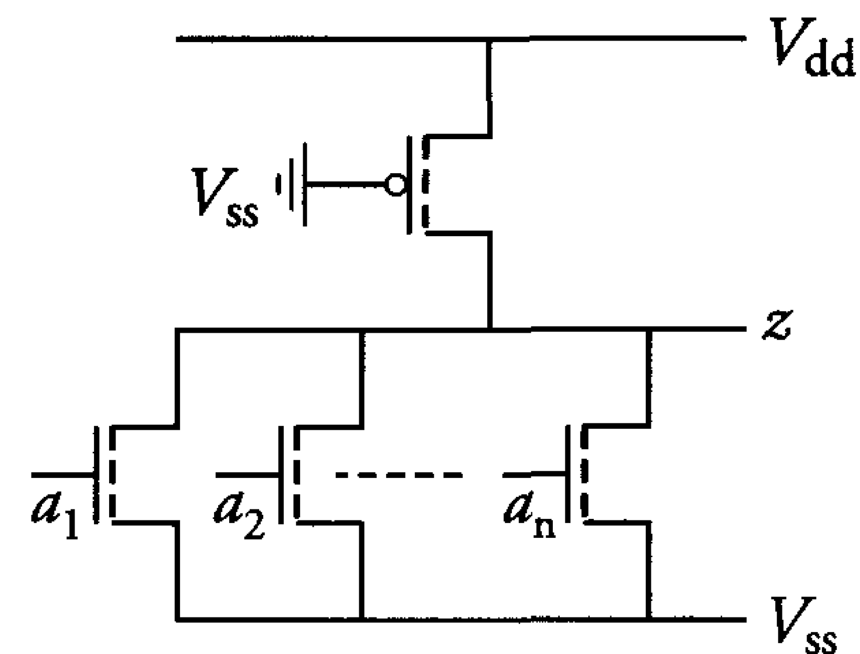


Figure 8.7: *n-Input pseudo-nMOS NOR gate*

For low-power applications, this is not an efficient way of implementation. In this case, the power can be reduced by replacing the grounded pMOSt by a clocked pMOSt. This may reduce the power by a factor equal to the clock duty cycle. For real low-power dissipation, this is not a good solution, because a *pseudo-nMOS logic* gate consumes about 10 to 20 times that of a full CMOS realisation. Therefore, to eliminate static power consumption, no pseudo-nMOS should be used at all.

8.4.3 Reduction of P_{dyn} by design measures

The dynamic dissipation was expressed by:

$$P_{dyn} = C \cdot V^2 \cdot a \cdot f$$

By means of design techniques, we are able to influence all parameters in this expression. We will therefore present several alternative measures for each parameter to reduce its contribution to the power consumption.

Power supply (V) reduction

A lower voltage generally means less performance and less chance for latch-up. Let's assume we have the following circuit on a chip, see figure 8.8.

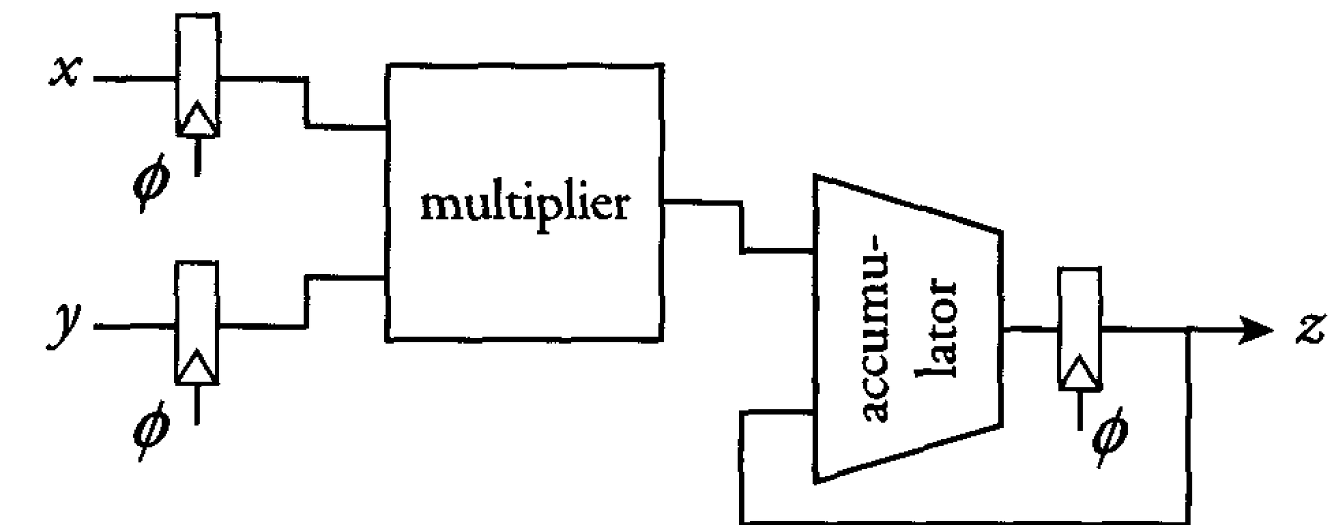


Figure 8.8: *A basic data path*

The total propagation delay time through the logic is equal to the sum of the propagation delay of the multiplier and accumulator. This total propagation delay determines the minimum duration T of the clock period. If we double this clock period, the propagation delay is allowed to be twice that of the original circuit. To achieve this doubling, we may reduce the supply voltage from 2.5 V to 1.8 V, for example. However, if the throughput is to be retained, two of the circuits can be connected in parallel and their inputs and outputs multiplexed (parallelism) or additional latches can be placed in between the logic functions to shorten the critical delay paths between two latches (pipelining).

A) *Parallelism*
Figure 8.9 shows a parallel implementation of the circuit. As a result of demultiplexing and multiplexing the signals, the same performance can be achieved as in the original circuit of figure 8.8, which runs at twice the clock frequency.

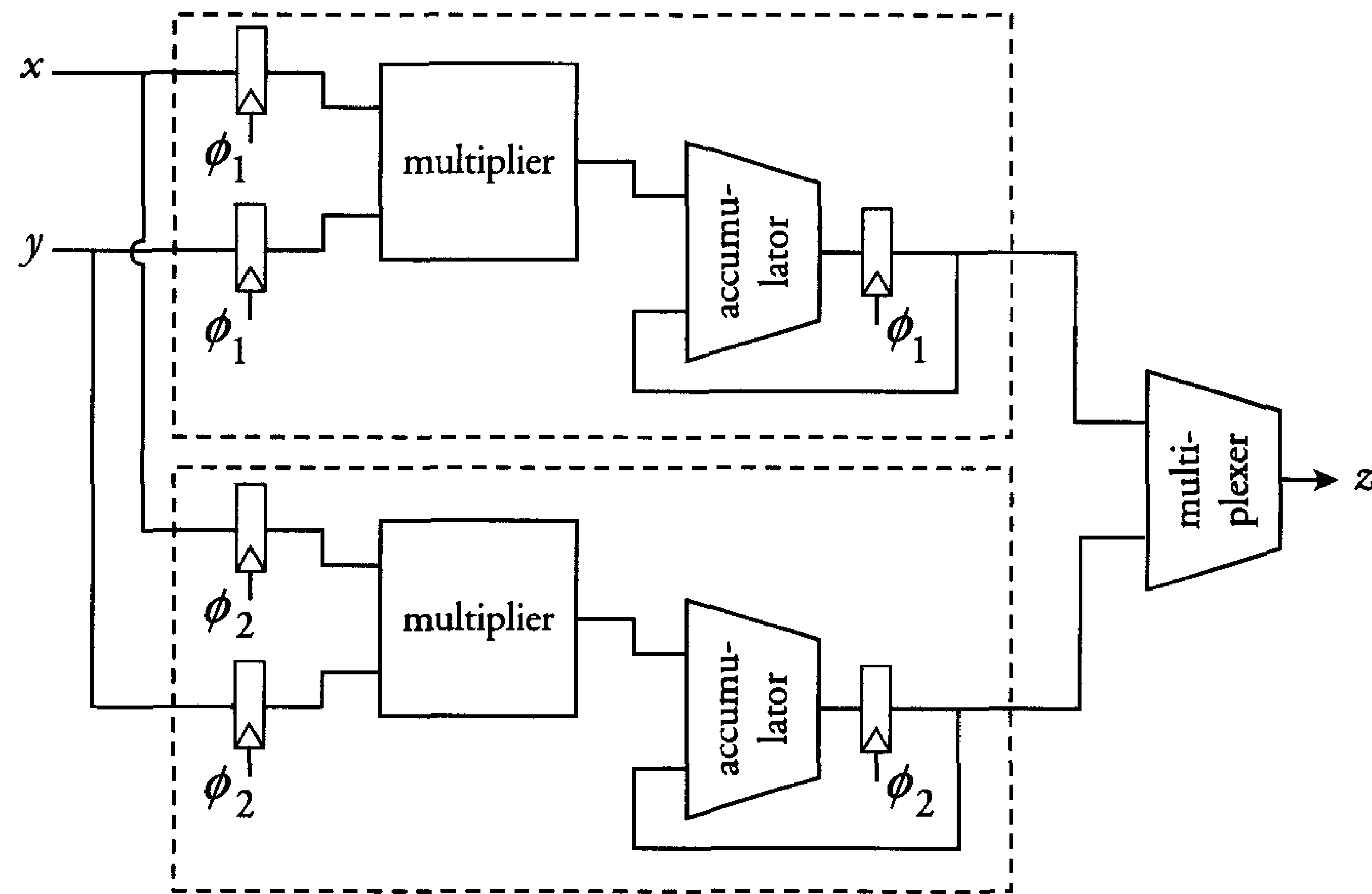


Figure 8.9: *Parallel implementation*

When we include multiplexers and additional wiring, this parallel architecture results in an increase of the total capacitance to be switched by a factor of about 2.25. The power dissipation comparison for the circuits of figure 8.8 and the parallel implementation in figure 8.9 then yields:

$$\begin{aligned} P_{\text{dyn}}(\text{basic data path}) &= C \cdot V^2 \cdot a \cdot f = P_{\text{ref}} \\ P_{\text{dyn}}(\text{parallel data path}) &= (2.25C) \cdot (\frac{1.8}{2.5}V)^2 \cdot a \cdot \frac{f}{2} = 0.58 \cdot P_{\text{ref}} \end{aligned}$$

Thus, the parallel implementation of the data path results in a power reduction of a factor of about 1.7, however at the cost of area overhead of more than a factor of two. This is sometimes not allowed, especially in the cheap high volume TV (and PC) markets. Another way to maintain performance at a reduced power supply voltage is pipelining.

B) *Pipelining*
In figure 8.8, the critical path is equal to:

$$T_{\text{crit}} = T_{\text{mpy}} + T_{\text{acc}} \Rightarrow f_{\text{max}_{\text{ref}}}$$

If the propagation delays of the multiplier and the accumulator are about the same, we can put a pipeline in between. Figure 8.10 shows the circuit with the additional pipelines.

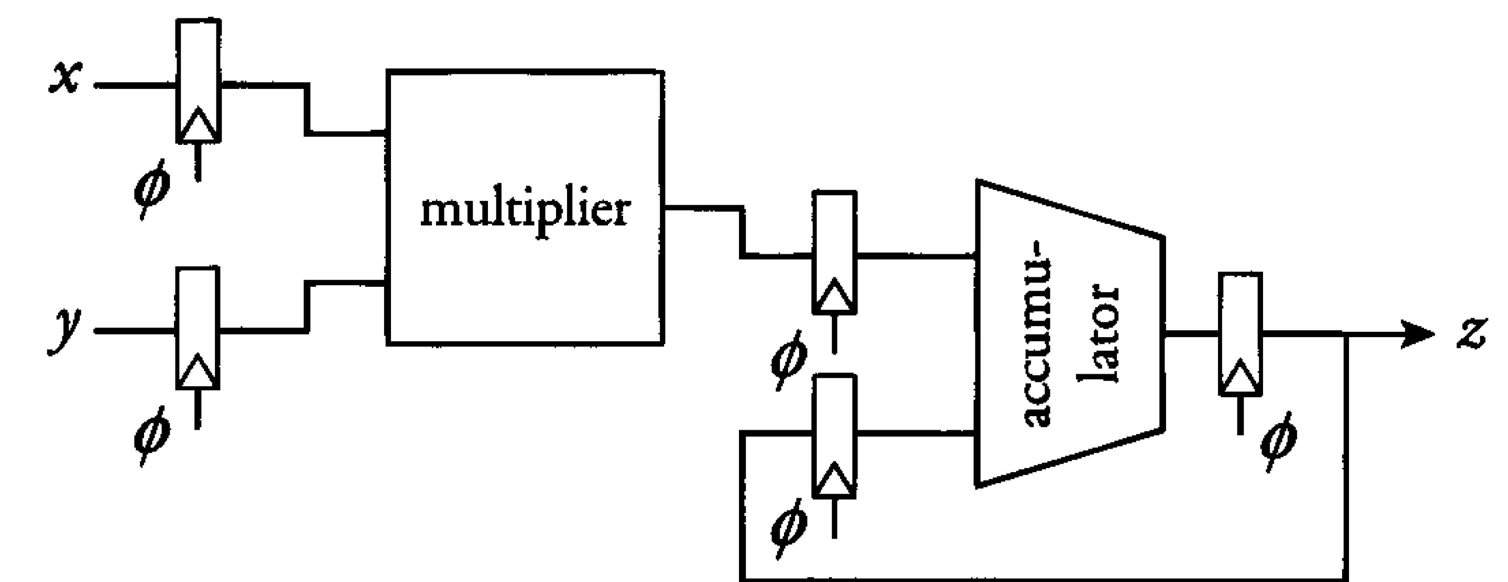


Figure 8.10: *Pipelined implementation*

Now, the critical path is:

$$\begin{aligned} T_{\text{crit}} = \max[\text{T}_{\text{mpy}}, \text{T}_{\text{acc}}] \quad &\Rightarrow \quad f_{\text{max}} > f_{\text{max}_{\text{ref}}} \\ \text{if } \text{T}_{\text{mpy}} \approx \text{T}_{\text{acc}} \quad &\Rightarrow \quad f_{\text{max}} \approx 2 \cdot f_{\text{max}_{\text{ref}}} \end{aligned}$$

The additional pipeline allows a frequency which is twice as high. Therefore, the voltage may reduce to about 1.8 V to maintain the same frequency again. As a result of the additional pipelines and multiplexer, the area increase will be about 20%. Comparing this pipelined architecture with the original one leads to the following result:

$$\begin{aligned} P_{\text{dyn}}(\text{basic data path}) &= C \cdot V^2 \cdot a \cdot f = P_{\text{ref}} \\ P_{\text{dyn}}(\text{pipelined data path}) &= (1.2C) \cdot (\frac{1.8}{2.5}V)^2 \cdot a \cdot f = 0.62 \cdot P_{\text{ref}} \end{aligned}$$

Thus, with only an area penalty of 20%, we almost get the same result as with parallelism. An alternative is the combination of parallelism and pipelining.

C) *Combination of parallelism and pipelining*
By using both parallelism and pipelining techniques, the critical path is reduced by a factor of four. This also results in a reduction of the requirement on speed by a factor of four. To achieve this speed requirement, the voltage can be reduced to only $0.7 \cdot V_{\text{ref}}$. Comparing this technique with the original one leads to:

$$P_{\text{dyn}}(\text{basic data path}) = C \cdot V^2 \cdot a \cdot f = P_{\text{ref}}$$
$$P_{\text{dyn}}(\text{parallel/pipelined}) = (2.25 \cdot 1.2C) \cdot (0.55 \cdot V)^2 \cdot a \cdot \frac{f}{2} = 0.4 \cdot P_{\text{ref}}$$

Therefore, by using this combination of techniques, we can achieve an improvement (reduction) in power of a factor of 3. However, this will lead to an increase in chip area by a factor of about 2.7. The choice between area and power is a matter of priority. However, a designer does not usually have the freedom to select the supply voltage level: he chooses a technology and then the supply voltage is "fixed": for a 0.25 μm CMOS process, the supply voltage is generally fixed at 2.5 V, because the library cells are characterised for this voltage.

D) *Real low-voltage design*
A real low-voltage design may be connected to a single battery supply of 0.9 to 1.1 V. This gives a reduction in power dissipation according to:

$$P_{\text{dyn}}(2.5\,\text{V}) = C \cdot 6.25 \cdot a \cdot f$$
$$P_{\text{dyn}}(1\,\text{V}) = C \cdot 1 \cdot a \cdot f$$

This results in an improvement of more than a factor of 6. However, V_{T}'s are often between 0.3-0.6 V in absolute values. This means that the supply voltage, V_{dd} can be less than:

$$V_{\text{dd}} < V_{\text{T}_\text{n}} + |V_{\text{T}_\text{p}}|$$

This results in hysteresis in the inverter characteristic of a CMOS inverter, see figure 8.11.

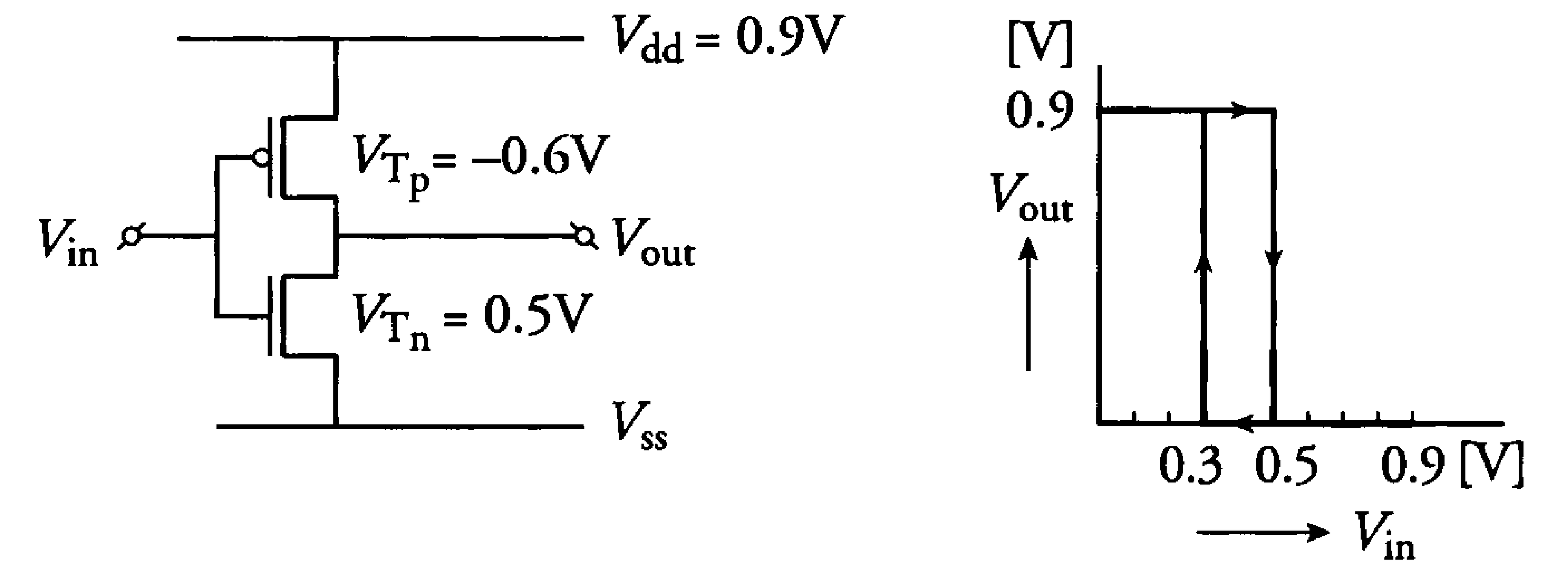


Figure 8.11: *CMOS inverter + characteristic, showing hysteresis when* $V_{\text{dd}} < V_{\text{T}_\text{n}} + |V_{\text{T}_\text{p}}|$

In this example, the following values are assumed:
$V_{\text{dd}} = 0.9\,\text{V}$,
$V_{\text{T}_\text{n}} = 0.5\,\text{V}$ and
$V_{\text{T}_\text{p}} = -0.6\,\text{V}$.

The operation of the inverter is as follows, when switching V_{in} from 0 to V_{dd} and back again:

- When $V_{\text{in}} = 0\,\text{V}$, the pMOS transistor is on and the nMOS transistor is off; the output is at the high level (0.9 V).
- When $0.3\,\text{V} < V_{\text{in}} < 0.5\,\text{V}$, both the pMOS and nMOS transistors are off, so the output remains (floating) at the high level.
- At V_{in}=0.5 V, the nMOS transistor switches on and the output V_{out} immediately goes to 0 V because the pMOS transistor remains off.
- When $0.5\,\text{V} < V_{\text{in}} < 0.9\,\text{V}$, the nMOS transistor remains on and the output remains at 0 V.
- When we switch V_{in} back to 0 V again, when $0.3\,\text{V} < V_{\text{in}} < 0.5\,\text{V}$, both the pMOS and the nMOS transistors are off. This means that the output remains at 0 V, but floating (high impedance state).
- When V_{in} becomes equal to 0.3 V, the pMOS transistor switches on and the output switches to 0.9 V.
- Finally, when $0\,\text{V} < V_{\text{in}} < 0.3\,\text{V}$, the pMOS transistor remains on and the output remains at 0.9 V.

Although these kinds of circuits ($V_{dd} < V_{T_n} + |V_{T_p}|$) are relatively slow, they have been used for a long time in battery-operated products, e.g. watches. One advantage of these circuits is that a short-circuit current can never flow, because one transistor always switches off before the other one switches on. Therefore, there is no short-circuit dissipation at all. Not every library is suited for low-voltage operation. This means that a new low-voltage library must be developed and characterised, including a RAM, a ROM and other generators. Moreover, because of the low-voltage supply, the threshold voltage (V_T) must be controlled very accurately. ($|V_T| \simeq 0.3...0.6$ V with $V_{T_{spread}} \approx 150$ mV. The optimum V_{dd} is then close to $V_{dd} = 1.5$ V.

E) *Voltage regulators*
Generally, ICs also contain low performance parts which could actually run at lower supply voltages. These can be supplied externally, or generated on the chip by means of a voltage regulator [5], see figure 8.12.

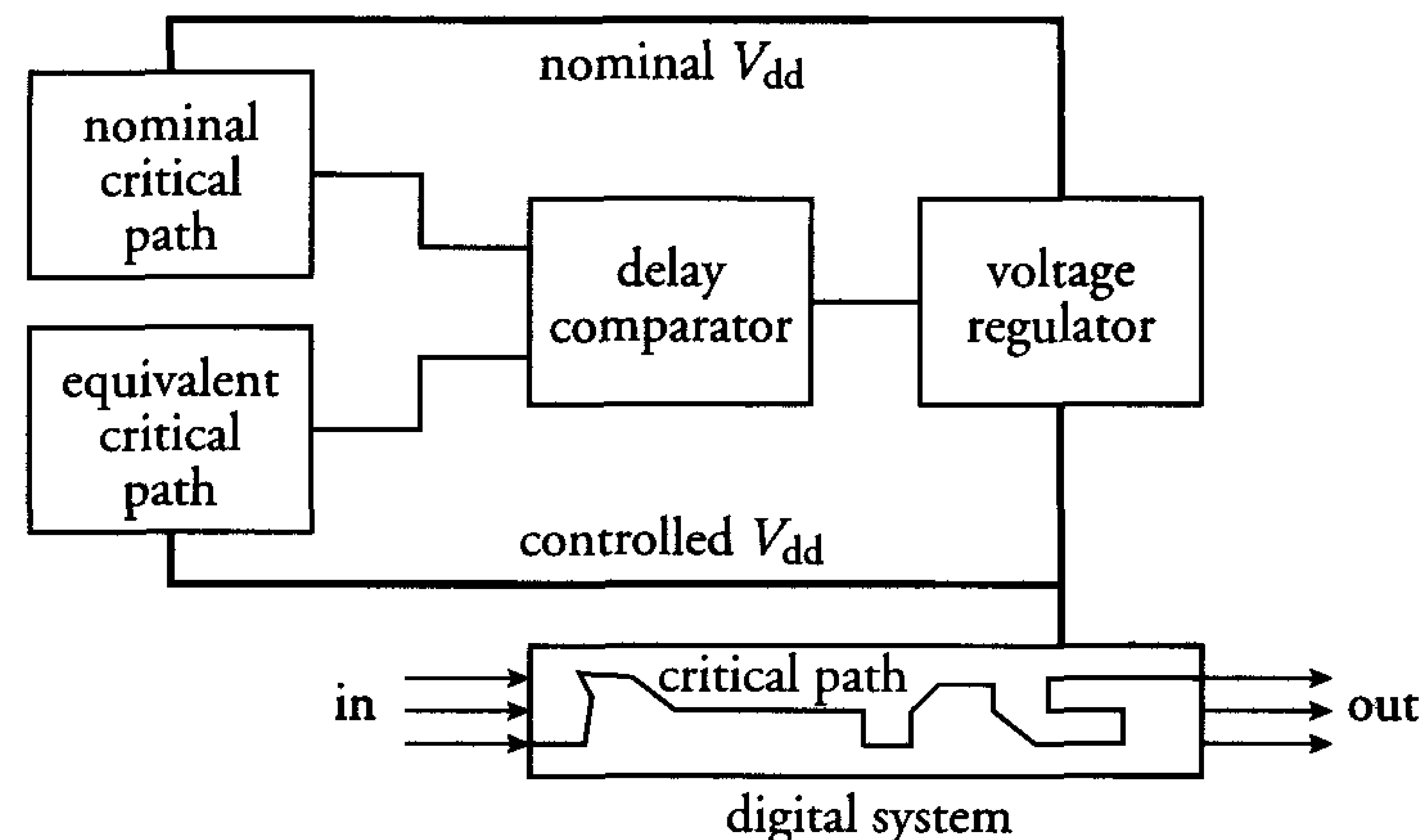


Figure 8.12: *Example of voltage regulator principle*

If such a voltage regulator is used, attention should always be paid to its power efficiency. A better alternative is to run the chip at the lowest required voltage and perform a voltage-up conversion only for the high-performance circuit parts. Such voltage-up converters are used in single cell hearing aids, for example. One can also use *DC-DC converters.* Here too, the power efficiency is an important factor

in the use of such circuits. At the moment, this efficiency is in the ninety percent range.

F) *Reduced voltage swing*
Currently, the bus widths, both on chip and off chip, are tending to grow to 32, 64 and even to 128 bits. This means that the number of simultaneously switching buses and/or outputs has already increased dramatically and the number will continue to increase. If the power dissipation becomes high with respect to other parts of the chip, then a lowering of the voltage swing on these buses (or outputs) has to be considered. As an example, we take a video signal processor, which has 72 outputs loaded with 30 pF each, and which runs at a clock frequency of 54 MHz. The total maximum dissipation at V_{dd}= 2.5 V when these outputs switch simultaneously is equal to:

$$P_{dyn} = C \cdot V^2 \cdot a \cdot f = 72 \cdot 30 \cdot 10^{-12} \cdot 6.25 \cdot \frac{1}{2} \cdot 54 \cdot 10^6 = 0.4\,\text{W}$$

If we could lower the output swing to 1 V, this power dissipation would reduce to 70 mW. Reduced voltage swing techniques are frequently used to reduce the power dissipation of large 32-bit (or 64-bit) processors.

Capacitance reduction

The total capacitance to be switched on an IC can be reduced or limited at two levels: at system level and at chip level. The decisions taken at system level might have more effect on the IC power than at chip level. This is because a different architecture for an ALU/multiplier or for a filter structure can have more area consequences for the total hardware. This is shown in the following example:

A) *System level*
Suppose we have to perform the following algorithm:

$$y(n) = \sum_{m=0}^{k-1} x(n-m)$$

A possible hardware implementation is shown in figure 8.13.

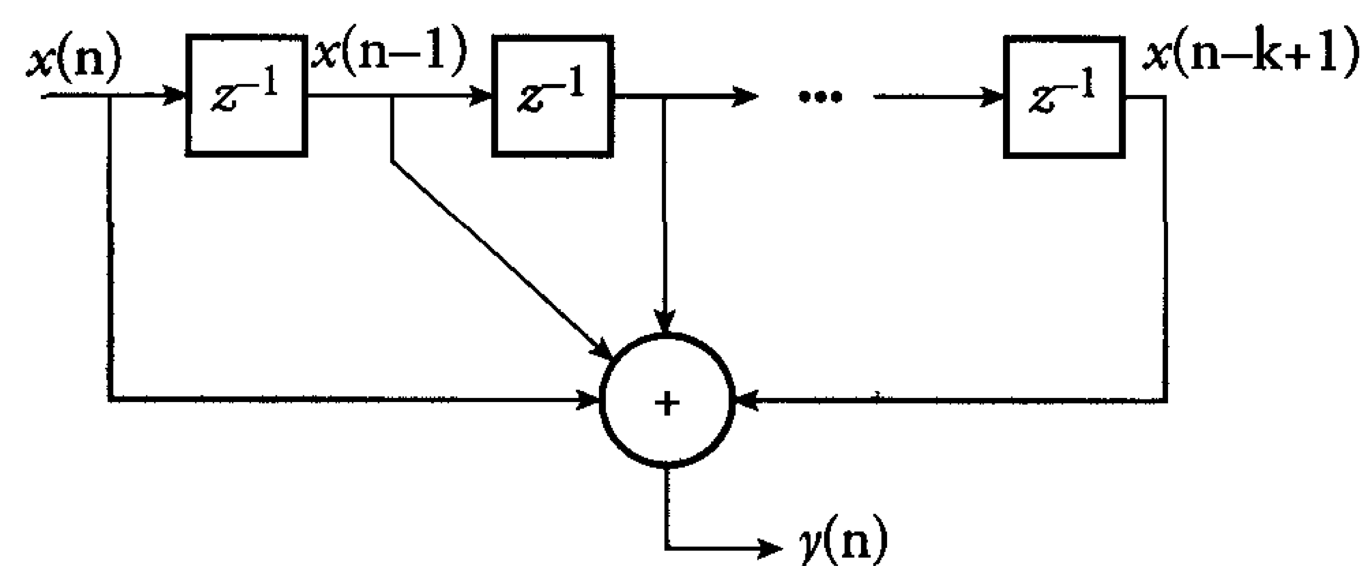


Figure 8.13: *Digital realisation of the running sum algorithm*

When m is large, many additions have to be performed. Here the hardware configuration will contain a lot of full adders to generate the sum and carry functions. The data has to ripple through a large number of full adders, leading to long propagation times and a limited clock frequency. A high-performance implementation would even require additional pipelines and/or carry-look-ahead techniques to improve speed. With regard to the power consumption, this implementation is very disadvantageous. Figure 8.14 shows an alternative recursive realisation:

$$y(n) = y(n-1) + x(n) - x(n-k)$$

Although it consists of two adders, each adder here has only two inputs, which means that much less hardware is involved.

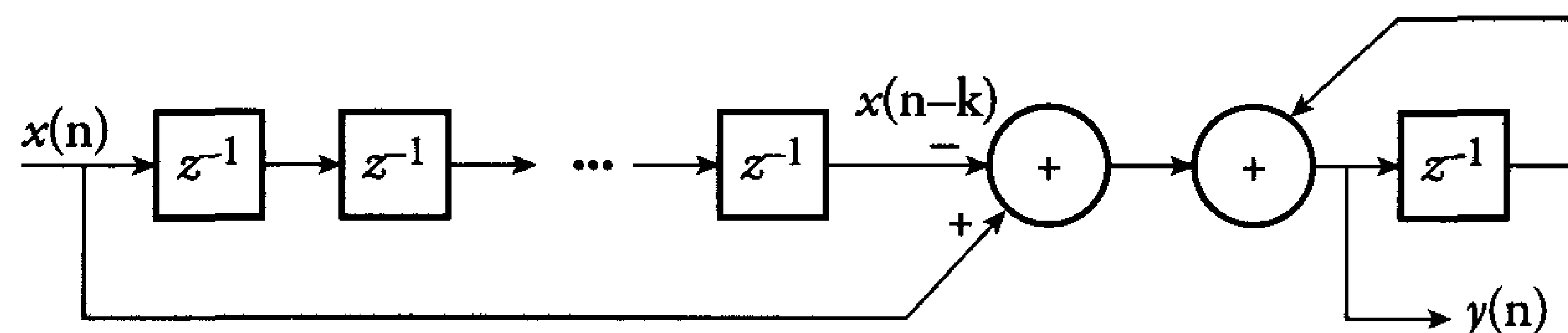


Figure 8.14: *Recursive realisation of the running sum algorithm*

From this example, we can conclude that the algorithm level is at least as important as the chip level for a low-power realisation of a certain function.

B) *Chip level*

At chip level, there are many options for lowering the power consumption by capacitance reduction. This ranges from libraries, via tools and circuit techniques, to layout implementation.

- Optimised libraries

  In many cases, low power means that a low voltage must be applied. This requires library cells with a low gate complexity (maximum three inputs). These cells suffer from less body effect and show a better performance than cells with higher complexity. Most libraries are designed for high performance. They contain relatively large transistors which consume power accordingly. Using these libraries for a low-power design is an overkill in both power and area. In a real low-power library, the transistor and cell sizes must be limited, such that both the parasitic junction capacitances of source and drain regions and the total interconnect lengths after routing will also be reduced. Source and drain regions can be reduced by adapting a very regular layout style.

  Flip-flops are probably the most frequently used cells of a library. In many synchronous chips, ten to fifty percent of the total layout area is often occupied by flip-flops. They therefore play a dominant role in the performance, the area, the robustness and the power consumption of a chip. It is clear that the flip-flops should be designed for low power, not only for their internal power consumption, but also for the clock driver power consumption. A low fan-in for the clock input combined with better clock skew tolerance (more output delay) allows smaller clock driver circuits, thereby reducing both power consumption and current peaks.

- Pass-transistor logic (transfer gate; pass gate; transmission gate)

  This logic already existed in the nMOS era. The most efficient circuits to be implemented in pass-transistor logic are multiplexers, half adder and full adder cells. The basic difference between this logic and conventional static CMOS logic is that a pass-transistor logic gate also has inputs on the source/drain terminals of the transistors. A major disadvantage of nMOS pass-transistor logic is the threshold voltage loss ($V_{\text{out}} = V_{\text{dd}} - V_{\text{T}_\text{n}}$) at high output level. When such a signal is input to a CMOS inverter, a leakage current flows in this inverter when $V_{\text{T}_\text{n}} \geq |V_{\text{T}_\text{p}}|$.

nMOS pass-transistor logic will thus not be an alternative for low-power design. For different reasons it is usually not feasible to control the threshold voltages (i.e. $V_{T_n} \geq |V_{T_p}|$) at the technology level. To compensate for the threshold voltage loss and for other disadvantages of nMOS pass-transistor logic, several pass-transistor logic styles have been presented. The most important ones will now briefly be discussed.

Complementary Pass-Transistor Logic (CPL) [6]
A CPL gate (figure 8.15) basically consists of two nMOS logic circuits, two small pMOS transistors for level restoration and two inverters for generating complementary outputs. Without the cross-coupled pMOS pull-up transistors, CPL would also show the same problems as the above-discussed nMOS pass-transistor logic.

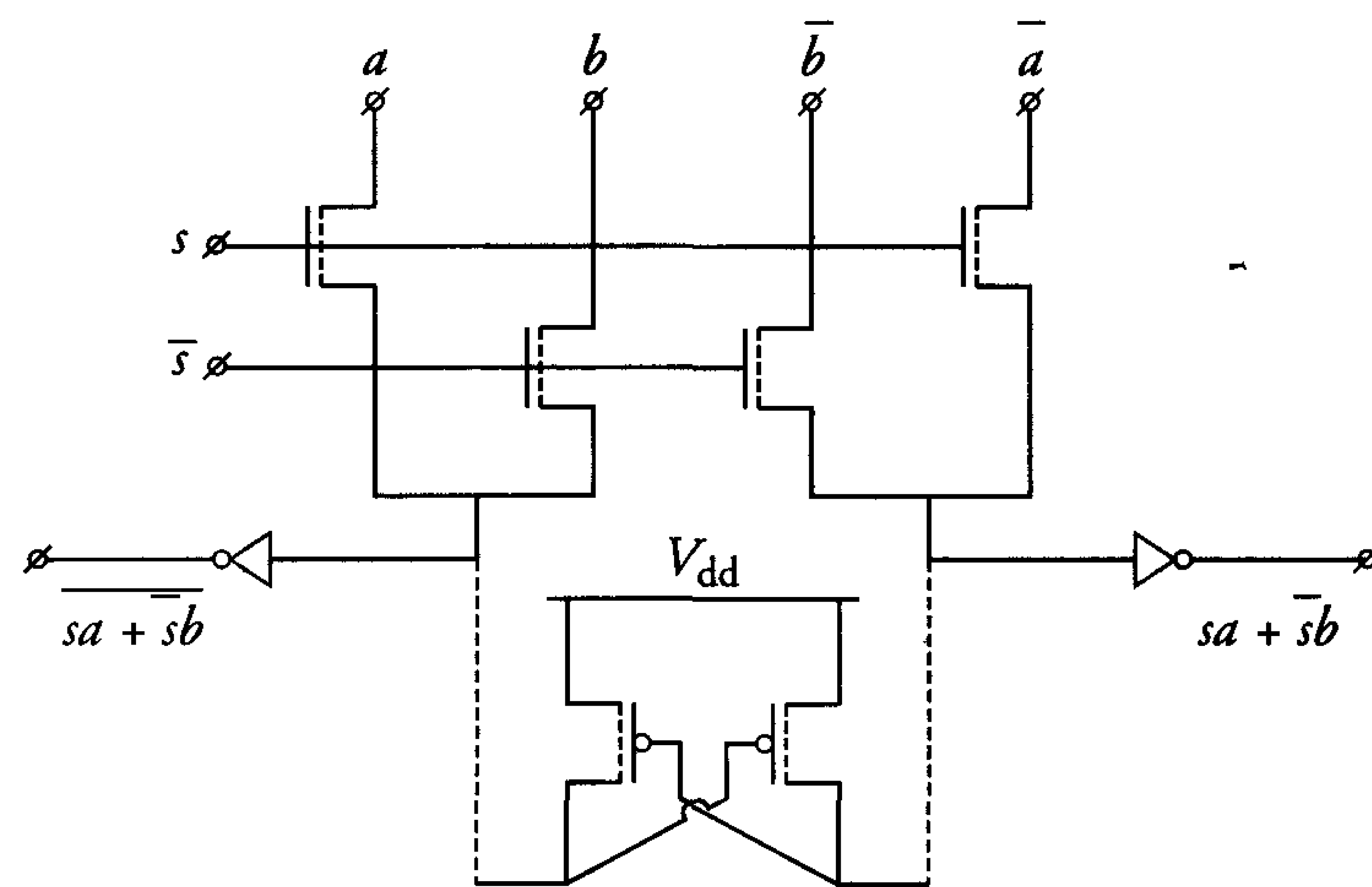


Figure 8.15: *2-input multiplexer in CPL*

Because of the complementary logic circuits, the fan-in and the complexity of a CPL gate approaches that of a conventional CMOS gate. Because of the availability and necessity of the complementary signals, much more routing area is required. Moreover, simple logic functions require a relatively high transistor count.

Double Pass-Transistor Logic (DPL) [7]
A DPL logic gate uses both nMOS and pMOS logic circuits in parallel, providing full swing at the outputs, see figure 8.16.

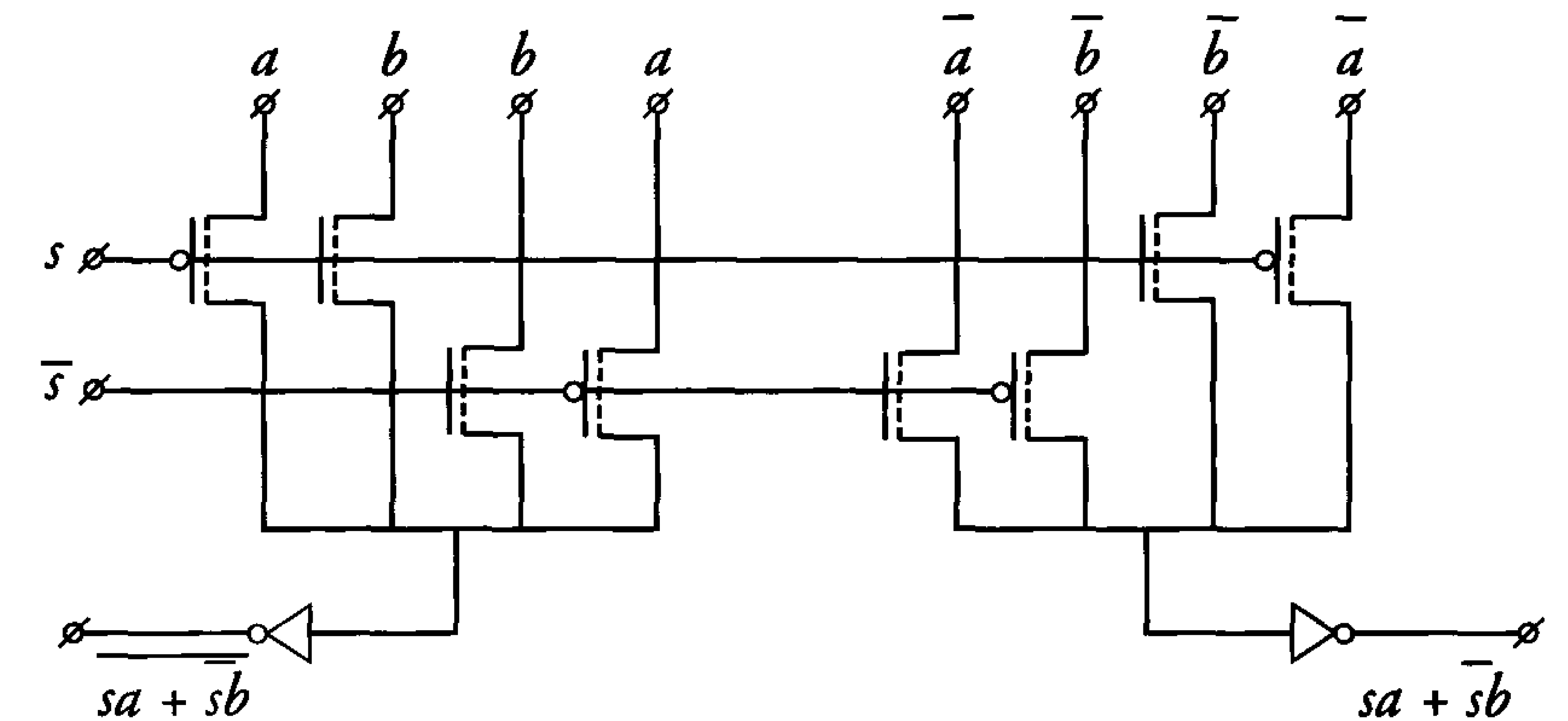


Figure 8.16: *2-input multiplexer in DPL*

Because of the necessity of complementary signals, this logic style has the same routing complexity as CPL. Although it needs no swing restore circuit, it contains more transistors, particularly in complex gates, and has a higher fan-in than CPL. This usually requires more switching power. It is therefore less attractive than other pass-transistor logic and standard CMOS logic.

Other styles of pass-transistor logic
There are several other styles of pass-transistor logic. One, *the Swing Restored Pass-Transistor Logic* (SRPL; [8]) is derived from CPL. Here, the output inverters are mutually cross-coupled (compare figure 8.15) and must be overwritten by the pass-transistor network. This makes this logic less robust for general implementation. In *Lean Integration with Pass-Transistor* (LEAP; [9]), single-rail pass-transistor logic is used. This basically contains an nMOS logic network and a level restore circuit, consisting of an inverter and a feedback pMOS pull-up transistor. This is slower than CPL. At supply voltages of $V_{dd} < 2V_{T_n}$, this pass-transistor style is no longer applicable because the output inverter can no longer be turned on.

Finally, new styles of pass-transistor logic are being invented continuously (DPTL[10]; EEPL[11]; PPL[12]). However, many are derived from previous styles with only minor changes. Whatever style of pass-transistor logic will be invented yet, they will all have the same disadvantages: they will either suffer from threshold voltage loss and need a level restore circuit or they will need the double rail approach (complementary inputs and outputs).

Conclusions

Although different pass-transistor logic families are presented in literature, showing better performance in terms of power delay products than conventional CMOS logic, the opposite is also published [13].

Initially, pass-transistor logic showed equal functionality with low transistor count. However, with reduced voltages, complex ICs and low-power focus, this advantage has been undone by the necessity of a level restore circuit and/or dual rail implementation. Except for half and full adder functions, conventional CMOS circuits perform better than any pass-transistor style where both power and robustness are concerned. As a result of increasing process variations and extending application environments, the robustness will play an especially dominant role in the development of deep-submicron (standard) cell libraries.

- Synthesize logic functions into larger cells.

  Usually, logic functions are mapped onto library cells. This, however, is rather inefficient in terms of area and power. The full-adder function might serve as a good example, where s is the sum function and c represents the carry:

$$
\begin{aligned}
s &= a\bar{b}\bar{c} + \bar{a}\bar{b}c + \bar{a}b\bar{c} + abc \\
c &= ab + ac + bc
\end{aligned}
$$

  In a standard cell library without a full-adder cell, the sum function would require four 3-input AND functions and one 4-input OR. With a dedicated full-adder library cell, the area will be roughly halved. Generally, a cell compiler, capable of optimising complex functions and creating logic gates, would be a good

tool for optimising both area and speed. However, good characterisation tools must then also be available to generate accurate timing views of these compiled cells.

- Use optimised synthesis tools.

  Good tools are required for an optimum mapping of complex logic functions onto the library cells. These tools must include reasonably accurate timing models. Usually, the less hardware is used, the less power will be consumed.

- Use optimised place & route tools.

  Many current CAD tools for place & route are area or performance driven. Part(s) of the circuits can have different weights for high performance. These require priority in the place & route process. With a focus at low power, power driven (activity/capacitance) place & route tools are required, resulting in minimum wire lengths.

- Use custom design, if necessary.

  Reduction of the interconnection lengths can be achieved by different layout styles. Especially cell abutment is a way to optimise data paths, bit slice layouts and multipliers, etc. Custom design must only be applied if the additional design time can be retrieved. Practically speaking, this only holds for high volume chips, or for chips with very tight power specifications, which cannot be achieved with other design styles.

- Make an optimum floor plan.

  Although this sounds very commonplace, it is not self-evident. During floor planning, the focus should be on wasting less area and on reducing bus and other global interconnections.

- Optimise the total clock network.

  Clock signals run globally over the chip and usually switch at the highest frequency (clock frequency f; data frequency $< f/2$). As discussed, the number of flip-flops and their properties are a dominant factor in the total clock network. The flip-flops should be optimised for low fan-in and a better clock skew tolerance so that smaller clock drivers could be used. Section 9.3.1 presents a robust flip-flop, which is also very well suited for low-power designs.

- Use well-balanced clock trees.

  Balanced clock trees are those in which drivers and loads are

tuned to one another, such that equal clock delays are obtained, anywhere in the chip. This reduces the clock skew, which allows for smaller clock drivers.

- Dynamic versus static CMOS.
  Chapter 4 presents implementations of static and dynamic CMOS logic gates. With respect to capacitance, a dynamic CMOS gate generally has less fan-in capacitance. This is because the function is usually only realised in an nMOS network, while the pMOSt only acts as a (switched) load. Because every gate is clocked, we get very large clock loads. Moreover, as a result of the precharging mechanism, the average activity in a dynamic gate is higher than its static counterpart. A more detailed look into the activity of static and dynamic CMOS logic is presented in the following paragraph.
- Memory design.
  To reduce the total capacitance to be switched in a memory, the memory can be divided into blocks (block select), such that they can be selectively activated (precharge plus read/write). Divided word lines means that less capacitance is switched during each word line selection. Wider words (64 bits instead of 32 bits) reduce the addressing and selection circuit overhead per bit.
  The precharge operation can be optimised by selectively precharging the columns (only those to be read or written) instead of all simultaneously.

Reduction of switching activity

Most of the switching activity of a circuit is determined at the architectural and register transfer level (RTL). At the chip level, there are less alternatives for lowering the power consumption by reducing switching activity.

This paragraph presents several of these alternatives, starting at the architectural level.

A) *Architectural level*
Choices made at the architectural and RTL level heavily influence the performance, the area and the power consumption of a circuit. This subsection summarises the effect that these choices have on the activity of the circuit.

- Optimum binary word length.
  The word length must be not only optimum in terms of capacitance but also in terms of activity, which means that only that number of bits is used that is really required to perform a certain function.
- Bit serial versus bit parallel.
  Figure 8.17 gives two alternative implementations for a 16 by 16 bit multiplier: a bit serial iterative multiplier and an array multiplier.

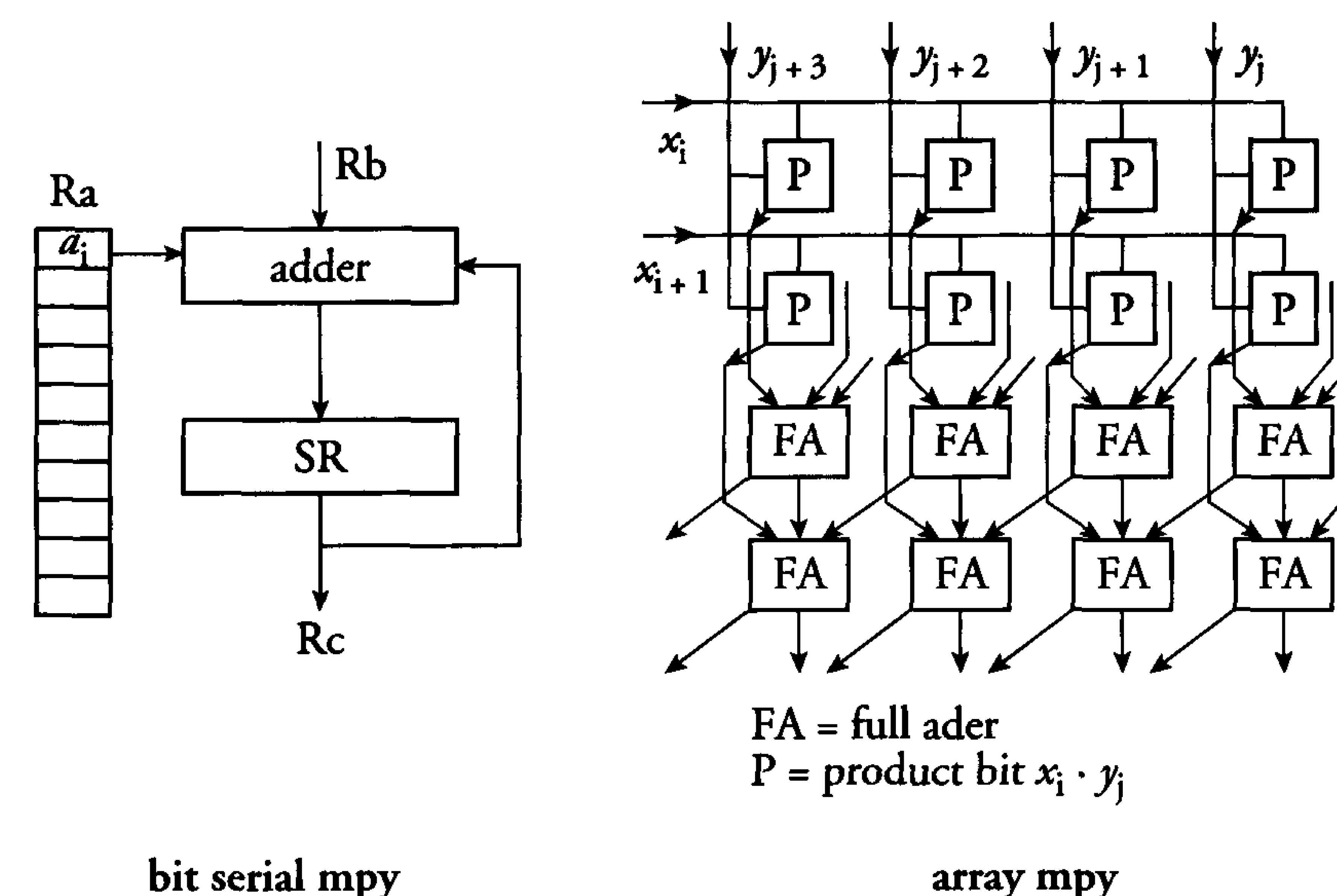


Figure 8.17: *Bit serial iterative and array multiplier*

The array multiplier only consists of logic that is really required for its function. In the bit serial approach, not only the required logic for multiplication is needed, but also the logic for additional control and registers. For a fair comparison, a complete multiplication must be taken. For the parallel multiplier, we have power*1 (period); for the bit serial one, we have power*16 (periods). From this example, we may conclude that a parallel implementation generally has less overhead than a bit serial one and will therefore consume less power.

- Optimise system power instead of chip power only.
  Complete systems use blocks such as DSP, A/D, D/A and memories, etc. As a result of the increasing communication bandwidth (data word length times frequency) of signals between these blocks, a lot of power would be wasted in the I/O circuit if each block was a separate chip. If possible, all functions should be on one chip. This will increase the chip power, and reduce the system power. A concentration of high-performance system parts and low performance system parts on different areas on one chip is attractive for power as well. The low performance parts could then run at lower frequencies, to save power.
- Number representation.
  The choice of the number representation can also have an effect on the power consumption, see also figure 8.18.

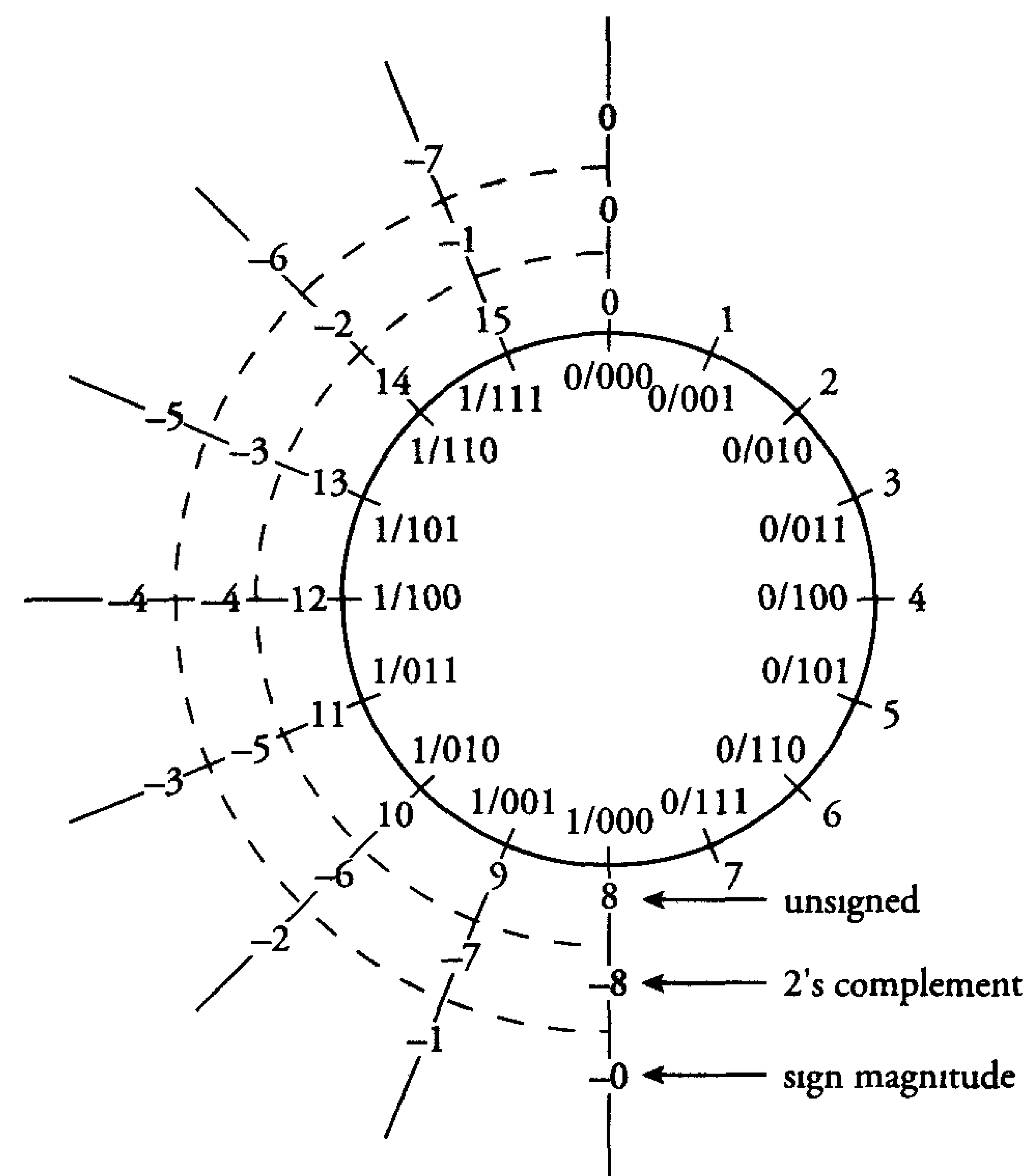


Figure 8.18: *Number representation diagram*

It is clear that unsigned code is only favourable for positive numbers. The most significant bit is then also used for magnitude representation. The two's complement notation shows problems (discontinuity) at the transition from $7 \rightarrow -8$. The diagram shows two discontinuities for the sign-magnitude notation: at the transition from $7 \rightarrow -0$ and also at the transition from $0 \rightarrow -7$. It is therefore more difficult when used in counters.

When small values are represented by many bits, the most significant bits in the two's complement notation adopt the value of the sign bit. If the signal is around zero, it will frequently switch from a positive to a negative value and vice versa. In the two's complement notation, a lot of sign bits will then toggle, while in the sign-magnitude notation only one sign bit will toggle, resulting in less power consumption. In the following example, the use of the two's complement notation and the sign-magnitude notation in adders and multipliers is compared.

Example:

8-bit adder/subtractor. The representation is shown in figure 8.19:

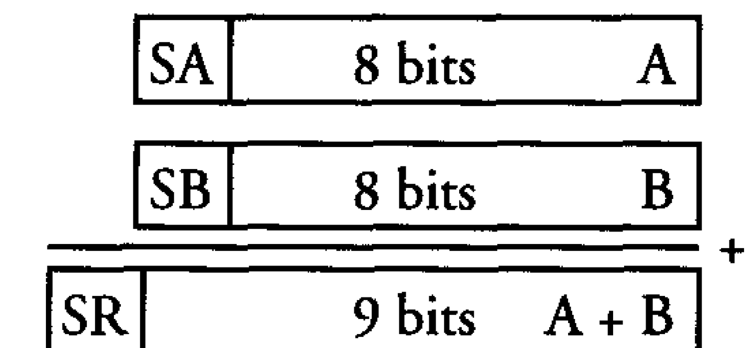


Figure 8.19: *Representation of an 8-bit adder/subtractor*

In the two's complement notation, the addition/subtraction operation does not give any problems. However, in the sign-magnitude notation, additional comparators must be used for a subtraction:

$$\begin{aligned} \text{if } A < B \;&\Rightarrow\; sum = B - A \\ \text{else} \;&\Rightarrow\; sum = A - B \end{aligned}$$

Implementation with synthesis and standard-cell place & route tools reveals a difference in silicon area of a factor of about three in favour of the two's complement notation.

Example:
Two's complement multiplication:

$$X = -X_{n-1} \cdot 2^{n-1} + \sum_{i=0}^{n-2} X_i \cdot 2^i$$

$$Y = \underbrace{-Y_{m-1} \cdot 2^{m-1}}_{\text{sign}} + \underbrace{\sum_{j=0}^{m-2} Y_i \cdot 2^j}_{\text{value}}$$

The result of multiplying X and Y is:

$$X \cdot Y = X_{n-1} \cdot Y_{m-1} \cdot 2^{n+m-2} + \sum_{0}^{n-2} \sum_{0}^{m-2} X_i Y_j 2^{i+j}$$
$$-(\sum_{j=0}^{m-2} X_{n-1} \cdot Y_j \cdot 2^{n-1+j} + \sum_{i=0}^{n=2} Y_{m-1} \cdot X_i \cdot 2^{m-1+i})$$

The realisation in an array multiplier requires the last two product terms to be skipped. A nice alternative is the Booth multiplier, in which half the number of full adders are replaced by multiplexers and where these two product terms are automatically skipped.

Example:
Sign-magnitude multiplication:

$$X = -1^{X_{n-1}} \cdot \sum_{i=0}^{n-2} X_i \cdot 2^i$$

$$Y = -1^{Y_{m-1}} \cdot \sum_{j=0}^{m-2} Y_j \cdot 2^j$$

and the product:

$$X \cdot Y = \underbrace{-1^{X_{n-1} \oplus Y_{m-1}}}_{\text{sign}} \sum_{i=0}^{n-2} \sum_{j=0}^{m-2} \underbrace{X_i \cdot Y_j \cdot 2^{1+j}}_{\text{magnitude}}$$

In this notation, the sign bit of the product is just a simple EXOR of the individual sign bits, while the magnitude is just the product of only positive numbers.

Conclusions on number representation

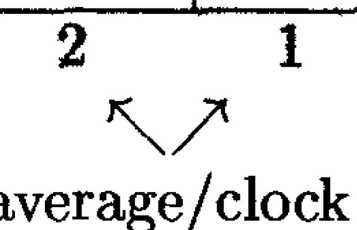
Although the sign-magnitude notation is convenient for multiplier implementation, the Booth algorithm array multiplier is more popular. Such a multiplier requires relatively little hardware and is thus suited for low power implementation.

The sign-magnitude notation is convenient for other applications. However, use is limited to representing absolute values in applications with peak detection, but even here it is still used more for number representation than for calculation. If only number representation is considered, the sign-magnitude notation shows less activity when the signal varies around zero. Note that, with compression techniques such as MPEG, a lot of zeros (000..00) are only represented by one bit. The use of compression techniques automatically reduces the power consumption.

- Optimum code.
  Even the code in which an operation is expressed can influence the power consumption. An example is shown in table 8.2

Table 8.2: Comparison of switching activity in a BCD counter and a Gray code counter

Standard binary code (BCD)			Gray code
	number of changing bits		
0 0 0	3	1	0 0 0
0 0 1	1	1	0 0 1
0 1 0	2	1	0 1 1
0 1 1	1	1	0 1 0
1 0 0	3	1	1 1 0
1 0 1	1	1	1 1 1
1 1 0	2	1	1 0 1
1 1 1	1	1	1 0 0
	2	1	



average/clock

Table 8.2 shows the switching activity of two 3-bit counters: a BCD counter and a Gray code counter. The table also shows that the BCD counter exhibits an average of twice the switching activity of the Gray code counter.

- Alternative implementations for arithmetic multiplier and adder circuits.
  Besides the previously-discussed options (bit serial versus bit parallel and number representations), there are many other alternatives that can influence the power consumption of arithmetic logic. Alternatives for multiplier implementation include: Booth multiplier, array multiplier and Wallace tree multiplier, etc. Alternatives for the addition process are carry select, carry ripple, carry save and carry look ahead techniques. With respect to power consumption, a general rule of thumb is: "every implementation that speeds up an arithmetic process will require additional power." The choice of an arithmetic implementation depends on the priorities in an application with respect to speed, area and power consumption. Therefore, no fixed prescribed choice can be given for low power here.
- Microprocessor and microcontroller architecture.
  Many products with low-power consumption use microprocessor cores: portable telecommunication, medical electronics, watches and games. Maintaining or improving the performance while reducing the power consumption is a continuous challenge for the designers of new products in these fields. Generally, an instruction in a RISC architecture needs more execution cycles than in a CISC architecture. Pipelined RISC microprocessors use one or two cycles per instruction, while the CISC microprocessor often uses 10-20 cycles. However, complex algorithms mapped on a RISC machine generally require more instructions than a CISC machine. The CISC architecture may have too much hardware for only simple algorithms, which leads to a kind of overkill.

  In these cases, CISC power consumption may be more. From literature, it appears that each architecture (whether RISC or CISC) can in itself be optimised for low power. No real winner can be distinguished here because both architectures have many parameters to be adjusted for optimum low power.

- Limited I/O communication.
  In many applications, many I/O pins are used for communication between processor and memory and/or A/D or D/A converters. To reduce activity, these blocks have to be integrated on one single die. This may increase the chip power, but it certainly reduces the system power.
- Synchronous versus asynchronous.
  In synchronous circuits, the data transfer to, on and from the chip is usually controlled by a global clock signal. However, this clock signal does not contain any information. In contrast, asynchronous circuits proceed at their own speed. Here, the output of one circuit is immediately used as an input to the next. The relatively large difference in delay paths may lead to random operation and requires a special design style and test strategy. Actually, there are two kinds of asynchronous circuits: asynchronous subfunction(s) of synchronous designs and purely asynchronous designs (self-timed circuits).
- Asynchronous subfunction (of synchronous design).
  A synchronous chip is nothing more than a collection of asynchronous circuits which are separated by flip-flops (registers). Thus, asynchronous blocks are embedded between registers. A 4-bit counter may serve as an example.

  Figure 8.20 shows an asynchronous implementation and two synchronous alternatives of this counter. In the synchronous versions, each flip-flop is clocked at the highest frequency, which consumes a lot of power. The synchronous counter with parallel carry consumes the most power because it has more hardware than the ripple carry counter. In the asynchronous counter version, only the first flip-flop (LSB) runs at the highest frequency, whereas the others act as frequency dividers (divide by two). This version therefore requires much less power (about 1/3) than the best of the synchronous versions.

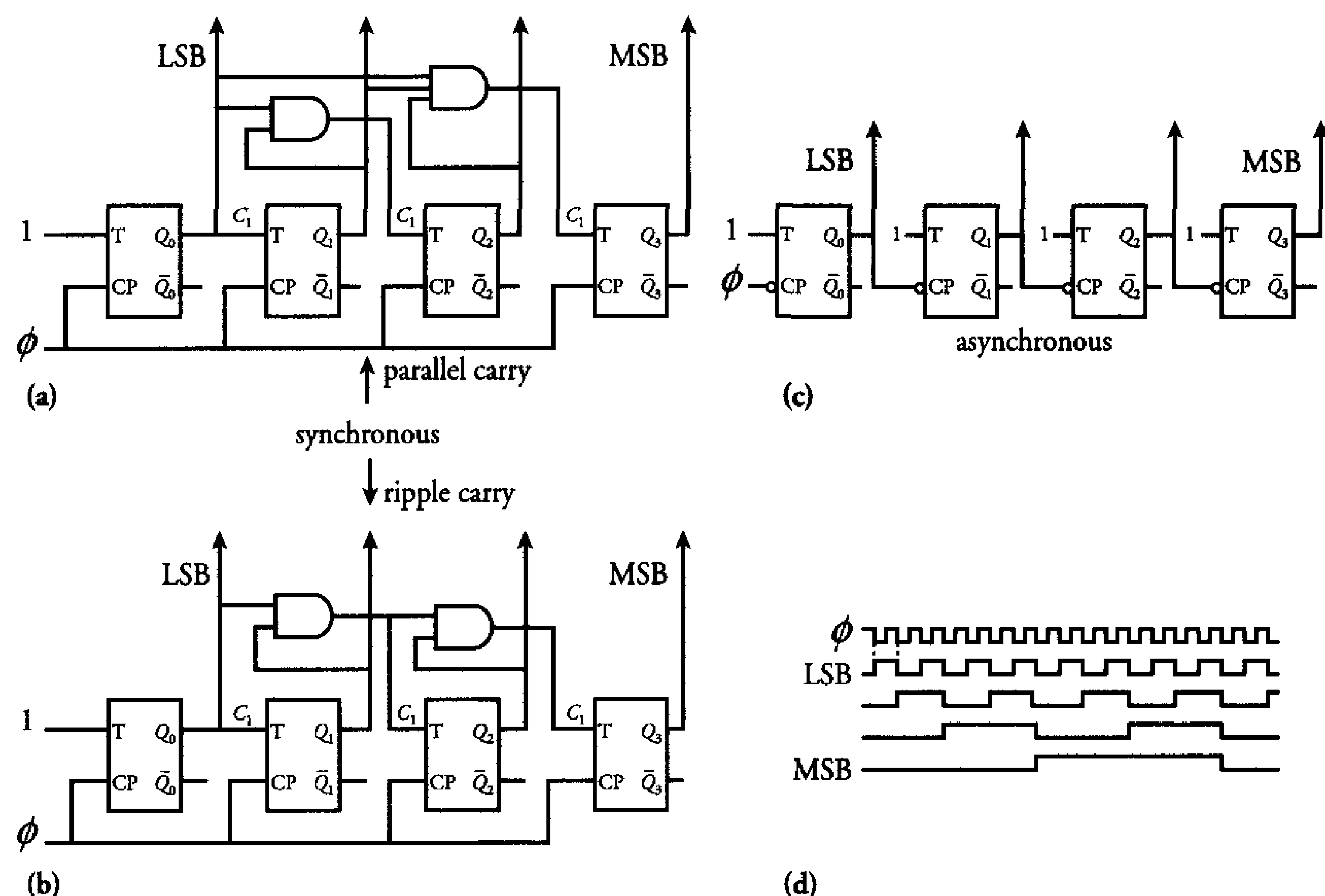


Figure 8.20: *Different versions of a 4-bit counter with timing diagram. a) synchronous with parallel carry b) synchronous with ripple carry c) synchronous and d) timing diagram*

- Pure asynchronous designs (self-timed circuits).
  A basic asynchronous design requires additional hardware to perform the necessary request (GO) and acknowledge (DONE) signals. Figure 8.21 shows a full-adder cell implemented as an asynchronous logic cell.

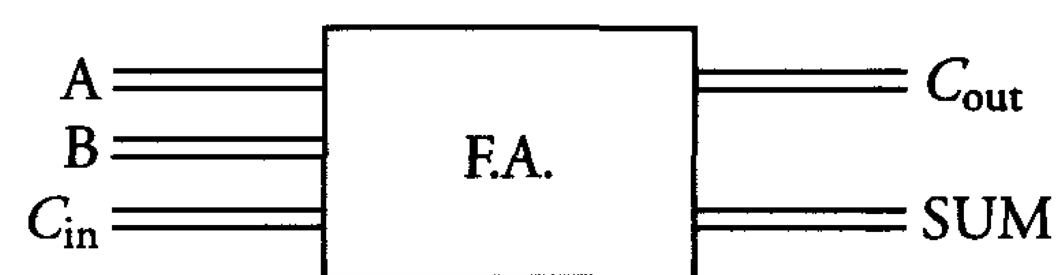


Figure 8.21: *Self-timed logic cell*

In this technique, an enormous area overhead must be spent to implement and route the additional logic that is associated with each request and acknowledge signal. This overhead is at least a factor two. An advantage is that no glitches can occur (see next

subsection B). Another way of implementing self-timed circuits is to generate the request and acknowledge signals at a higher level of circuit hierarchy, see figure 8.22.

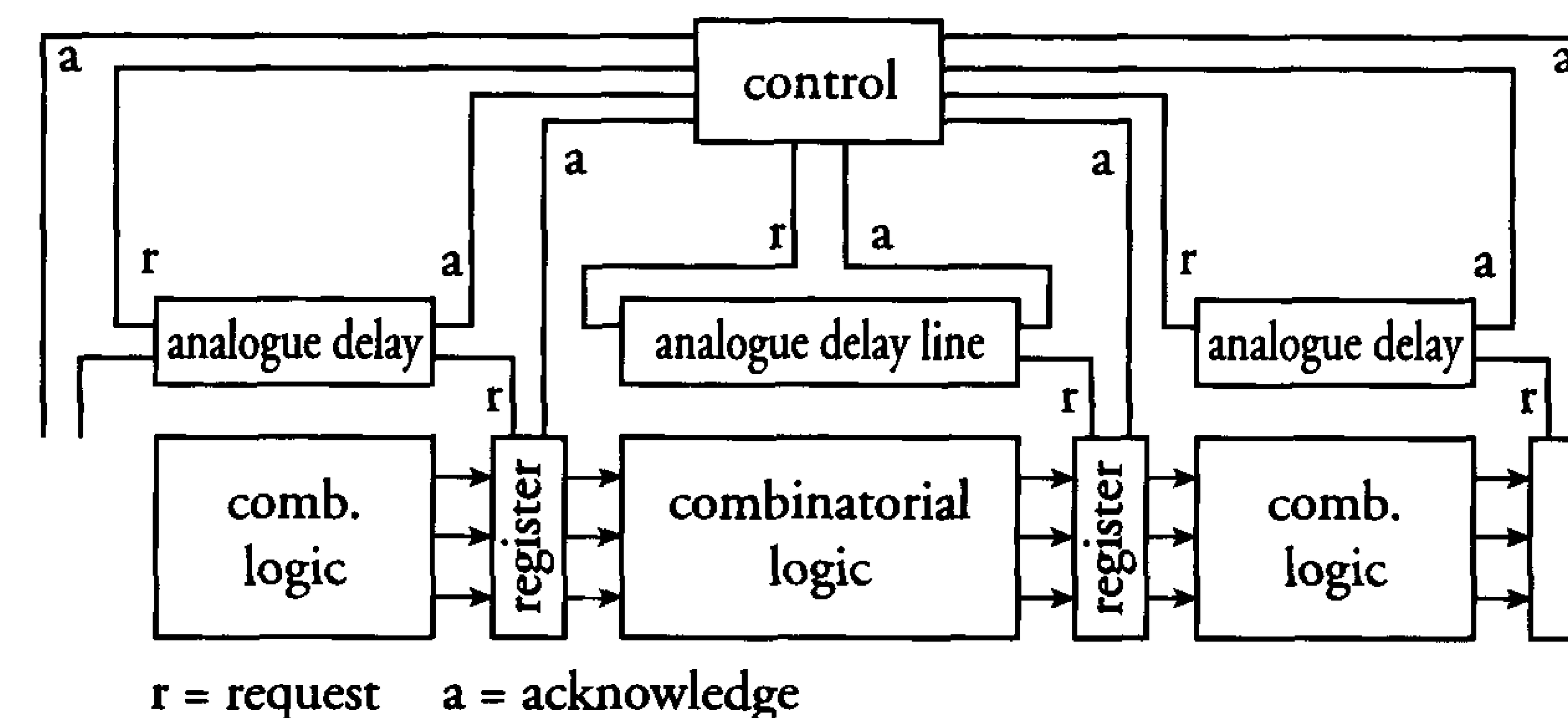


Figure 8.22: *Self-timed circuit by using analogue delay that mimics combinatorial logic delay*

In this way, a higher component efficiency is achieved. A major disadvantage is that the dummy delay lines must be designed to be marginally slower than the slowest path through the combinatorial logic. This combinatorial logic also shows glitches as in synchronous logic, see next subsection B.

The key to this form of self-timed logic is the ability to model the combinatorial logic delay with a very small circuit (inverter chain). Self-timed techniques are also used in synchronous systems, for instance, to generate the clocks needed in smaller parts of the chip. In dynamic RAMs, dozens of self-timed clocks are generated on chip. A final discussion on power consumption of synchronous and asynchronous circuits leads to the following statement:

'Although asynchronous circuits are only active when necessary and thus operate at reduced power, these need not be *the* implementation for low-power circuits.'

Synchronous logic, optimised for low power, can achieve a power level that approaches that of asynchronous circuits. Up to now, however, synchronous logic has only been optimised for high speed (and, in some cases, for small area). Certain circuits are

particularly suited for asynchronous implementation. However, for those that are not, the power consumed by the control circuit and the large test circuit can be greater than the advantage gained by having no clocks.

Currently, several microprocessor design houses are quietly replacing relatively small portions of their systems with asynchronous units. Hewlett-Packard has added an asynchronous floating-point multiplier to its 100 MHz RISC processor. These approaches are probably the wave of the future: asynchronous sub-units residing in a synchronous framework [15].

- Optimised memory design.
  The previously-discussed comparison can also be used in the realisation of memories. To reduce internal memory activities, self-timed techniques are used to generate a lot of different clocks or acknowledge signals which should be active according to some sequence. The alternative to performing one single operation (such as activate precharge, deactivate precharge, select word line, activate sense amplifier and select column, etc.) in one clock period means that a lot of clock periods are needed for only one read or write operation. This would be at the cost of increased power consumption.

B) *Implementation level.*

- Reduce glitching.
  Static CMOS circuits can exhibit glitches (also called dynamic hazards, critical races or spurious transitions) as a result of different propagation delays from one logic gate to the next. Consequently, a node can have multiple unnecessary transitions in a single clock cycle before it reaches its final state. Figure 8.23 gives an example.

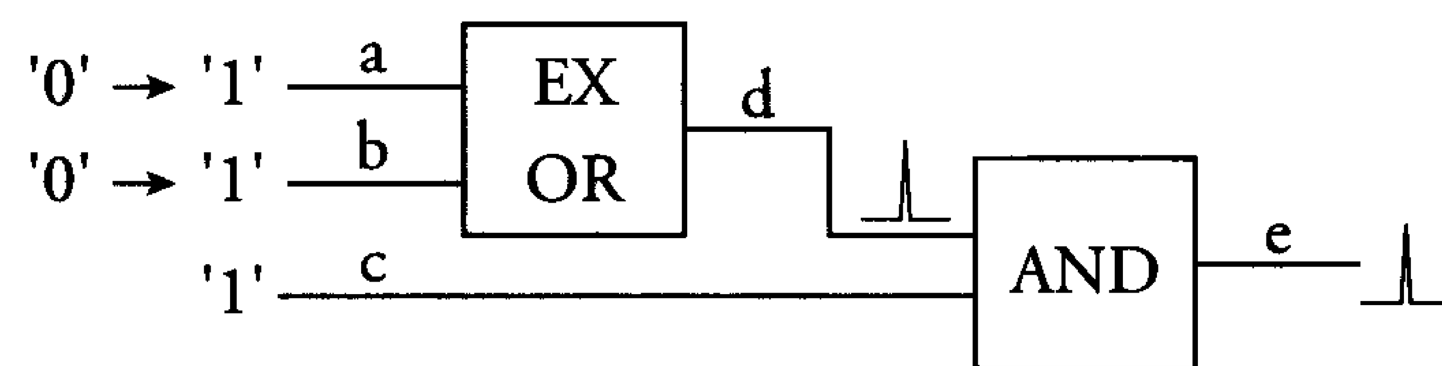


Figure 8.23: *Unnecessary transitions in a simple logic circuit*

Suppose the (a,b) inputs of an EXOR gate switch from (0,0) to (1,1). In both situations, output d of the EXOR should remain low ('0'). However, because of a different delay in the switching of the input signals, the state diagram of the (a,b) inputs might follow the following sequence (0,0)→(0,1)→(1,1). Therefore, the (a,b) inputs are (0,1) for a very short period of time, resulting in a temporary '1' at output d. This glitch also propagates through the next AND gate.

Such unnecessary transitions dissipate extra power. The magnitude of this problem is related to the kind of circuit to be realised. As a result of the occurrence of glitches, an 8-bit ripple carry adder with random input patterns consumes about 30% more power. For an 8*8-bit array multiplier, this number is even 150%, for a 16*16-bit array multiplier 326% and for standard cell implementation of a progressive scan conversion circuit, it is even 379%! Therefore, a large power saving could be achieved in such circuits if all delay paths were balanced.

Different architectures can lead to a different percentage of unnecessary transients. A 16*16 bit Wallace tree multiplier has only 16% glitches, compared to the above 326% for a 16*16-bit array multiplier. The Wallace tree multiplier has far more balanced delay paths.

Finally, another way of reducing the number of glitches is to use retiming/pipelining to balance the delay paths.

- Optimise clock activity.
  There are two reasons why clock signals are very important with respect to power dissipation. The first is that clock signals run all over the chip to control the complete data flow on the chip in a synchronised way. This means that clock capacitance caused by both very long tracks and a large number of flip-flops can be very large. In complex VLSI chips, the clock load can be as high as 500 pF. The second reason is that the clock signal has the highest frequency (the maximum switching frequency of data signals is only half the clock frequency). The total power consumed by the clock network depends heavily on the number of connected flip-flops and latches.

  Figure 8.24 shows the relative clock power consumption as a function of the average activity on a chip. This is expressed as a fraction of the total power consumption.

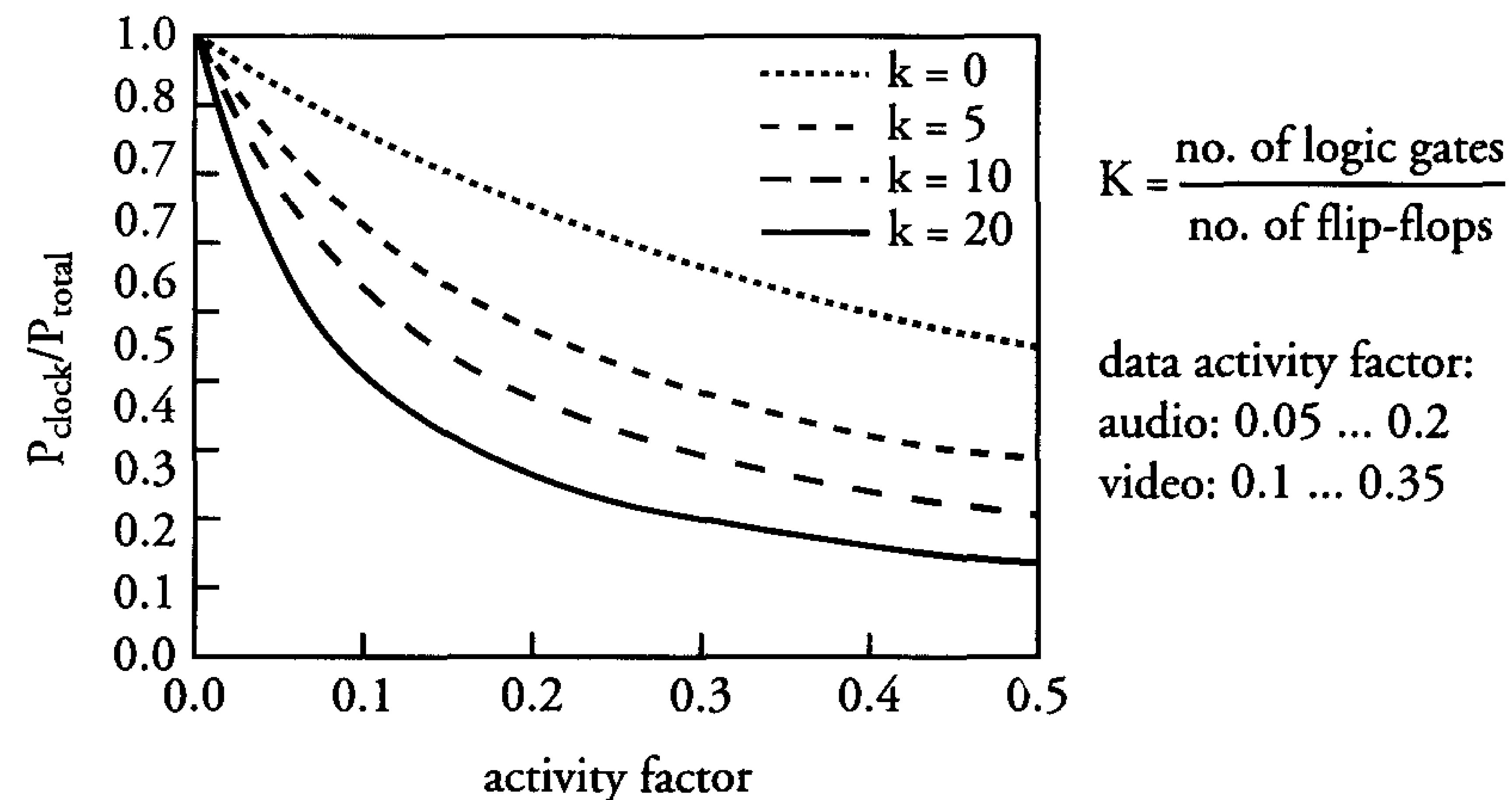


Figure 8.24: *Relative clock power consumption as a function of the activity*

Because the clock dissipation can be as high as 20-50% of the total chip dissipation, its activity should be reduced. This can be done because clock transitions carry no information. There are several ways to reduce clock activity. Including the use of *dual-edge triggered flip-flops.* If a flip-flop could be triggered on both edges of the clock pulses instead of on only one edge, it would be possible to use a clock at half frequency for the same data rate, thereby reducing the power dissipation of the total clock network.

A flip-flop that acts at both edges of the clock pulse is called a dual-edge triggered (DET) flip-flop, whilst the conventional positive and negative-edge triggered flip-flops belong to the category of *Single-Edge Triggered* (SET) flip-flops. However, the use of DET flip-flops has been limited up to now by the high overhead in complexity that these flip-flops require. Both the SET and DET flip-flops have two latches. Basically, in a DET flip-flop (see figure 8.25) the two latches are arranged in parallel, while in a SET flip-flop, see figure 8.25(a), they are placed serially [14]. DET and SET flip-flops show comparable maximum data rates, however, DET flip-flops either require additional silicon area, or they are more difficult in use with respect to timing aspects [19,20].

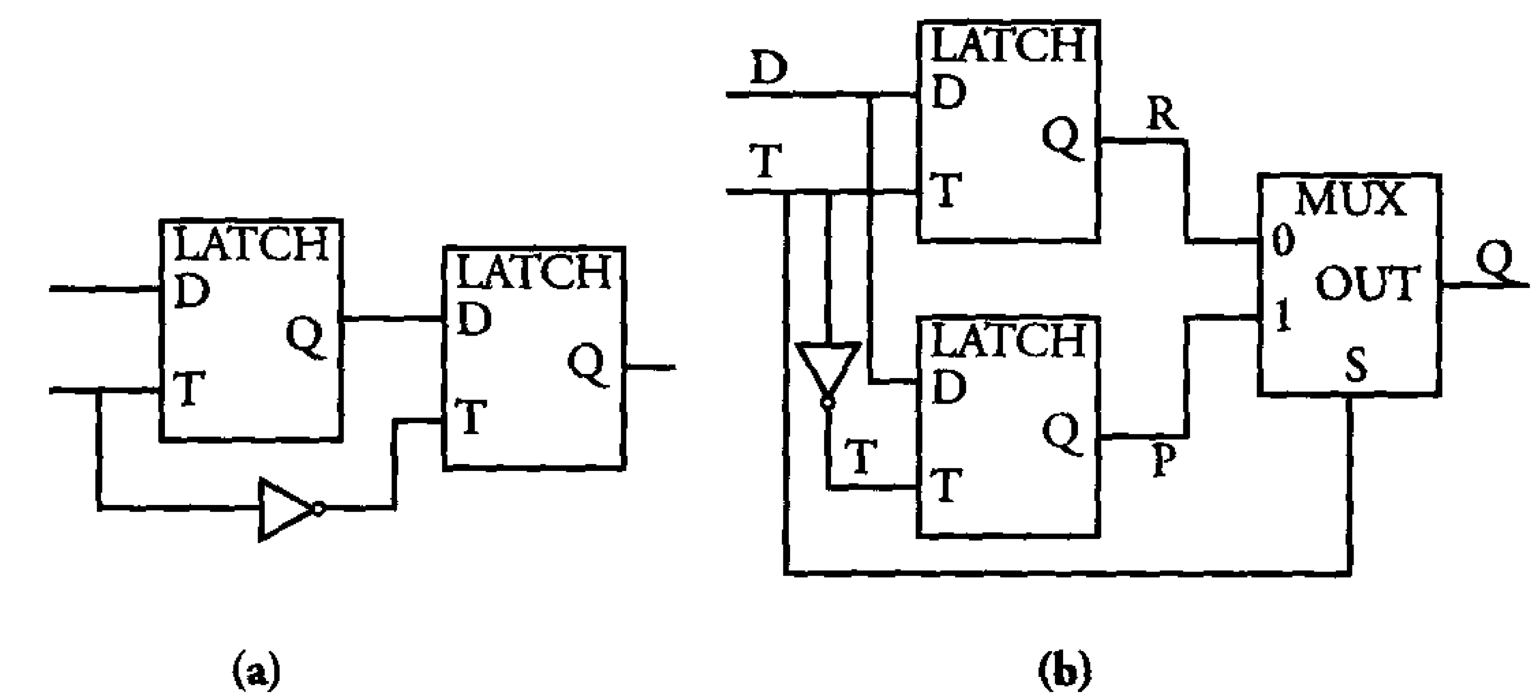


Figure 8.25: *Schematic diagram showing a SET and a DET flip-flop*

Since the clock contribution to the overall chip power consumption depends on the data activity, power savings of 15 to 45 percent are possible, at the cost of some additional flip-flop area (10 to 20%), when using DET flip-flops.

In conventional synchronous designs, the complete system is clocked at the highest frequency, even when some portions of the chip could operate on fractions of this frequency. In some cases, clock dividers are used to generate these lower frequencies. From a low-power point of view, we should start from the opposite direction.

This means that we supply the chip with the lowest required frequency and generate higher clock rates locally, if needed. This can be achieved by PLL-like circuits. In this way, the globally distributed clock would run at the minimum clock frequency and the higher clock frequencies would only be used where they are really needed. This might reduce the global clock activity drastically and also reduce the functional activity.

Another approach to reduce the total chip activity is to switch the clock off temporarily for certain functional blocks, or even for the complete chip during moments that no useful operations are executed. In this respect, different names are used for the same issue: gated clocks, stop-the-clock, sleep mode and power-down mode, etc.

A representative example is a coefficient ROM, whose power consumption can be relatively large. In many cases, such a ROM

is often used for less than 1% of the time. Forcing this block to the power-down mode, for instance, by switching off its clock saves 99% of its total power consumption.

When a signal processor enters the power-down mode, all its internal memory and register contents must be maintained to allow the operation to be continued unaltered when the power-down mode is terminated. Depending on the state of some control register(s), external devices can cause a wake-up of the DSP, e.g. when terminating an input operation. The processor enters the operating state again by reactivating the internal clock. The program or interrupted routine execution then continues.

A disadvantage of gated clocks (sleep modes, etc.) is that some logic operation has to be performed on the clock signal. This causes an additional delay for the internal gated clock, which may result in timing problems during data transfer between blocks that run at the main clock and those that run at a gated clock. Therefore, compensated delays must be used in those blocks that do not use a gated clock. Generally, gated clocks decrease the design robustness with respect to timing.

- Dynamic versus static CMOS.
  The decision to implement a circuit in dynamic or static CMOS logic not only depends on power considerations. Aspects of testability, reliability, ease of design and design robustness are also very important here. In the comparison of dynamic and static CMOS realisations, several differences show up with respect to power. As precharge and sample periods in dynamic CMOS circuits are separated in time, no short-circuit dissipation will occur. Also, the absence of spurious transitions (hazards) reduces the activity of dynamic CMOS. However, precharging each node every clock cycle leads to an increase of activities.
  EXAMPLE: (assume all input combinations are uniformly distributed), consider table 8.3:

Table 8.3: Function table of a 2-input NOR and an EXOR gate

2-input NOR		EXOR	
a b	z	a b	z
0 0	1	0 0	0
0 1	0	0 1	1
1 0	0	1 0	1
1 1	0	1 1	0

Because each output in a dynamic CMOS chip is high during precharge, the output will be discharged in 75% of the input combinations of a 2-input NOR $\Rightarrow$ activity factor 0.75. For the EXOR: activity factor 0.5. In static CMOS, power is only dissipated when the output goes high:

$$\begin{aligned} \text{NOR}: \quad & P_{0\to1} = P(0) \cdot P(1) = 3/4 \cdot 1/4 = 3/16 \\ \text{EXOR}: \quad & P_{0\to1} = P(0) \cdot P(1) = 1/2 \cdot 1/2 = 1/4 \end{aligned}$$

Usually, the logic function in dynamic CMOS is realised with an nMOS pull-down network, while a pMOS transistor is used for precharge. This leads to small input capacitances, which makes dynamic logic attractive for high-speed applications. Besides the higher activity factor, the additional clock loads to control the precharge transistors also leads to much higher dissipation. The use of dynamic logic is not as straightforward and common as static logic. In terms of design robustness and ease of design, static CMOS is favourable as well. Finally, when power reduction techniques (such as power-down modes, in which the clock is stopped) are being implemented, dynamic CMOS is not applicable because of its charge leakage. Generally, it can be stated that dynamic logic is not a real candidate for low-power (low-voltage) realisation.

- Connect high-activity input signals close to the output of a logic gate.

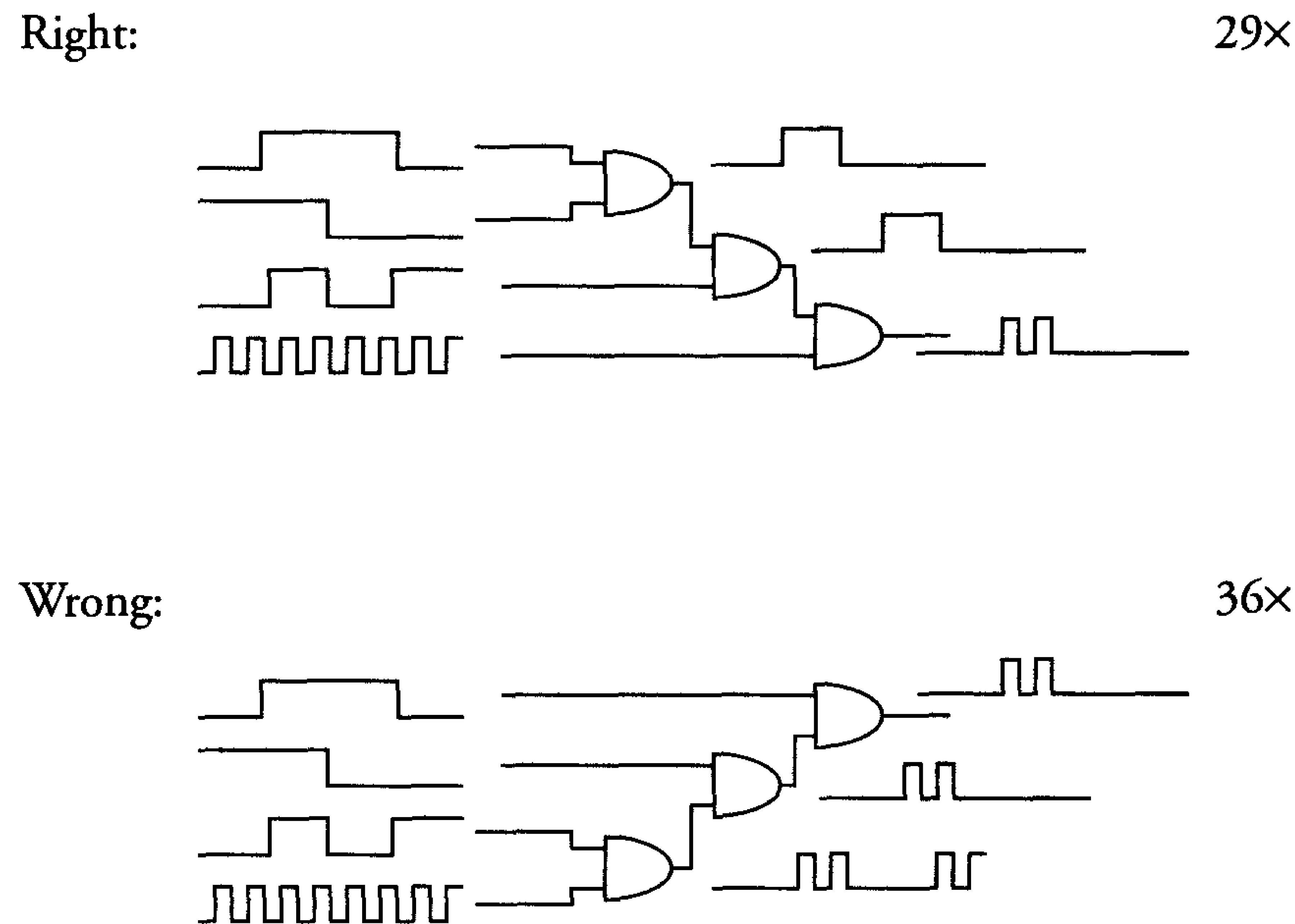


Figure 8.26: *Reduction of total activity by ordering signals*

Figure 8.26 shows that connecting signals with high activity close to the output of the propagation chain will reduce the total power consumption of that chain.

- Exploit the characteristics of library cells.
  Here again, when there are signals showing high activity, it is obvious that these will cause less power dissipation when they are connected to the low-capacitance inputs of logic gates. Figure 8.27 shows an example.

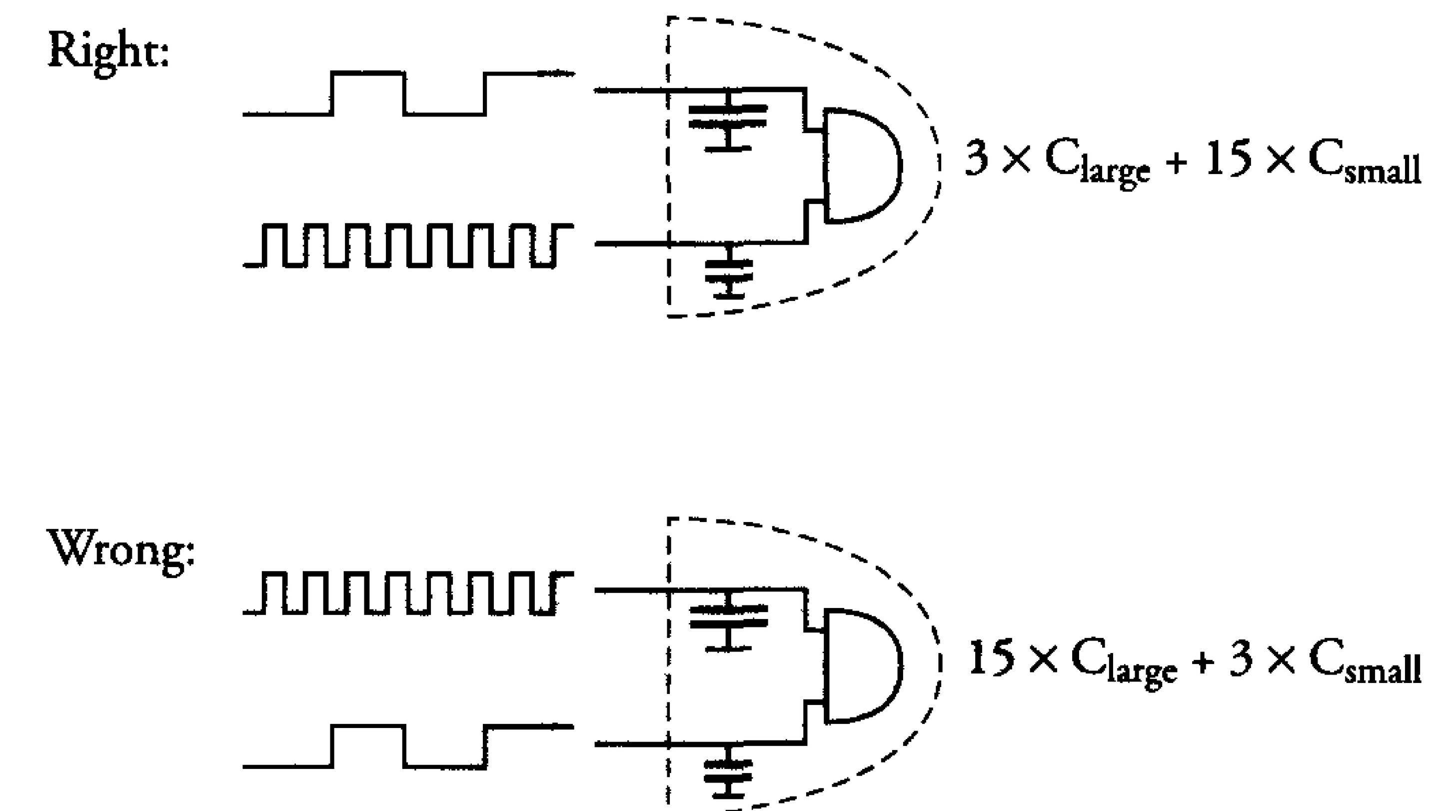


Figure 8.27: *Reduction of power dissipation by matching high-activity signals with low-capacitance inputs*

It should be clear that the power savings of these last two items can only be achieved by dedicated software programs, which perform some statistics on the signal activity inside a logic block.

8.5 Computing power versus chip power, a scaling perspective

The scaling process and its impact on the performance, reliability and signal integrity of MOS ICs is extensively discussed in chapter 11. However, the scaling process with respect to the system performance of digital signal processors (DSPs) requires a different approach.

An important parameter reflecting this system performance is the computing power of a DSP. Generally, this computing power (U) is equal to:

$$U = n \cdot f$$

where n represents the number of transistors and f is the frequency.

The power dissipation of the DSP is equal to:

$$P = n \cdot f \cdot C \cdot V^2$$

From these two equations, it can be derived that the computing power per Watt dissipation is equal to:

$$U/[\mathrm{W}] = \frac{1}{C \cdot V^2}$$

In the following discussion on scaling, V_T effects and velocity saturation are neglected. If the scaling factor between two successive process generations is s (usually $s \approx 0.7$), then the number of transistors will increase to:

$$n_\mathrm{s} = n/s^2$$

and the frequency to:

$$f_\mathrm{s} = p/s^2 \cdot f$$

Where p equals the voltage scaling factor ($V_\mathrm{p} = p \cdot V$), as this factor may differ from s.
The capacitance C scales to:

$$C_\mathrm{s} = s \cdot C$$

Combining the previous equations results in the following expressions concerning computing power and chip power impact: the computing power scales to:

$$U_\mathrm{s} = n_\mathrm{s} \cdot f_\mathrm{s} = p/s^4 \cdot U$$

and the power dissipation per unit area with:

$$P_\mathrm{s} = n_\mathrm{s} \cdot f_\mathrm{s} \cdot C_\mathrm{s} \cdot {V_\mathrm{p}}^2 = p^3/s^3 \cdot P$$

Therefore, the computing power per Watt after scaling increases to:

$$U_\mathrm{s}/[\mathrm{W}] = \frac{1}{s \cdot p^2} \cdot U/[\mathrm{W}]$$

Remarkably, voltage scaling has more impact on the computing power per Watt than the process scaling.

From the 0.35 μm CMOS process generation onwards into the deep-submicron processes, the voltage scale factor p is about equal to the

process scale factor. This means that, neglecting the second-order effects, the computing power per Watt for future generations of DSPs will increase to:

$$U_\mathrm{s}/[\mathrm{W}] = \frac{1}{s^3} \cdot U/[\mathrm{W}]$$

Each DSP generation will therefore be much more power efficient. Second-order effects have a more negative impact on the transistor performance and thus on the DSP efficiency. However, even after such a reduction in efficiency improvement, a lot of new DSPs are still expected to enter the market with much more power efficiency.

8.6 Conclusions

With respect to conventional CMOS processes and design styles, large power savings could be achieved because they were optimised for speed and area. Power can be reduced in different ways. However, the largest power savings can be achieved by reducing the supply voltage. Fortunately, from 0.35 μm CMOS technology onwards into the deep-submicron processes, supply voltage scaling is a must.

In technology development, a few measures can be taken to reduce power: limit the leakage currents and limit the parasitic capacitances.

In the design, however, there are many options for reducing the total capacitance and activity on a chip. However, a complete and clear set of design rules cannot be given, because the use of many of these options depend on the application. This chapter is meant to present most of these options and to provide the designer with a low-power attitude.

Finally, although several alternative low-power CMOS design styles have been presented at conferences and magazines during the last decade, static CMOS logic is still favourable in many ways. It is very robust with respect to transistor scaling and supply voltage reduction. Besides this, design integrity is becoming a key issue in deep-submicron VLSI design, which also makes static (complementary) CMOS the best candidate for many process generations to come.

8.7 References

[0] Kaushik Roy, Sharat C. Prasad.,
'Low-power CMOS VLSI Circuit Design'
John Wiley & Sons INC, 2000

[1] K. Seta, et al., '50% Active-Power saving without speed degradation using standby power reduction (SPR) Circuit',
IEEE Digest of Technical papers, pp 318,319, Feb. 1995

[2] T. Kuroda, et al.,
'A 0.9 V, 150 MHz, 10 mW, 4 mm^2, 2D Discrete Cosine Transform Core Processor with variable Threshold Voltage (V_T) Scheme',
IEEE Journal of Solid-State Circuits, pp 1770-1779, Nov. 1996

[3] M. Izumikawa , et al.,
'A 0.25 μm CMOS 0.9 V, 100 MHz, DSP Core',
IEEE Journal of Solid-State Circuits, pp 52-61, Jan. 1997

[4] H.J.M. Veendrick, 'Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits',
IEEE Journal of Solid State Circuits, Vol. SC-19,
No. 4, August 1984, pp 468-473

[5] Von Kaenel et al.,
'A Voltage Reduction Technique for Battery-Operated Systems',
IEEE JSSC. Vol. 25, Oct. 1990, pp 1136-1140

[6] K. Yano et al., 'A 3.8 ns CMOS 16x16-b multiplier using complementary pass-transistor logic',
IEEE JSSC, Vol. 25, April 1990, pp 388-393

[7] M. Suzuki et al.,
'A 1.5 ns 32b CMOS ALU in double pass-transistor logic',
Digest ISSCC, Feb 1993, pp 90-91

[8] A. Parameswar et al.,
'A swing restored pass-transistor logic-based multiply and accumulate circuit for multimedia applications',
IEEE JSSC, Vol. 31, June 1996, pp 805-809

[9] K. Jano et al., 'Top-down pass-transistor logic design',
IEEE JSSC, Vol. 31, June 1996, pp 792-803

[10] J.H. Pasternak and C. Salama, 'Differential pass-transistor logic', IEEE circuits & Devices, July 1993, pp 23-28

[11] M. Song et al., 'Design methodology for high speed and low power digital circuits with energy economized pass-transistor logic (EEPL)', Proc. 22nd ESSCIRC Digest, 1996, pp 120-123

[12] W.H. Paik et al., 'Push-pull pass-transistor logic family for low-voltage and low-power', proc. 22nd ESSCIRC Digest, 1996, pp 116-119

[13] R. Zimmermann and W. Fichtner, 'Low-Power Logic Styles: CMOS Versus Pass-Transistor Logic', IEEE JSSC, Vol. 32, July 1997, pp 1079-1090

[14] R.Hossain, et al., 'Low Power Design Using Double Edge Triggered Flip-Flops', IEEE Trans. on VLSI, Vol.2, No.2, June 1994

[15] C. Maxfield, 'To be or not to be asynchronous that is the question', EDN, December 7, 1995, pp 157-173

[16] D. Berndt, 'Maintenance-Free Batteries, A handbook of battery technology', Research Studies Press, LTD, 1997

[17] A. Bellaouar and M.I. Elmasry, 'Low-Power Digital VLSI Design, Circuits and Systems', Kluwer Academic Publishers, 1995

[18] A. Chandrakasan, et al., 'Low-power digital CMOS design', IEEE Journal Solid State Circuits, April 1992, pp 473-484

[19] Jerry Yuang, et al., 'New Single-Clock CMOS Latches and Flipflops with Improved Speed and Power Savings', IEEE Journal Solid State Circuits, January 1997, pp 62-69

[20] A.G.M. Strollo, et al., 'Low power double edge-triggered flip-flop using one latch', Electronics Letters, Vol. 35, 4 February 1999, pp 187-188

8.8 Exercises

1. Why must every designer always have a low-power attitude?
2. Which of the different power contributions is the larger and why?
3. How could the sub-threshold leakage power dissipation be reduced?
4. In optimizing a complete library for low power, which of the library cells would you focus most of your attention to?
5. What is the greatest advantage of constant-field scaling with respect to power dissipation?
6. What would be the difference in activity factor between a static and dynamic CMOS realisation of the next boolean function: $z = \overline{abc}$
7. Repeat exercise 6 for $z = \overline{a + b + c}$

Chapter 9

Circuit reliability and signal integrity in deep-submicron designs

9.1 Introduction

With shrinking feature sizes and increased chip sizes, the average delay of a logic gate is now dominated by the interconnection (metal wires) rather than by the transistor itself. Most of the potential electrical problems, such as cross-talk, critical timing, substrate bounce and clock skew, etc. are related to the signal propagation and/or high (peak) currents through these metal wires.

Currently, high-performance complex VLSI chips may contain millions of transistors that realise complete (sub)systems on one single die. For the design of these ICs, a lot of different tools are used, see chapter 7. The sequence in which these tools are used, from the upper hierarchy levels down to the layout level, is called the "design flow".

At the moment, IC design flows have been automated so much that "first time right silicon" is considered as natural. However, keeping control over all the tools used in the design flow (the high-level description language, the synthesis tools and the verification tools, to name a few) requires the complete attention of the designers. Thus, even when designers are familiar with the physical aspects of complex ICs, the potential electrical problems do not get the attention that they require, particularly in deep-submicron technologies.

First silicon (especially of high-performance ICs) therefore shows first-time-right functionality but often at lower or higher supply voltages and/or at lower frequencies than required. Actually, at a time where designers are drifting away from the physical transistor level into abstract high-hierarchy levels of design, exactly the opposite would be required to get current and future VLSI chips operating electrically correctly. Many ICs are therefore no longer "correct by design" but are "designed by corrections".

Therefore, this chapter focuses on such reliability aspects as latch-up, electrostatic discharge (ESD) and electromigration. Hot carrier effects, which are also important with respect to reliability, are treated in chapter 2. Besides this discussion on reliability, maintaining signal integrity is also an increasingly important requirement for proper circuit operation of deep-submicron ICs.

This chapter is intended to present the basics of the electrical consequences of the scaling process into the deep-submicron era. Knowledge about these consequences and taking the right measures during the design improves its robustness and increases the chance of true first-time-right products.

9.2 Design for reliability

9.2.1 Introduction

The reliability of packaged integrated devices generally deals with their long-term behaviour. Reliability tests include physical, mechanical and electrical stress tests to allow rapid evaluation of the IC's sensitivity to related effects, such as temperature changes, humidity, latch-up, ESD, electromigration and hot-carrier degradation. The first two effects in this list are related to packaged dies and discussed at the end of this chapter. The other four effects are related to the quality of the design and therefore every designer has to design for reliability.

9.2.2 Latch-up in CMOS circuits

The presence of nMOS and pMOS transistors in a CMOS process leads to parasitic thyristors, as shown in figure 9.1. These thyristors switch on when a sufficiently positive voltage is present in the substrate (A) or when a voltage that is sufficiently negative with respect to V_{dd} is present in the well (B). Some nodes in a circuit will then assume a fixed logic

level. This undesirable effect is called latch-up and leads to incorrect circuit behaviour.

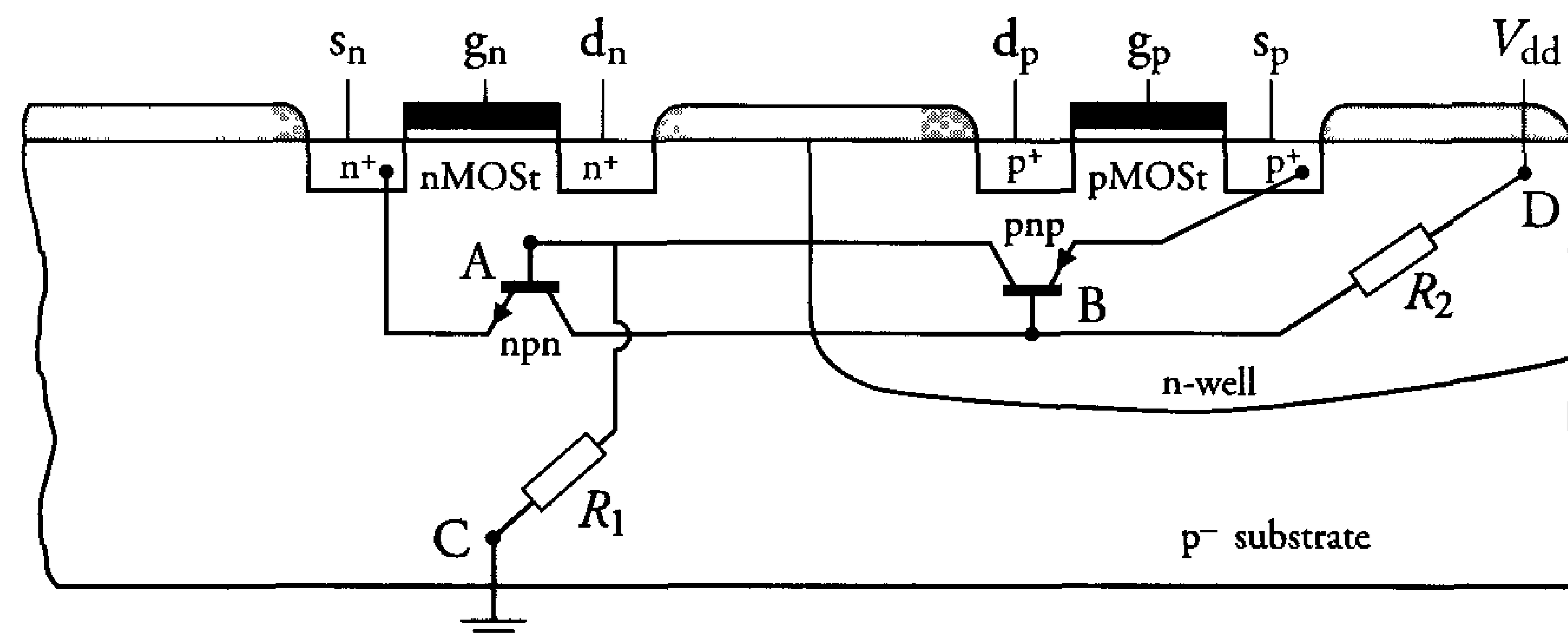


Figure 9.1: *Parasitic thyristors in CMOS*

Latch-up does not occur in CMOS processes of the Silicon-On-Insulator (SOI) or SIMOX types, as discussed in chapter 3. For the remaining CMOS processes, the problem is fundamentally the same. The latch-up phenomenon will now be explained with the aid of the basic cross-section of an n-well CMOS configuration, shown in figure 9.1. The parasitic pnp and npn transistors that comprise the thyristor are schematically presented in this figure. Resistor R_1 represents the series resistance of the p^- substrate between the substrate connection C and a random location A in the substrate. Resistor R_2 represents the series resistance between an n-well connection D to the V_{dd} and a random location B in the well. The equivalent circuit for the parasitic thyristor is shown in figure 9.2.

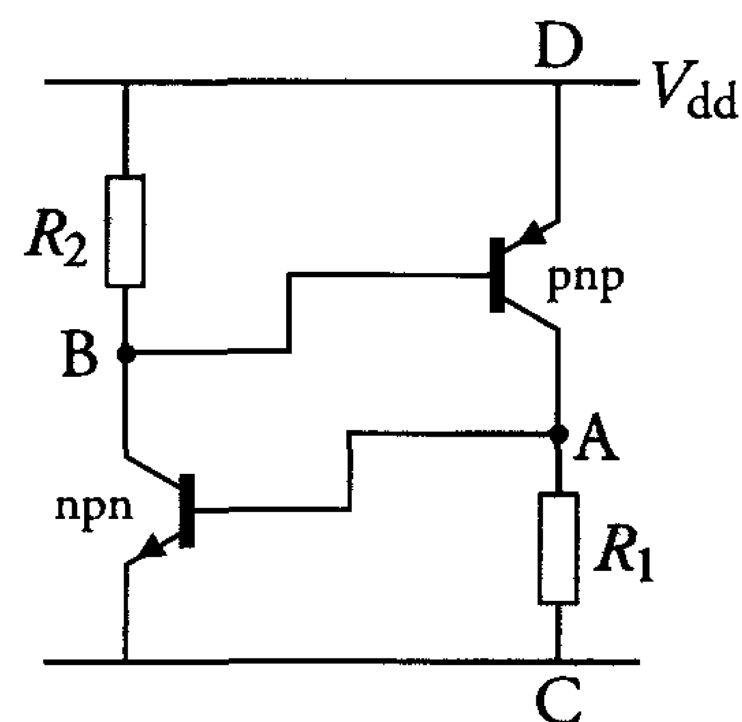


Figure 9.2: *Equivalent circuit diagram of the CMOS parasitic thyristor*

Source S_n of the nMOS transistor of figure 9.2 is assumed to be connected to V_{ss} and source S_p of the pMOS transistor is assumed to be connected to V_{dd}. Substrate currents through R_1 could cause the potential at A to rise to a point where the npn transistor conducts. The potential at B then falls and the pnp transistor conducts. Consequently, the potential at point A rises further. If the resulting feedback gain is greater than 1, the thyristor behaves like a latch and maintains its state. The currents that then flow through it are ultimately only limited by the series resistances R_1 and R_2. These large currents continue to flow until the supply voltage V_{dd} is removed. Clearly, this latch-up effect has disastrous consequences for circuit operation and must be avoided.

Remedies to avoid latch-up

Latch-up can be avoided in CMOS circuits by applying the following technological and/or design remedies.

- Minimise the substrate and well resistances.
  The substrate and well resistances in figures 9.1 and 9.2 are R_1 and R_2, respectively. The potentials at A and B will differ little from nodes C and D, respectively, when these resistances are of minimal value. The parasitic thyristor is then unlikely to turn on. Resistances R_1 and R_2 can be technologically minimised by using high substrate and n-well dopes. This poor solution causes an increase in the parasitic capacitance values. The back-bias effect, or K-factor, also increases. A better alternative is to apply the following rules during the design phase:

  1. Use extra substrate contacts to V_{ss} to reduce R_1 and extra well contacts to V_{dd} to reduce R_2 (at least one every five transistors).
  2. Circuits that carry large currents (such as buffers and I/O circuits, etc.) are particularly likely to cause latch-up. The distance between the nMOS and pMOS transistors in these circuits should be relatively large. If possible, the nMOS transistors should be enclosed by a p^+ guard ring that is connected to V_{ss} while the pMOS transistors should be enclosed by an n^+ guard ring that is connected to V_{dd}.

- Apply a back-bias to the substrate.
  When the p^- substrate in figure 9.1 is connected to a negative volt-

age (e.g. -2.0 V) instead of to V_{SS}, the base voltage V_A of the npn transistor will be lowered. Therefore, this npn transistor cannot be turned on easily.

- Use a p^- epitaxial layer on a p^+ substrate, as shown in figure 9.3. This layer is only a little deeper than the n-well (several micrometers) and facilitates the use of a relatively high dope for the p^+ substrate. A large part of the pnp collector current will therefore flow through this substrate and only a small part will flow into the base of the npn transistor. The npn transistor is then largely excluded from the latch circuit and cannot be turned on easily.

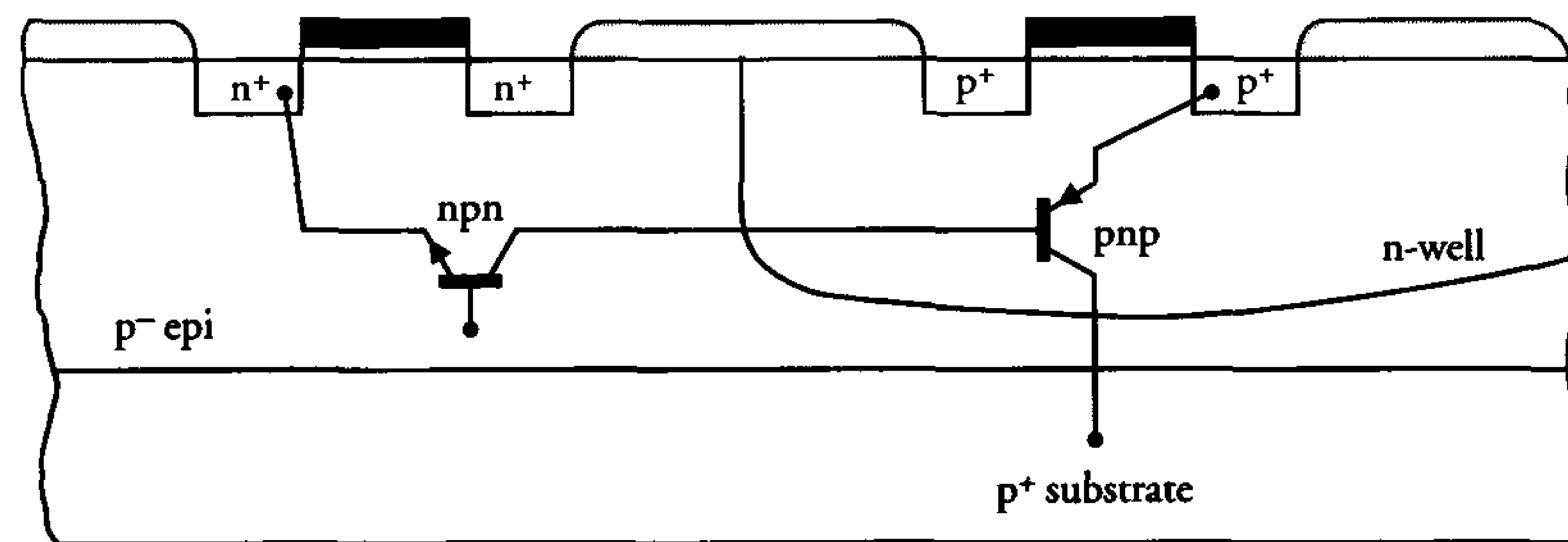


Figure 9.3: *n-well CMOS with thin p^- epitaxial layer on p^+ substrate*

- Use Silicon-On-Insulator (SOI) or SIMOX processes.
CMOS circuits that have to achieve a high standard of reliability (e.g. military and space applications) can be manufactured in SOI and SIMOX processes (chapter 3). The npn and pnp transistors in these processes are electrically isolated from each other. As early as 1982, Toshiba Corporation used a Silicon On Sapphire SOI process for the commercial manufacture of a 4kx1 static RAM.

The application of one or more of the above remedies has increased latch-up immunity to a very high level. In future technologies, the latch-up phenomenon is likely to disappear inside electronic circuits, as the supply voltage will be reduced every generation, see also chapter 11. However, at the chip I/Os, the requirements on latch-up remain high, as will be clear from the following remarks on testing.

Testing latch-up immunity and future trends

The highest chance of occurrence of latch-up is during testing. Standard testing requirements include immunity to 100 mA in epi-wafer ICs and 50 mA in homogeneous wafer ICs. This means that, with epi-wafer material, 100 mA can be supplied to the output of an output buffer even though no output transistor is conducting. This 100 mA now flows directly into the substrate, thereby raising the substrate voltage and possibly turning the thyristor on, see figure 9.1. In practice, tests are done with 150-200 mA at a maximum ambient temperature of 75 °C.

9.2.3 Electrostatic discharge (ESD) and its protection

Introduction

Integrated circuits are exposed to many possible sources of damage, both during and prior to their application. The principle cause of damage is electrostatic discharge (ESD). The duration of ESD is very short, normally lasting less than 100 ns, but it may result in very large power spikes. The high impedance of MOS input circuits makes them particularly vulnerable to physical damage when they are exposed to these spikes. This may result from operations such as handling unpackaged dies and bonding, etc. during an IC's production. It may also occur during testing, application or maintenance.

The damage caused by ESD is irreversible. The human body is one of the main sources responsible for ESD. A person walking on a carpet, for instance, and wearing shoes with highly insulating soles can build up a voltage in excess of 15000 V. The resulting charge can be transferred via ESD to electronic circuits. It is very important that precautions are taken to prevent ESD damage during IC production. In addition, protective measures must be included in an IC's design to ensure that it can withstand acceptably large ESD pulses. On-chip MOS protection circuits are used to increase the immunity of an IC to ESD pulses. These circuits are designed to provide input and output circuits with low impedance shunt paths, which prevent the occurrence of excessive voltages on the chip.

ESD test models and procedures

ESD sources are emulated in several different manners. Some manufacturers use the "zero-resistance method", in which a capacitor with a voltage of 300 V is connected directly to an IC's pin without a series resistor. The subsequent effects are measured. The "fingertip method", where the pin is approached with a charged finger, is also used. The method relies on ionisation of the air between the fingertip and the pin. ESD then takes place through an electric arc. A low voltage of a few hundred volts on the finger means that the distance between it and the pin has to be very small before a discharge occurs. A very high capacitance of a few hundred picofarads then exists between the finger and the pin. When the finger voltage is very high, however, the arc will occur at a larger distance and a low capacitance between the finger and the pin. The "human body method" is a popular alternative to the above ESD tests. It usually uses the "human body model" shown in figure 9.4, together with a DUT (device under test). This model is specified in the internationally accepted MIL-STD-883C standard.

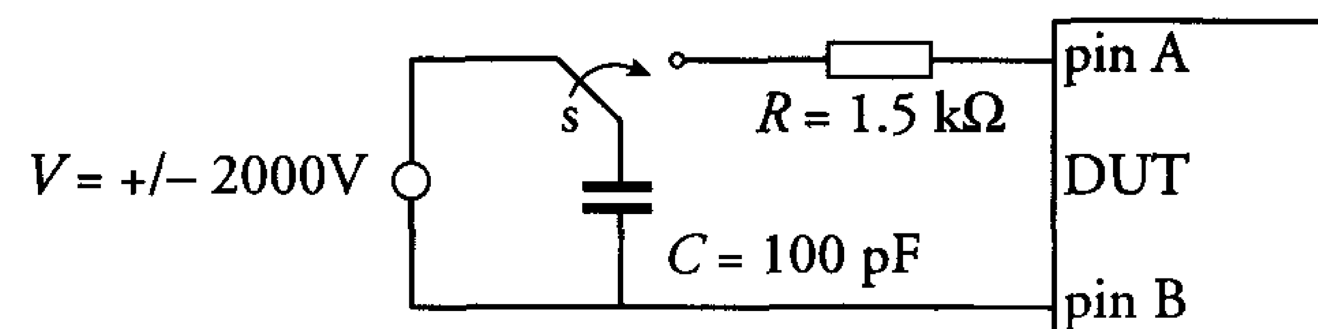


Figure 9.4: *An ESD test based on the human body model*

The test is normally done on an ESD tester. This human body model has not changed much over the last decade. Pins A and B of the DUT are connected to the positive and negative terminals, respectively, of the human body model. These pins may be input, output, ground or supply pins. All of the other DUT pins are unconnected. Capacitor C is charged when the switch is in the left position. When the switch is placed in the right position, C discharges via resistor R through pins A and B. Generally, three positive and negative pulses are applied at 1 second intervals to every type of pin with respect to all types of pins. Table 9.1 shows an example of a possible ESD test sequence for an input pin, an output pin and the supply (V_{dd}) and ground (V_{ss}) pins.

Table 9.1: Different ESD test states

State	DUT	
	pin A	pin B
1	input	V_{ss}
2	V_{ss}	input
3	input	V_{dd}
4	V_{dd}	input
5	output	V_{ss}
6	V_{ss}	output
7	output	V_{dd}
8	V_{dd}	output
9	input	output
10	output	input
11	V_{dd}	V_{ss}
12	V_{ss}	V_{dd}

Of course, a complete sequence includes all pins on the DUT. In this sequence, only positive voltages need to be applied to the capacitor of the human body model. The required negative pulses are achieved by reversing the pin connections to the model's terminals.

Stressed pins are tested after application of each ESD pulse series. If no failure is observed for a sequence through all pins, then the ESD voltage level is increased by 100 V and the sequence is repeated. The process continues until a failure occurs or the required maximum voltage of 2000 V or more is reached.

Sometimes, even values of 3000 V to 4000 V are required for certain applications. The ESD test is complete when a failure is observed or when all pins on the DUT have been stressed as described. Generally, the following criteria are used to determine failure:

- An increase of 100 mV at a current of 5 mA in the voltage across the resistor R in figure 9.4.
- A change of more than 5% in the forward voltage drop and breakdown voltage of the diode characteristic.
- An increase of more than 10% in the I_{ddq} leakage current.
- Incorrect functional operation or a violation of device specifications.

The MOS input protection circuit

MOS input protection circuits usually comprise a voltage spike filter and diode clamps. A typical implementation is shown in figure 9.5.

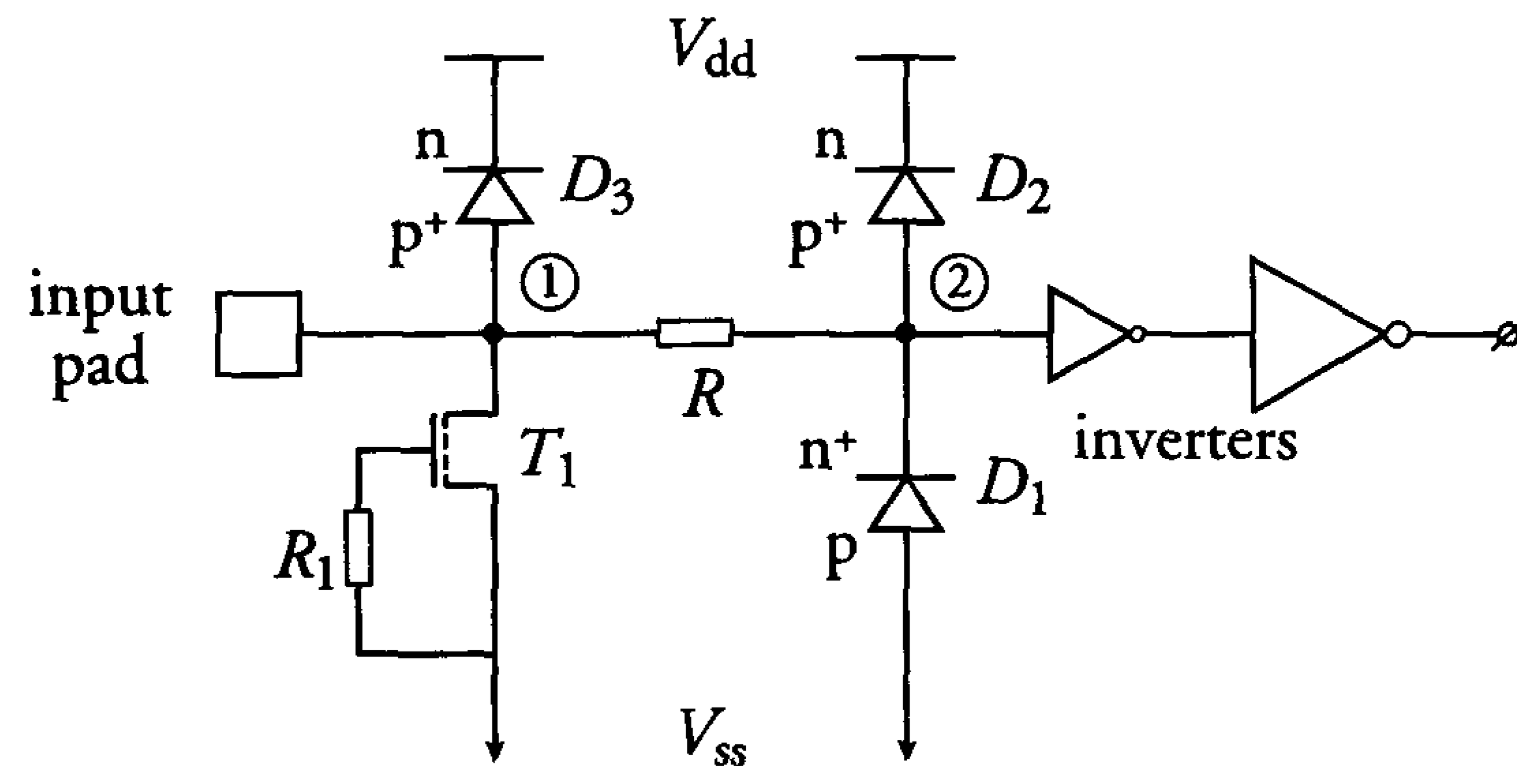


Figure 9.5: *Typical MOS input protection circuit*

The effective capacitance at node 2 comprises parasitic track and pn-junction capacitances. These capacitances and the resistor R form a low-pass filter, which protects the internal circuit from spikes on the input signal. The MOS diode T1 protects the internal circuit from input voltages of excessive negative amplitude. The bipolar diode D3 is forward biased when the input voltage exceeds V_{dd}. The voltage on node 1 is then clamped, or fixed, at $V_{dd} + 0.7\,V$. This internal voltage is clamped to $V_{ss} - V_{T_n}$ for input voltages below $V_{ss} - V_{T_n}$.

Resistor R also serves as a current limiter; its value depends on the type of input for which the protection circuit is intended. For normal signals, R will be between $100\,\Omega$ and $1\,k\Omega$. Clock inputs, however, often have to drive large capacitances at high frequencies. Here, the value of R must be between $50\,\Omega$ and $300\,\Omega$ to ensure that the RC delay is not excessive. Diodes D1 and D2 serve as additional protection and clamp the voltage at node 2 at $V_{ss} - 0.7\,V$ and $V_{dd} + 0.7\,V$, respectively.

The behaviour of a MOS input protection circuit depends very much on its size and layout and on various process parameters. Therefore, the design of such circuits should be done in co-operation with specialists in the field of protection devices. Only one protection device is necessary for each process generation.

The MOS output protection circuit

High currents prohibit the use of a relatively large series resistance in MOS output protection circuits. Normally, the output transistors incorporate such large diffusion n^+ and p^+ areas that there is no need for additional protection. An example of such a CMOS output driver is shown in figure 9.6.

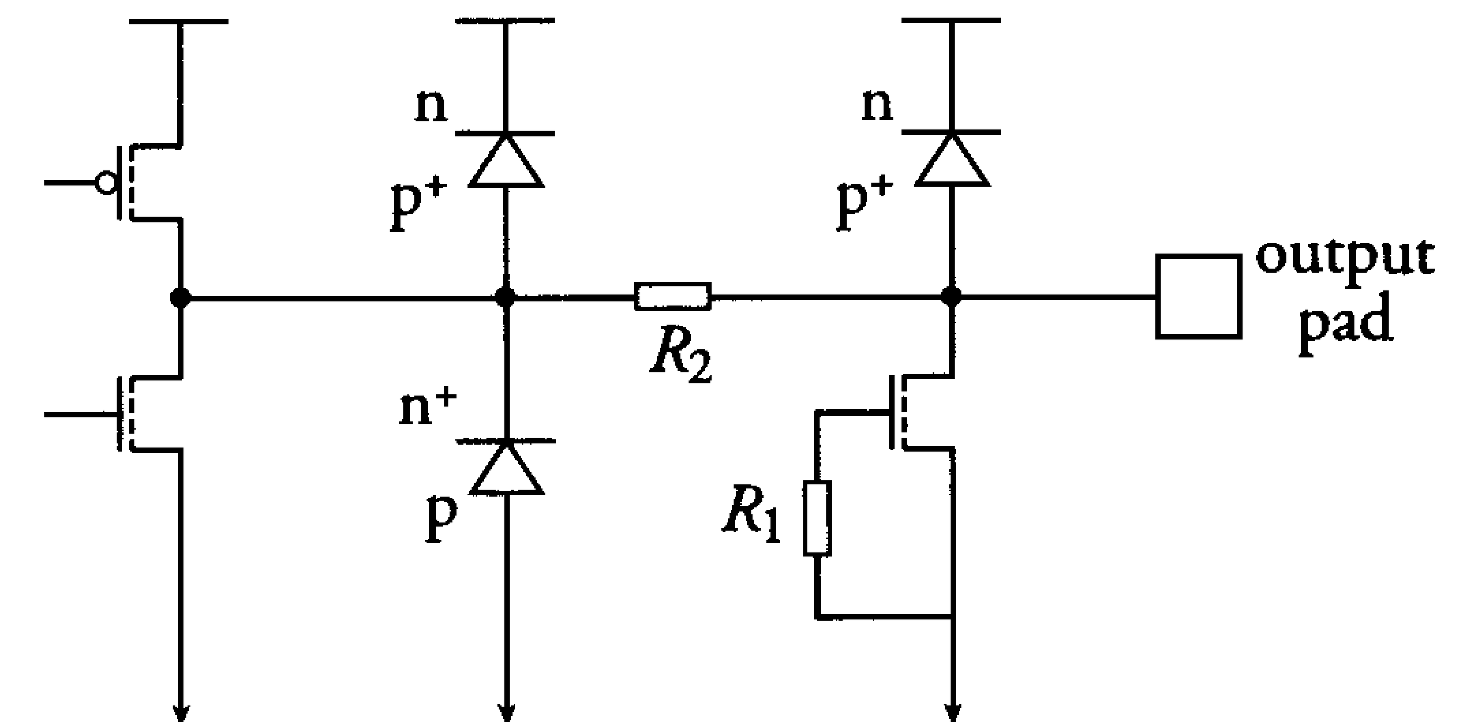


Figure 9.6: *Typical CMOS output driver stage with protection circuit*

The p-type substrate and the n^+ drain area of the lower nMOS transistor form a diode junction which clamps negative voltage spikes to about $V_{ss} - 0.7\,V$. Analogous to this, the p^+ drain of the pMOS and the n-well together form a diode junction, which clamps positive voltage spikes to about $V_{dd} + 0.7\,V$. When the pMOS transistor is used in the CMOS output driver, the increased risk of latch-up requires extra attention. The additional components, such as the bipolar diode, the MOS diode and both resistors, serve to improve the quality of the output protection circuit.

Each manufacturing process has its own specific design rules for ESD protection of CMOS output buffers. These rules should be consulted before the design of an output buffer is started. There are also some more or less general design rules which greatly reduce the risk of damage to an output by an ESD pulse. These rules are defined below for an output buffer in a CMOS process without silicided, or salicided, source and drain areas. An output transistor which adheres to these rules is shown in figure 9.7.

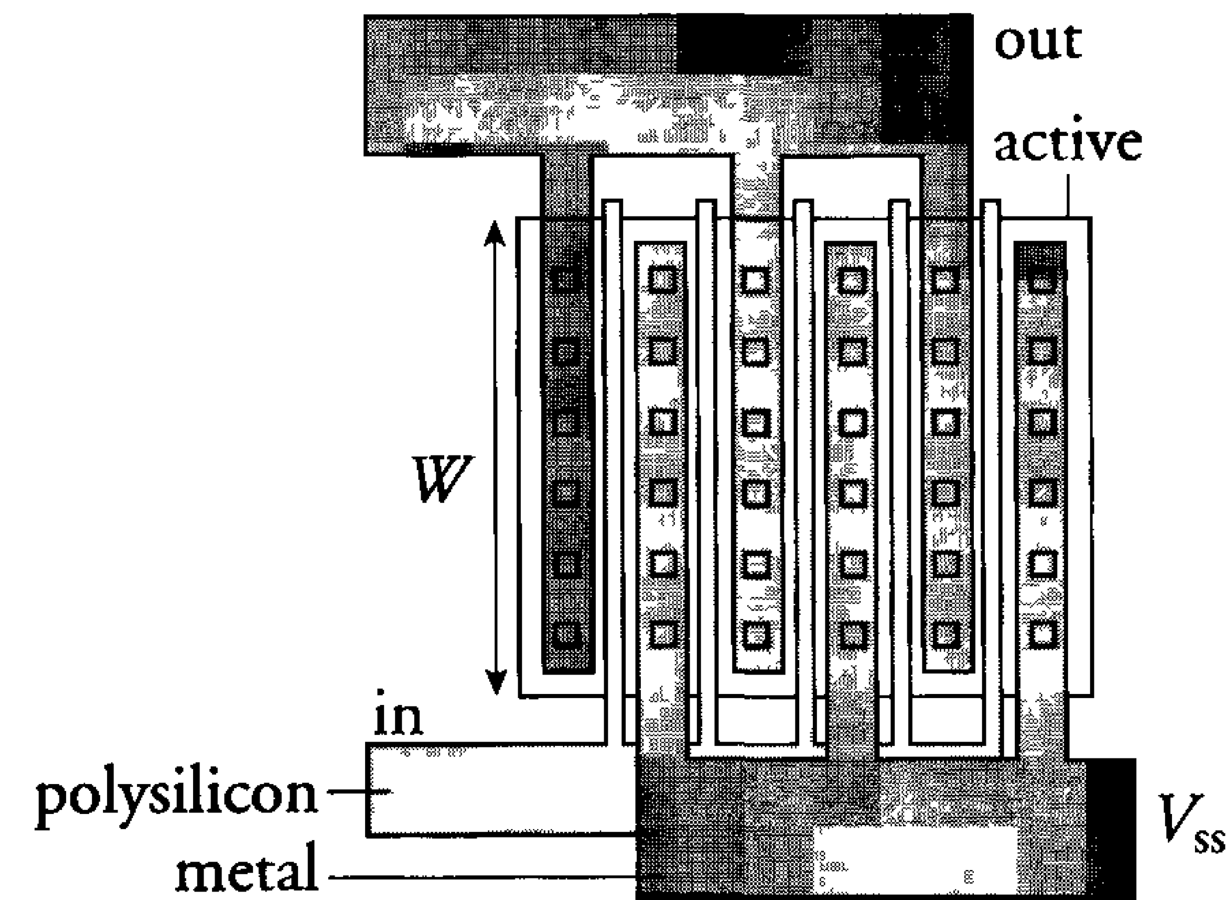


Figure 9.7: *Typical layout configuration of an output transistor*

Some general design rules:

- The channel length of the output transistors should be minimum, as a result of the better activation offered by the npn transistor.
- Many large parallel transistors are used to achieve good current homogeneity. The width W of these transistors must be larger than a minimum value.
- There are restrictions on the number and positions of contact holes required for uniform output current distribution.
- The layout of the output may not comprise a matrix. The gates of the output transistors should be parallel.
- ESD performance is increased by not using silicide in the active areas of an output transistor. Silicide makes current crowding more difficult.
- The aluminium of the output node should not cross the gates of the output transistors. This prevents capacitive coupling, which could damage other parts of the internal circuit.

Generally, the above design rules should yield output buffers that meet the specifications of the MIL standard 883c model for ESD testing. However, similar to input protection circuits, the design of output protection circuits should be done in co-operation with specialists. Standard input

and output circuits that include the necessary protection devices are often available for a chosen process, and IC developers are usually obliged to use them.

9.2.4 Electromigration

The increase in current density associated with scaling may not only have a detrimental impact on circuit performance, as previously discussed, but also on the IC's reliability. The flow of the resulting high currents with their many charge carriers may cause metal ions to be transported through the interconnection layers. Because the current physically moves material from a certain location to another location, we get open circuits or voids on locations where material is removed, and hillocks where material is added. This *"electromigration effect"* damages the layer and results in the eventual failure of the circuit. Electromigration therefore dramatically shortens the lifetime of an IC.

Figure 9.8 shows photographs of the electromigration effect on a metal track. The three photographs were taken during an accelerated electromigration stress, after 1 minute, 5 minutes and 9 minutes, respectively.

Electromigration is avoided by ensuring that current densities do not become excessive. A simple rule of thumb states that the maximum current density permitted in an aluminium track is roughly $1\,\text{mA}/\mu\text{m}^2$ at 125 °C. Therefore, depending on the thickness of the track, different maximum currents are allowed in the different metal layers. An example of values for the thickness of three metal layers in a $0.25\,\mu$m CMOS process is: metall: 600 nm, metal2 to metal5: 700 mm and metal 6: 1000 nm.

The above maximum currents are average DC currents. AC currents and peak currents have a much lower electromigration effect. However, the maximum allowed peak current is about 25 times the corresponding maximum average current.

Also, currents through contact holes and vias are limited to avoid electromigration of the contact conductor. As contacts and vias in a $0.25\,\mu$m process are about $0.4 \times 0.4\,\mu\text{m}^2$, a good value for the maximum current here is about 0.5 mA/contact at 125 °C. The peak current for contacts is also limited to 25 times the corresponding maximum current.

Current design tools hardly include the electromigration requirements with respect to track widths. It is therefore always the designers' responsibility to check the widths, especially of supply, ground and clock lines

to guarantee a robust design, not only with respect to electrical performance but also with respect to reliability and lifetime of the IC.

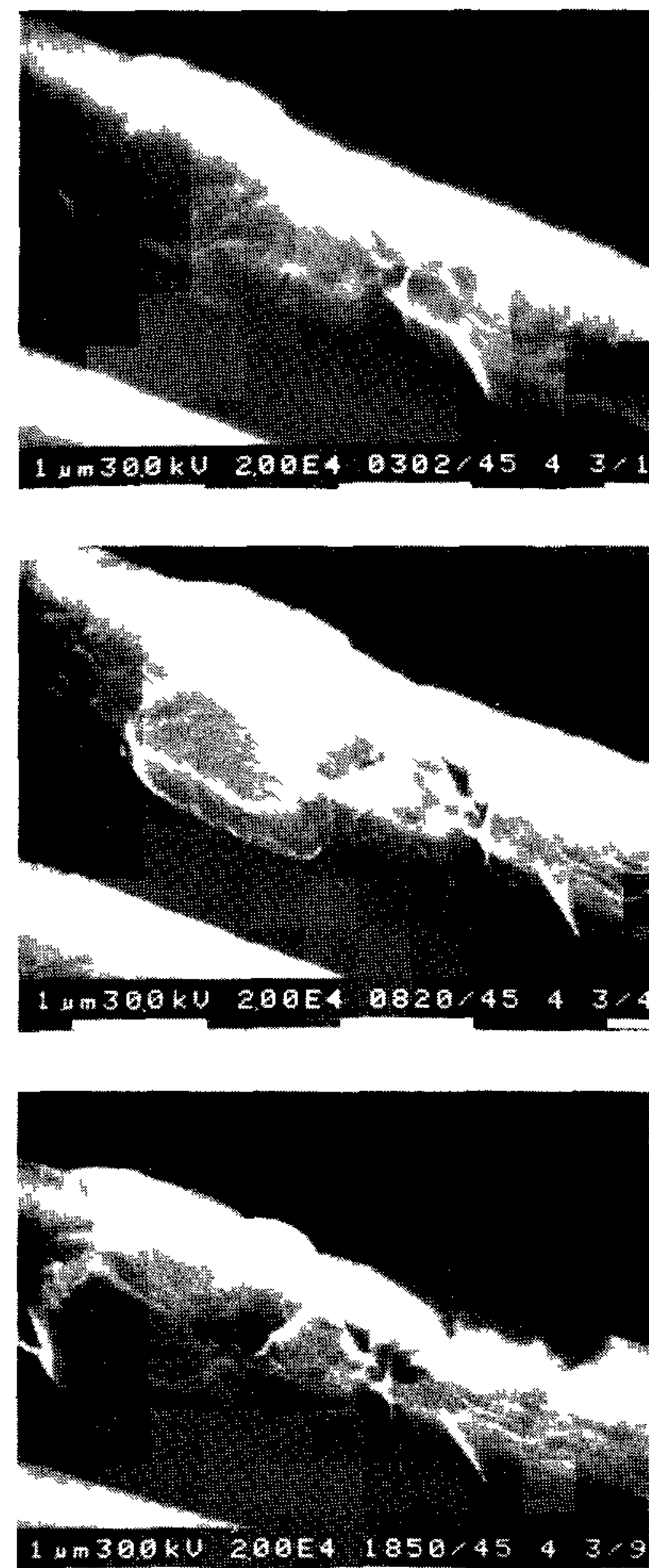


Figure 9.8: *Photographs of the effect of accelerated electromigration stress on a metal track*

9.2.5 Hot-carrier degradation

Impact ionisation can generate high-energy electrons and holes, which can cross the $Si\text{-}SiO_2$ interface barrier when their energy is at least 3.2 eV and 3.8 eV respectively. This injection can cause degradation of transistor current-voltage characteristics. This is called hot-carrier degradation, and is more extensively discussed in chapter 2.

9.3 Design for signal integrity

9.3.1 Introduction

The increase in complexity of ICs over the last decades has enabled the integration of millions of transistors on one single die. This has resulted in complete systems on a chip (SOC). Not only the functionality of a board, but also its problems concerning clock skew, supply network and supply decoupling, interference and EMC will be integrated on the chip as well. The increased manifestation of these effects on a chip is threatening the signal integrity of deep-submicron ICs in different ways. Many of the problems are related with the back-end of the manufacturing process: the formation of the interconnections, which starts dominating the IC's performance and signal integrity. This section tries to evaluate these problems and offers measures to maintain signal integrity at a sufficiently high level.

9.3.2 Clock distribution and critical timing issues

Very complex designs may contain over several millions of transistors on silicon die areas of one to several square centimetres. Most VLSI designs contain synchronous logic, which means that data transfer on the chip is controlled by means of one or more clock signals. These clock signals are fed to latches, flip-flops and registers, which temporarily store data during part of the clock period.

Current VLSI chips may contain several thousands to several hundred thousands of these latches and the total wire length of the clock signals may exceed several metres. To achieve high system performance, the clock frequency is often maximised. This combination (a large clock load and a relatively high clock frequency) is the cause of many on-chip timing problems. There are many different clocking strategies for synchronous logic.

The following sections discuss potential timing problems, most of which are related to the clock signals.

Single-phase clocking

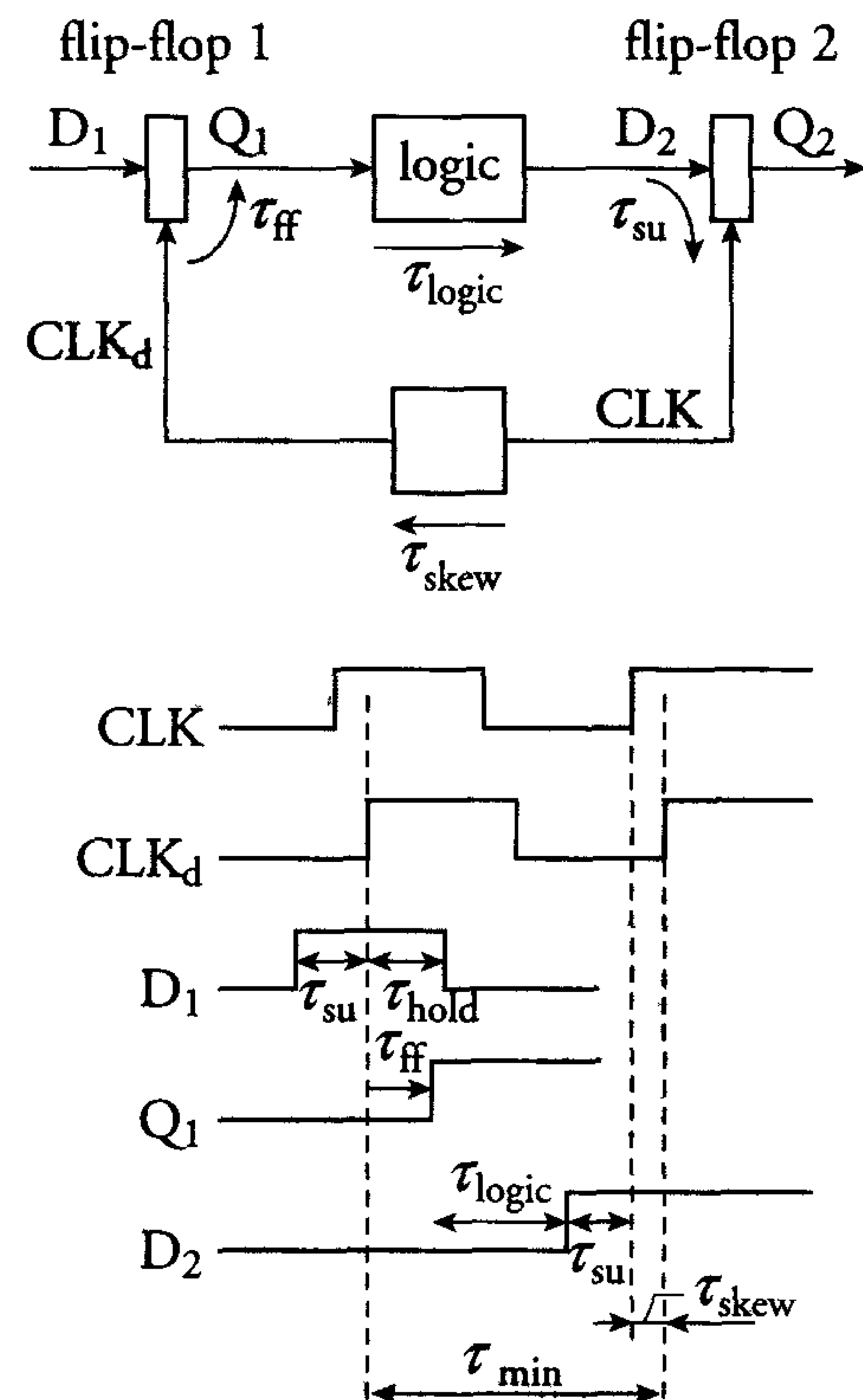


Figure 9.9: *Single-phase clock system and its timing diagram*

From figure 9.9, we can derive that the minimum cycle time is given by:

$$\tau_{\min} = \tau_{\text{ff}} + \tau_{\text{logic}} + \tau_{\text{su}} + \tau_{\text{skew}} \tag{9.1}$$

where

τ_{ff} is the flip-flop delay from clock to output,
τ_{logic} is the propagation delay through the logic and
τ_{su} is the setup time of the data of flip-flop 2.
τ_{skew} is the maximum amount of time that the clock of flip-flop 2 can be earlier than that of flip-flop 1.

Especially τ_{logic}, which is dominant in equation 9.1, must be carefully simulated to be sure that the required frequency (clock period) will be achieved.

In many designs, the (pipe line and/or scan) registers are implemented by using series connections of flip-flops. Especially in the scan mode during testing (see chapter 10), flip-flops are directly connected to other flip-flops. In figure 9.10, a flip-flop of logic block 1 can be directly connected to a flip-flop of logic block 2.

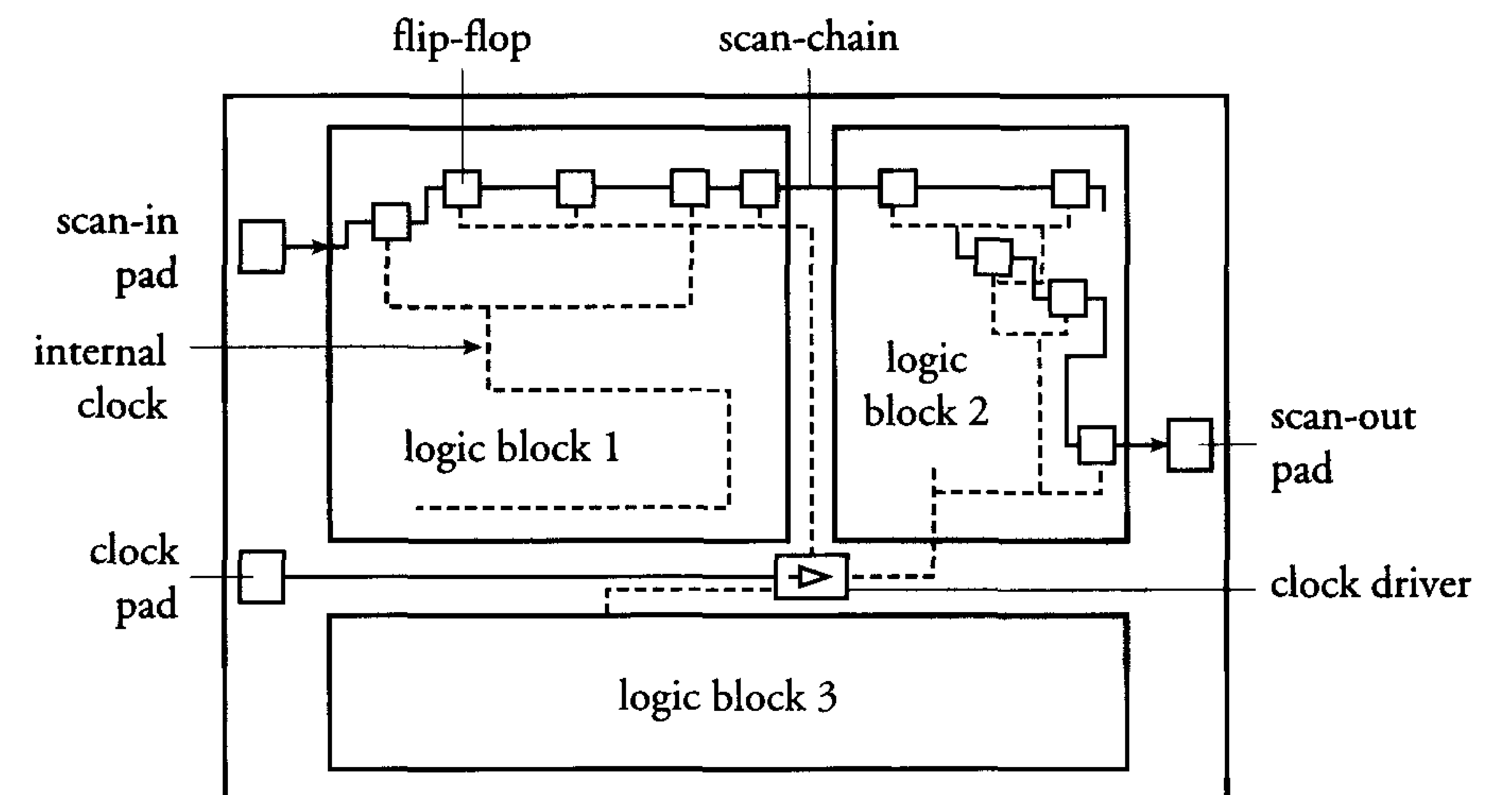


Figure 9.10: *Example of a scan-chain in a complex VLSI circuit*

With a direct connection, the propagation time of the data between these flip-flops can be very short. As the clock is routed through these blocks automatically, its time of arrival at the first flip-flop of logic block 2 can be later than the arrival time of the data. This will result in a race, which can also occur in two-phase clocked registers. Therefore, each (scan) register should be carefully checked with respect to the above critical timing situation. If necessary, additional delay by using several inverters should be included in the critical path in the scan chain. The above is an example of a race between flip-flops.

Generally, there is a variety of single-phase clocked flip-flops in a library. As many of these flip-flops need two clock phases, one or both are generated inside the flip-flop by means of inverters. Figure 9.11 shows an example of a race within a single-phase clocked flip-flop.

Figure 9.11: *Example of single-phase clock with critical timing*

Although the flip-flop uses two clock phases internally, there is only one clock routed to that flip-flop. When the ϕ clock is routed throughout the logic block, its signal shape may look like a sine wave rather than a square signal. In this example, the master clock is delayed with respect to the slave clock. Large rise and fall times of the ϕ clock could increase the $\overline{\phi}$ delay such that the $\overline{\phi}$ clock may stay on for a while after ϕ has switched to high level.

If this delay, $\tau_{\overline{\phi}}$, is larger than that of inverter 1, τ_{invl}, the flip-flop becomes transparent and the data can ripple through it at one clock edge (race). This kind of flip-flop should not be used in a proper design because the (routed) clock signal will almost always look more like a sine wave than a square signal. A more robust version of this flip-flop would be one in which both internally-used clocks are also generated internally by means of two inverters. In this way, both internal clock signals would have a square shape.

Different versions of flip-flops will be available in a library, each with its own characteristics in terms of area, power and integrity. Generally, the one with the largest number of transistors has the highest signal integrity. With regard to the previous characteristics, the flip-flop depicted in figure 9.12 is a nice one.

Figure 9.12: *Efficient single-phase clock flip-flop*

This flip-flop, which is a modified version from [1], can also be used in a low-power design: it has a small number of transistors, it is controlled by a single-phase clock and it has a relatively low clock load. Here also, the clock may be generated locally in the latch by means of an inverter.

Two-phase clocking

One way to design synchronous systems with potentially less critical timing is to use two-phase clock flip-flops. The non-overlapping property of the clocks prevent transparency either within a flip-flop or between flip-flops. The non-overlapping time between the clocks in figure 9.13 allows larger clock skew, however at the cost of performance.

Figure 9.13: *Two-phase clock system*

The non-overlapping time must be added to the cycle time and it therefore reduces the performance. Because two clock signals must be routed through the chip, the routing area of the chip will be increased with respect to single-phase clocking.

Figure 9.13 shows a synchronous two-phase clock system. When ϕ_1 is high, the master is listening and the slave is talking. When ϕ_2 is high, the master is talking and the slave is listening. When the difference in clock delays (clock skew) is larger than the $\tau_{\text{non-overlap}}$, both master and slave will be talking or listening at the same time. This results in bad communication and incorrect operation.

Additional margin can be created by increasing the non-overlapping time. In a two-phase clocked system, the clock signals are usually generated by a central clock generator circuit. Figure 9.14 shows an example of a two-phase clock generator circuit.

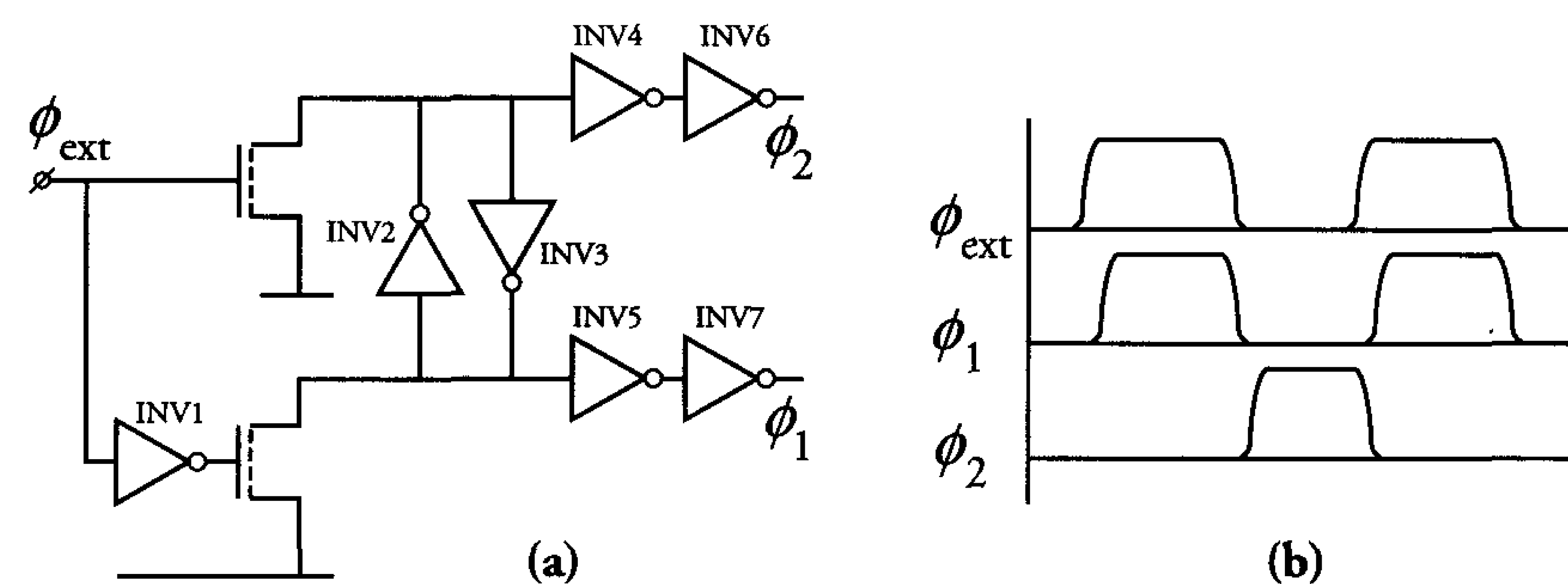


Figure 9.14: *(a) Two-phase clock generator (b) Clock waveforms*

Particularly the sizing of inverters 4,5,6 and 7 is used to define (or change) the non-overlapping time between ϕ_1 and ϕ_2.

Causes of clock skew

The previous subsections show various situations in which clock skew can dramatically reduce the integrity of operation of VLSI circuits. There are two sources of clock skew: different lengths of clock paths and/or different loads of the clock drivers.

Different lengths of the clock paths cause different total resistance and capacitance of the clock tracks. This results in different RC times for the clock signals. Although these clocks are routed in metal, they can still show a mutual clock skew of hundreds of picoseconds to even several nanoseconds in a very badly routed clock system. Care must be taken because most flip-flops in a library show clock skew sensitivities of only several hundreds of picoseconds to 1 ns. Control of clock routing in modern tools is therefore a must to guarantee proper operation of synchronous VLSI designs. Different loads of the clock phase drivers is another potential cause of clock skew. Figure 9.15(a) shows a distributed clock driving network.

Although clock signals ϕ_1, ϕ_2 and ϕ_3 should be equal, they may show some differences because of different loads. This can be originated by a different number of latches and flip-flops that they have to drive, or by a different capacitive load of the tracks.

Suppose capacitance C_3 (which represents all latches and flip-flops connected to ϕ_3) is larger than C_2. When Q_2 is fed to the S input of FF3, it might simultaneously ripple through this flip-flop as well, when its clock (ϕ_3) is delayed to a certain extent with respect to ϕ_2. Such anomalous circuit behaviour reduces the voltage margin with which the circuit should normally operate.

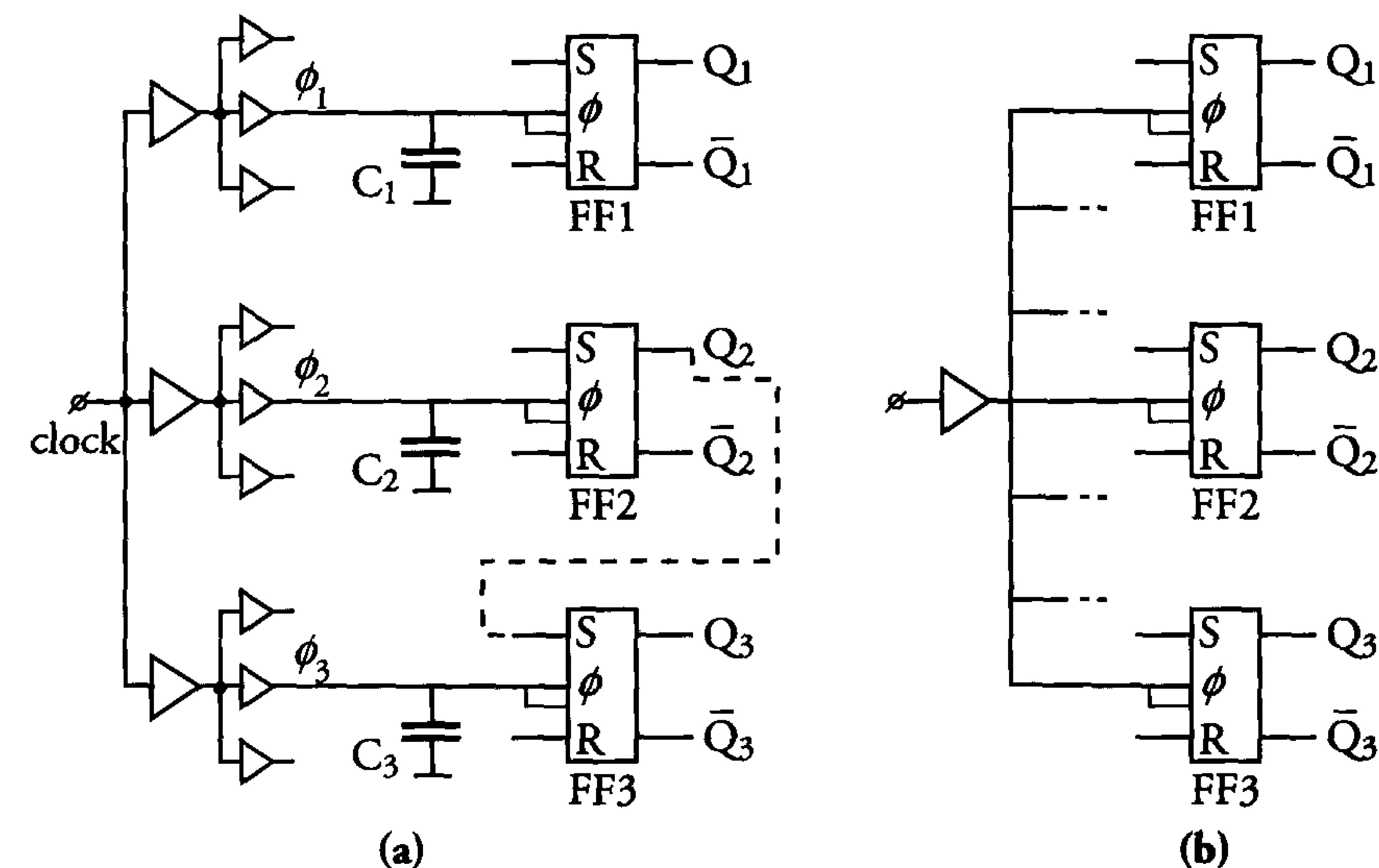


Figure 9.15: *(a) Distributed clock driving network (clock tree approach) (b) Clock trunk approach*

In the clock tree approach, it is extremely important that the clock branches are equally loaded (balanced clock tree). This must be verified by tools, particularly in high performance complex circuits. Current

tools offer a well-balanced clock tree synthesis, which enhances the quality of clock timing. An important advantage of this clock tree approach is the distribution of the different small clock drivers over the logic blocks. The use of distributed clock drivers also puts the clock drivers right there where they are needed. Distributed clock drivers keep the current loops short and they also do not switch simultaneously, but distributed over a small time frame. Moreover, they can use the intrinsic decoupling capacitance which is available in a logic standard cell block. This reduces the dI/dt fluctuations, which are responsible for most of the supply/ground bounce in VLSI designs.

In many designs, a clock trunk approach (see figure 9.15b) is still used, with only one global clock. In many synchronous designs, the total dissipation of the clock-related circuit may vary from 10% to even 50% of the total IC dissipation. It is obvious, then, that the clock system will also generate a large part of the total supply bounce. On the other hand, clock skew is now only determined by the delay across the clock wire. However, for design integrity, the balanced clock-tree approach is preferred.

Especially large library blocks (such as memories and microprocessor cores) may include their own clock drivers and/or routing. If this is the case, communication between these blocks and the rest of the chip may become very critical and must be thoroughly verified and simulated. Because these blocks have different internal clock delays, these delays must be compensated for in the clock architecture. This can either be done by using compensating delays (which must be simulated per logic or memory block) or by using a PLL per logic or memory block, which can synchronise the clock arrival times at the flip-flops. The clock phase synchronisation and clock skew of these blocks is discussed in section 9.3.3.

Other timing problems

Particularly in low-power CMOS circuits, some logic blocks (or sometimes even the complete chip) may often be inactive for certain periods of time. Such a chip may contain different clock domains, of which the mode of operation (active or stand-by) is controlled by a gated clock. In many cases then, the main clock is used as input to some logic gate to perform some logic function on the clock signal (gated clock). Figure 9.16 shows an example:

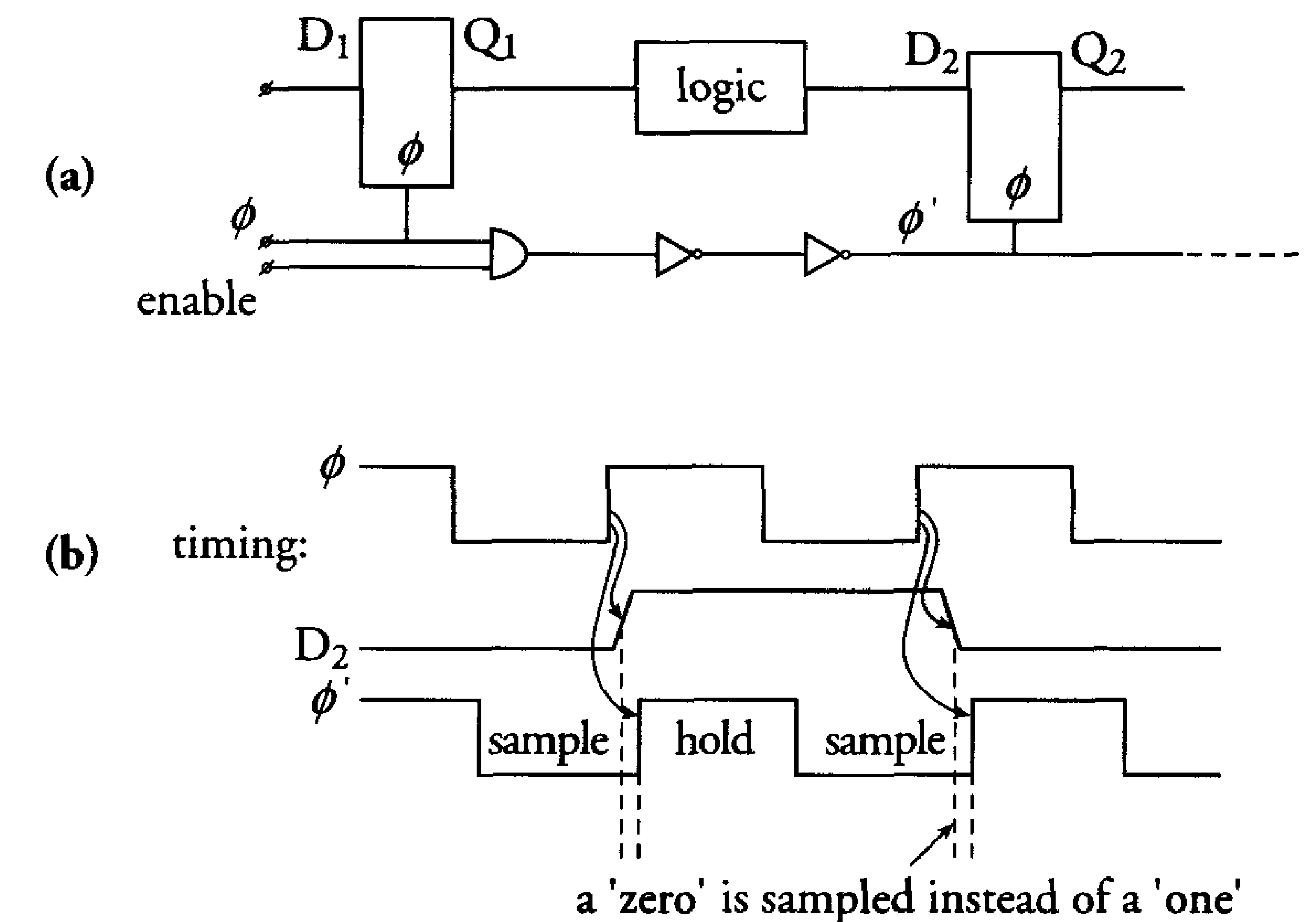


Figure 9.16: *(a) Example of a local clock-enabled circuit and (b) The corresponding timing diagram*

When the delay between the clock ϕ and clock ϕ' is longer than the data delay between the output Q_1 of one flip-flop and the input D_2 of the next flip-flop, this "new" data sample will be clocked into this flip-flop by the "old" clock and a race will occur.

Such clock-enabled signals are also often used in the design of memory address decoding circuits and are very critical with respect to timing margins.

Finally, timing problems could also occur when the data delay (caused by the logic and interconnection delay) between two successive latches or flip-flops becomes equal to or larger than one clock period. Figure 9.17 shows an example. When the total propagation time through the logic from Q_1 to D_2 exceeds the clock period, the data at D_2 can arrive after the sample period of flip-flop 2 has been terminated. It will then be sampled in the next clock period, resulting in improper synchronisation. Timing simulation to find critical delay paths is therefore a must in CMOS VLSI design and is part of the design flow.

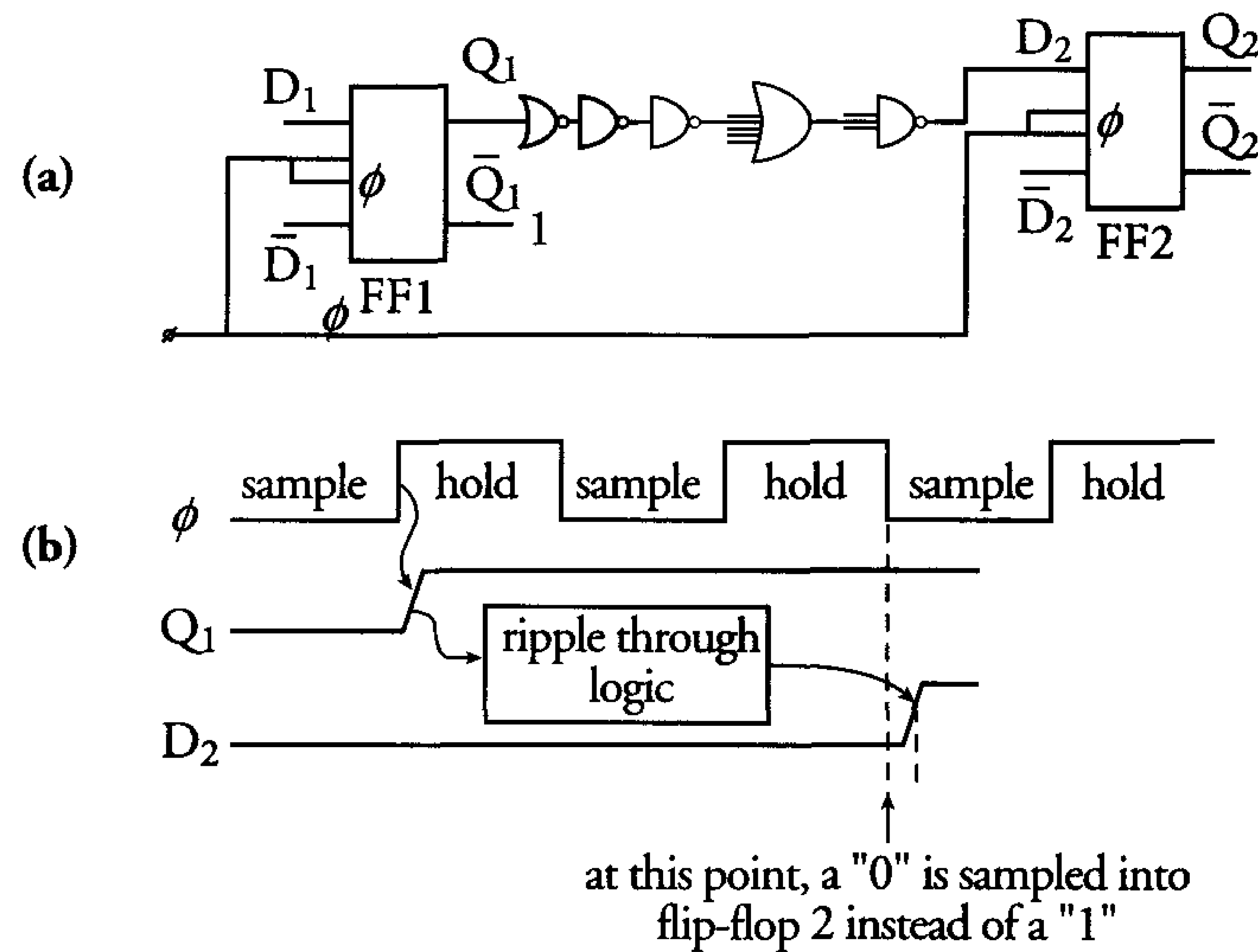


Figure 9.17: *(a) Example in which the data delay exceeds a clock period and (b) Its corresponding timing diagram*

Slack borrowing and time stealing

When a data path uses more than a full clock cycle in a single clock system, or more than half a clock cycle in a two-phase clock system, this is referred to as *cycle stealing.*

Slack borrowing refers to the case where a logical partition utilizes time left over (slack time) by the previous partition [10]. Important to note is that it can be used without the adjustment of circuitry and/or clock arrival times. This precludes the use of edge-triggered circuitry (dynamic logic and flip-flops). *Time stealing* refers to the case where a logical partition steals a portion of the time allotted to the next partition. This can only be obtained by adjusting the clock arrival time(s). Using one of these concepts to solve timing problems in (ultra) high-speed designs forces the designer to match certain design rule requirements. A well documented list of such design rules can be found in [10].

Source-synchronous timing (clock forwarding)

In a source-synchronous interface, data and clock signal propagation between transmitter and receiver are matched. This technique is currently used in high-performance microprocessors and SDRAM interfaces

[11,12], but is also a potential candidate for on-chip chip time-of-flight compensation.

9.3.3 Clock generation and synchronisation in different (clock) domains on a chip

With IC complexities exceeding tens of millions of transistors, the total effort required to complete such complex VLSI designs is immense. This stimulates the reuse (IP) of certain logic blocks (cores) and memories. Current heterogeneous systems on chip may not only incorporate many clock domains, but can be built up from cores that are designed at different sites or vendors with different specifications. Because each core has a different clock skew from the core's clock input terminal to the farthest away flip-flop, the clock phase of each core has to be synchronised with the main clock. This subsection discusses the generation of multiple clocks and the synchronisation of clocks in systems that use different cores.

On-chip multiple clock generation

On-chip multiples of the clock can be generated by phase-locked loops (PLLs). Figure 9.18 shows a basic phase-locked loop concept.

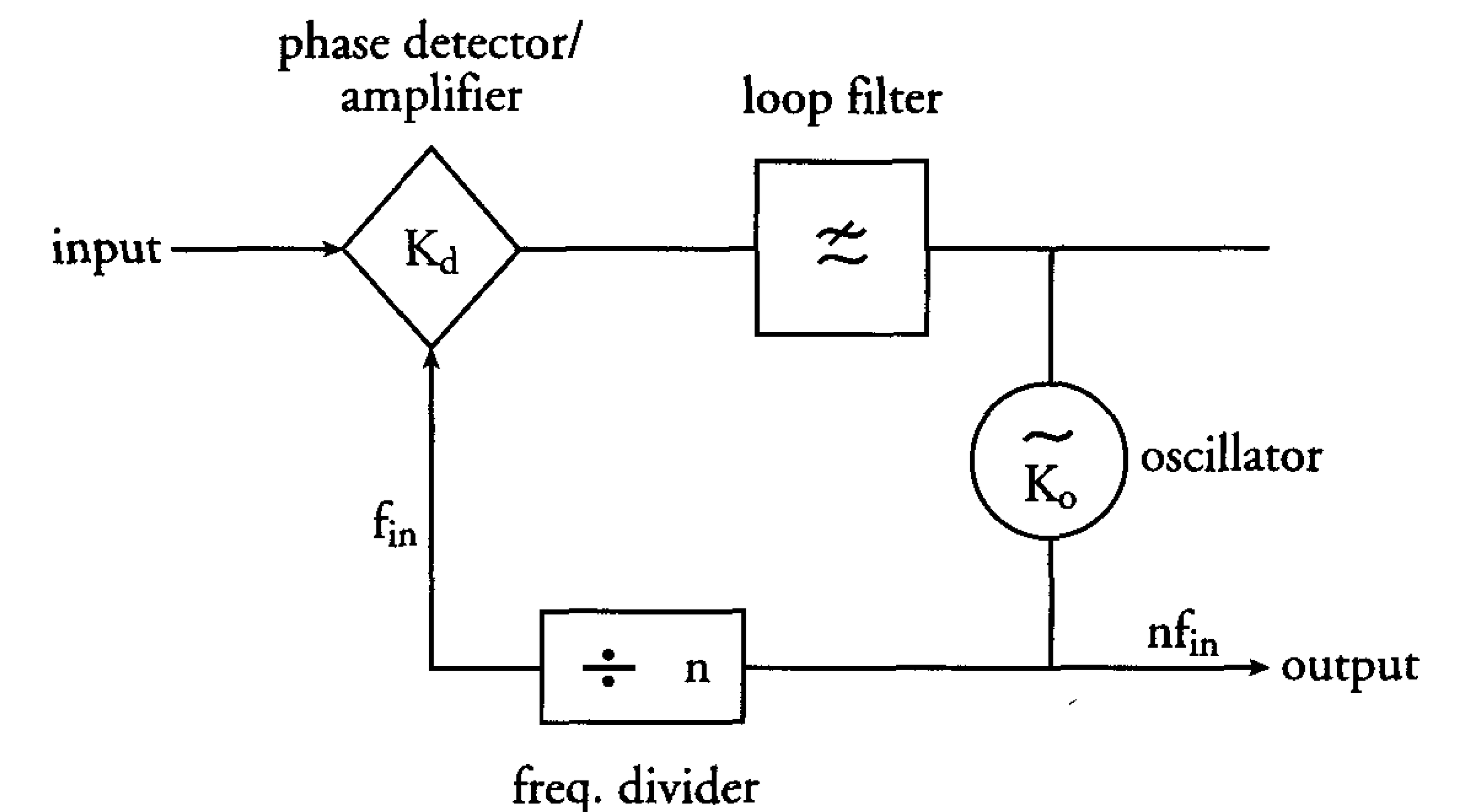


Figure 9.18: *Basic concept for a phase-locked loop*

The Voltage-Controlled Oscillator (VCO) - current-controlled oscillators (CCOs) are also used - is basically an oscillator whose frequency is determined by an externally applied voltage. This frequency is a multiple of

that of the input. The phase detector is sensitive to differences in phase of the input and VCO signals. A small shift in the frequency of the input signal changes the control voltage of the VCO, which then controls the VCO frequency back to the same value as that on the input signal. Thus, the VCO remains locked to the input. Based on this principle, a PLL can be used to generate an output frequency which is a multiple of the input frequency. The output frequency equals n times the input frequency. As current complex ICs require many different clock domains, multiple frequencies must be generated on chip. Figure 9.19 shows an example of a multi-clock generator based on PLL.

Figure 9.19: *Multi-clock generator, based on PLL*

In this example, the PLL output frequency equals $n \times m_1 \times f_{\text{in}}$. Using different divisions (m_i), many different clocks can be generated. The PLL, by nature, automatically makes these clocks in phase with the input.

Clock-phase synchronisation in multiple core environments

Because of differences in the clock arrival times at the flip-flops of different cores, these delays must be compensated for, to allow proper communication between different cores. There are several methods of synchronising the clock phase at the actual flip-flops in each core.

The first method is *adaptive skew control*. In this approach, the clock network of each core (domain) is extensively simulated. The clock skew in each core is then made equal to the worst case clock skew by using a chain of inverters. The length of this inverter chain is then adapted to the required additional delay in the specific core clock path.

The second method uses the PLL concept. The PLL property of locking one signal phase to the phase of another reference signal makes the PLL also suitable for the compensation of clock skew in different cores, see figure 9.20.

Figure 9.20: *Clock phase synchronisation to compensate for different clock skews inside different cores*

Node A represents the clock terminal of the core, and node B represents the clock terminal of the actual flip-flop in that core. The clock phase at the flip-flop will then be locked to the input reference signal, which is usually the chip's main clock. In this way, the clock tree delay (which might be different in all cores) can be easily compensated for. Moreover, when the frequency dividers in figure 9.20 are made programmable, then the same PLL can be used in all cores, even when they run at different frequencies.

Sometimes, reusable cores are only available with fixed instances and only in GDSII (layout description) format. In these cases, the clock tree must be thoroughly simulated and a delay chain, which mimics the core's internal clock delay path, replaces the clock tree between nodes A and B (figure 9.20) in the feedback path. The PLL must be placed outside the core. Disadvantages with the use of PLLs are:

- Because of high internal frequencies, PLLs can consume relatively high power.
- PLLs are difficult to start and stop.
- Especially the start-up takes a relatively long time.

- Multiple-clock concepts and the use of PLLs for clock generation and synchronisation makes testing very difficult. During testing, such PLLs must be set to the right mode first before the test procedure can be started.

Finally, to synchronise the clock phases to compensate for the different clock skews in different cores, Delay-Locked Loops (DLLs) can also be used, see figure 9.21.

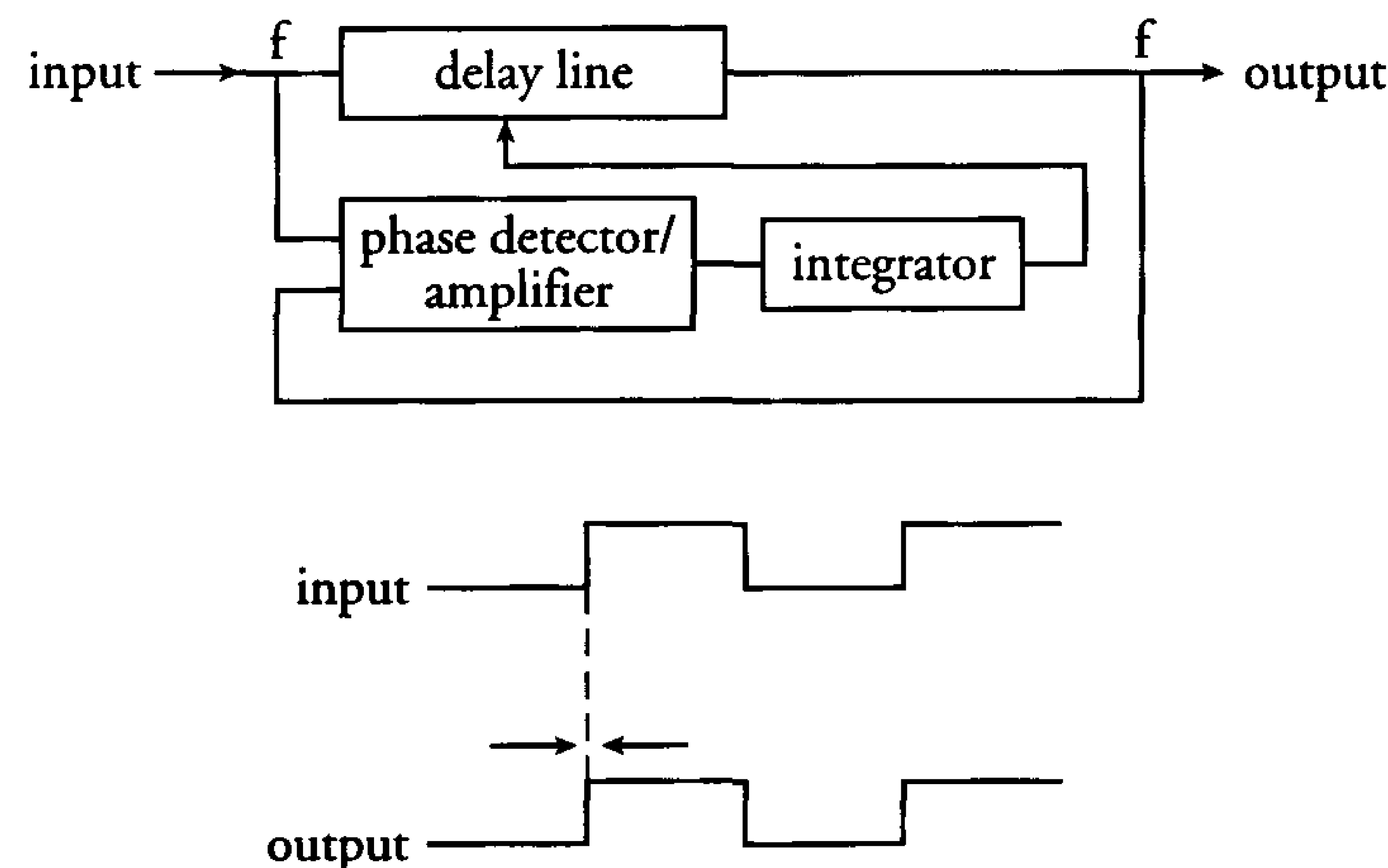


Figure 9.21: *Basic concept of a delay-locked loop and its timing*

The delay of the delay line can be controlled by the output voltage of the integrator. In this concept, the output signal is delayed over one complete clock period with respect to the input. If the delay is less, then the phase detector produces a signal which increases the delay of the delay line, via the integrator. The output signal in such a DLL has the same frequency as the input, and this concept of the DLL cannot be used to multiply the frequency.

Because the VCO or CCO in a PLL generates frequencies that depend on the supply voltage, clock jitter can occur when there is supply noise. Also, the delay in a DLL is susceptible to supply noise. Control of the clock jitter is therefore one of the most important constraints in the design of a PLL and DLL. For the synchronisation of the clock phases of all cores in a heterogeneous chip, each core needs its own PLL (DLL).

9.3.4 Phenomena related to large current fluctuations

Differences between ICs and their packaging are diminishing with the advent of high-performance VLSI chips. Couplings (be they resistive, capacitive or inductive) grow with decreasing feature sizes and increasing chip sizes.

The increase in current densities and driving capabilities of on-chip and off-chip drivers (e.g. clock and output buffers) can have dramatic effects on the chip behaviour. The increased current levels cause larger voltage drops along resistive lines and lead to supply and ground bounce during the simultaneous switching of logic and/or drivers. Also, the self-inductance of the supply and ground lines on chip and their bonding wires may even worsen this effect.

The following subsections discuss the above effects in more detail.

Supply and ground bounce (supply noise)

Over generations of ICs, the bus widths, the number of outputs and the clock load, etc. have increased dramatically. For example, microprocessor bus widths have increased from four to eight bits in the late seventies to 64 or even 128 at the present time. Consequently, the total load capacitance that has to be simultaneously charged or discharged has increased proportionally. As a result of the higher resistance of the power supply network (larger chip and block sizes) and the increased current slew rates (dI/dt), the peak currents can no longer be completely supplied by the power network, without showing large voltage bounces on its supply lines.

The associated voltage drops can be caused by the line resistance and the self-inductance of the board wiring, the bonding wires, the package leads and the on-chip supply lines themselves.

Currently, wide metal lines are used to supply the different functional blocks on an IC. When this is done properly, the influence of resistive voltage drops on the chips' behaviour is negligible.

However, the effect of the self-inductance is more dramatic. According to Faraday's law, any change in the magnetic flux (caused by a fast current change) causes an opposite self-induced electromotive force (*emf*). This *emf* is defined as:

$$emf = -\frac{d\phi}{dt} \tag{9.2}$$

The self-inductance L of a circuit is defined as a constant representing the relation between the total magnetic flux and the current I:

$$L = \frac{\phi}{I}$$

Both these equations lead to the following equation for the electromotive force:

$$emf = -L\frac{\mathrm{d}I}{\mathrm{d}t}$$

The effect of this electromotive force is two-fold. On the chip, it generates voltage drops in supply and ground lines with an amplitude:

$$\Delta V = \frac{\mathrm{d}\phi}{\mathrm{d}t} = L\frac{\mathrm{d}I}{\mathrm{d}t}$$

The self-inductance L is formed by the bond wires and the package leads. The value of L varies per pin with the different lengths of the bond wires (about 10 nH/cm) and with the different bonding structures. A dual in-line IC has the largest lead inductance (2-50 nH), because of the absence of a ground plane (usually) and because of its longer lead lengths. TAB (Tape Automated Bonding) usually performs better (0.5-10 nH) because of the presence of a ground plane and shorter and thicker leads.

Packages using multilayer ceramic substrates with power and ground planes, such as PGAs (Pin-Grid Arrays), also have small inductance values for their pins. In BGAs (Ball-Grid Arrays), the package pins are replaced by small solder balls, which are directly attached to the board. In such direct attachments (including TAB or flip-chip technologies), one level of packaging (bonding, package leads or pins) is eliminated, resulting in very low inductance values. As an example, flip-chip technology, in which the chip is directly attached to the substrate by its solder bump connections, has the lowest inductance values (0.15 - 1 nH).

The increase in total chip capacitance, combined with faster logic and driving circuits, causes very high values of $\mathrm{d}I/\mathrm{d}t$. Currently, supply current changes of a hundred milliamperes per nanosecond to 1 ampere per nanosecond are no exceptions. If a $\mathrm{d}I/\mathrm{d}t$ of 100 mA/ns is to be supplied by only one bonding wire (with a self-inductance $L \approx 10\,\mathrm{nH/cm}$ and a length of 1 cm), then the associated supply or ground bounce is equal to:

$$\Delta V = 10 \cdot 10^{-9} \cdot \frac{100 \cdot 10^{-3}}{10^{-9}} = 1\,\mathrm{V}$$

This effective voltage drop is determined by the number of nodes that switch simultaneously. It temporarily reduces the effective supply voltage because it introduces additional delays, which results in slower circuit operation. When the amplitude of this supply and/or ground bounce is large ($> 1\,\mathrm{V}$), it can manifest itself through glitches in the logic circuits. Especially on ICs with electrically-separated supply networks, the communication between circuits (e.g. core and I/O circuits) connected to different supply domains may be dramatically disturbed. Figure 9.22 shows an example:

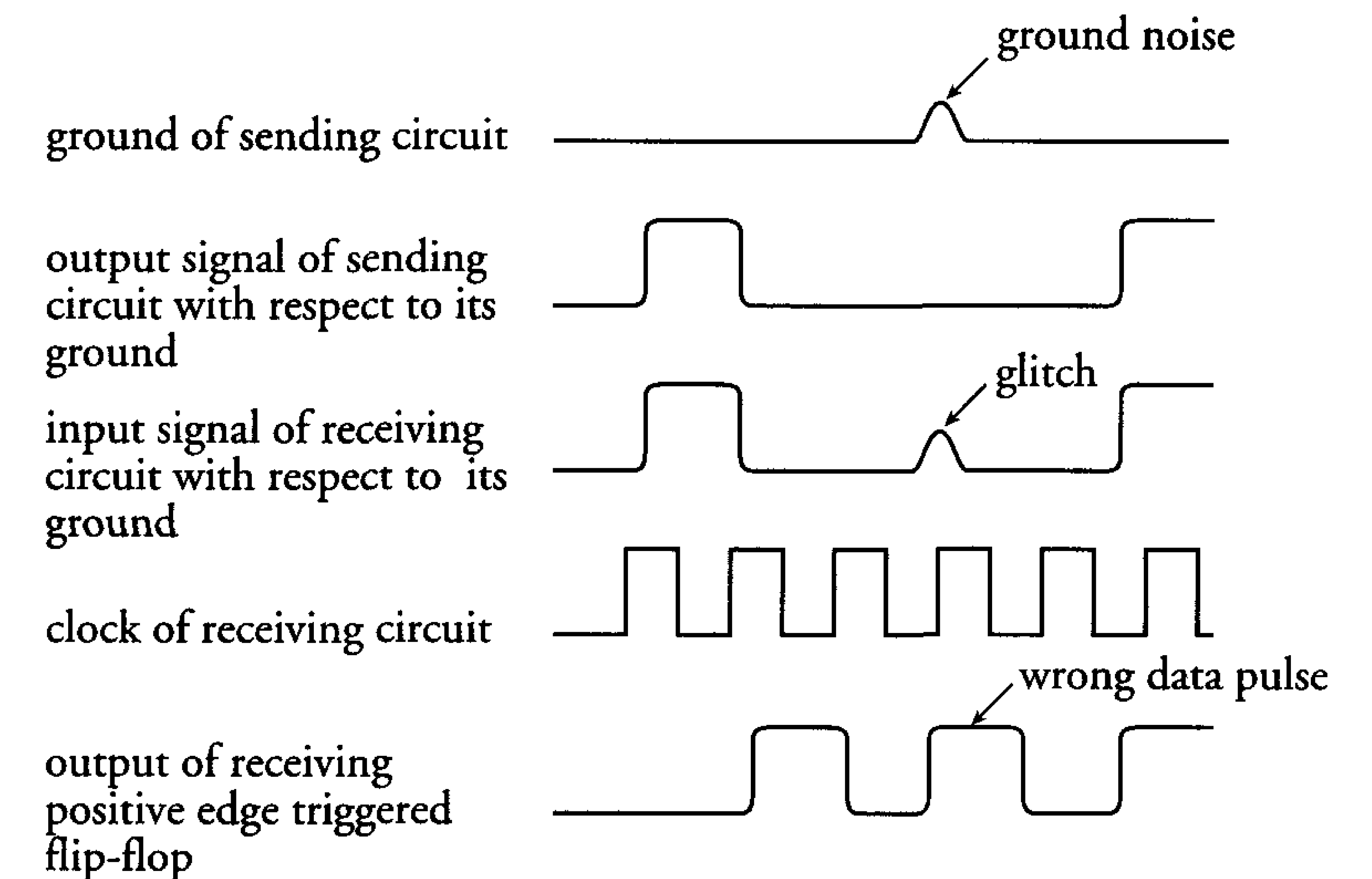


Figure 9.22: *Effect of ground bounce in an IC with multiple electrically-separated supply and ground lines*

A supply or ground bounce also manifests itself in the signals of the connected circuits. If the amplitude of such a glitch is high enough, it may be seen as a high level by the receiving flip-flop, connected to another supply "rail" with low bounce noise, and sampled as such. The original ground bounce pulse is perceived as a glitch at the receiving flip-flop [5].

Input circuits of an IC are also very susceptible to ground bounce. In many applications, these inputs are specified as TTL input, which means that a voltage below 0.8 V must be recognised as a "zero" and voltages of 2 V and higher must be recognised as a "one". A TTL high level of 2 V at the input is only about 1.5 V above the threshold voltage. When a ground bounce of 0.5 V occurs during sampling, the actual input voltage

is only 1 V above the threshold voltage. This can result in anomalous operation of the input circuit, or in an increased set-up time, which no longer fullfils the specification.

The above discussions show the dramatic effect that supply and/or ground bounce can have on the operation of digital circuits. However, the effect on combined analogue/digital circuits is even worse and is of a much more complex character.

Clock generating circuits and oscillators are often based on ring oscillators and/or phase locked loops (PLL). In many of these designs, the basic frequency determinator circuit is a voltage controlled oscillator (VCO). The frequencies of both a ring oscillator and a VCO depend heavily on their supply voltages. Whenever ground and/or supply noise occur, this will be interpreted as a change in the supply voltage and, as a result, a change in the frequency.

In video systems, for example, such instable frequencies appear as clock jitter and synchronisation problems on the screen. To reduce the level of supply and ground noise in digital circuits, on-chip decoupling capacitance must be added. This is discussed later in this section. One alternative is to supply the power via multiple V_{dd} and V_{ss} bond pads. In this way, the self-inductances in the supply lines reduce to:

$$L_{total} = \frac{L}{n}$$

where n is the number of bond pads per supply or ground.

In modern ASICs with up to 350 or more bond pads, tens of pins are used for supply and ground connections.

Design measures can also be applied to prevent the output buffers from switching simultaneously. This is often achieved by switching outputs sequentially at very short time intervals (several picoseconds).

In a chip with 64 outputs (with a fan-out of 30 pF each), the physical distance between the first and the 64th output pad may be in the order of 10 mm. If all these outputs receive their input signal from almost the same location on the chip, it means that the first signal only has to propagate through a short interconnection to the first output pad. The last (64th) one, however, has to propagate through an interconnection wire of about 10 mm. The difference in propagation delay between the first and the last signal is then about 225 ps (in a 0.25 μm CMOS technology).

Figure 9.23 shows the simulation results of such a tailored driver switch-on network:

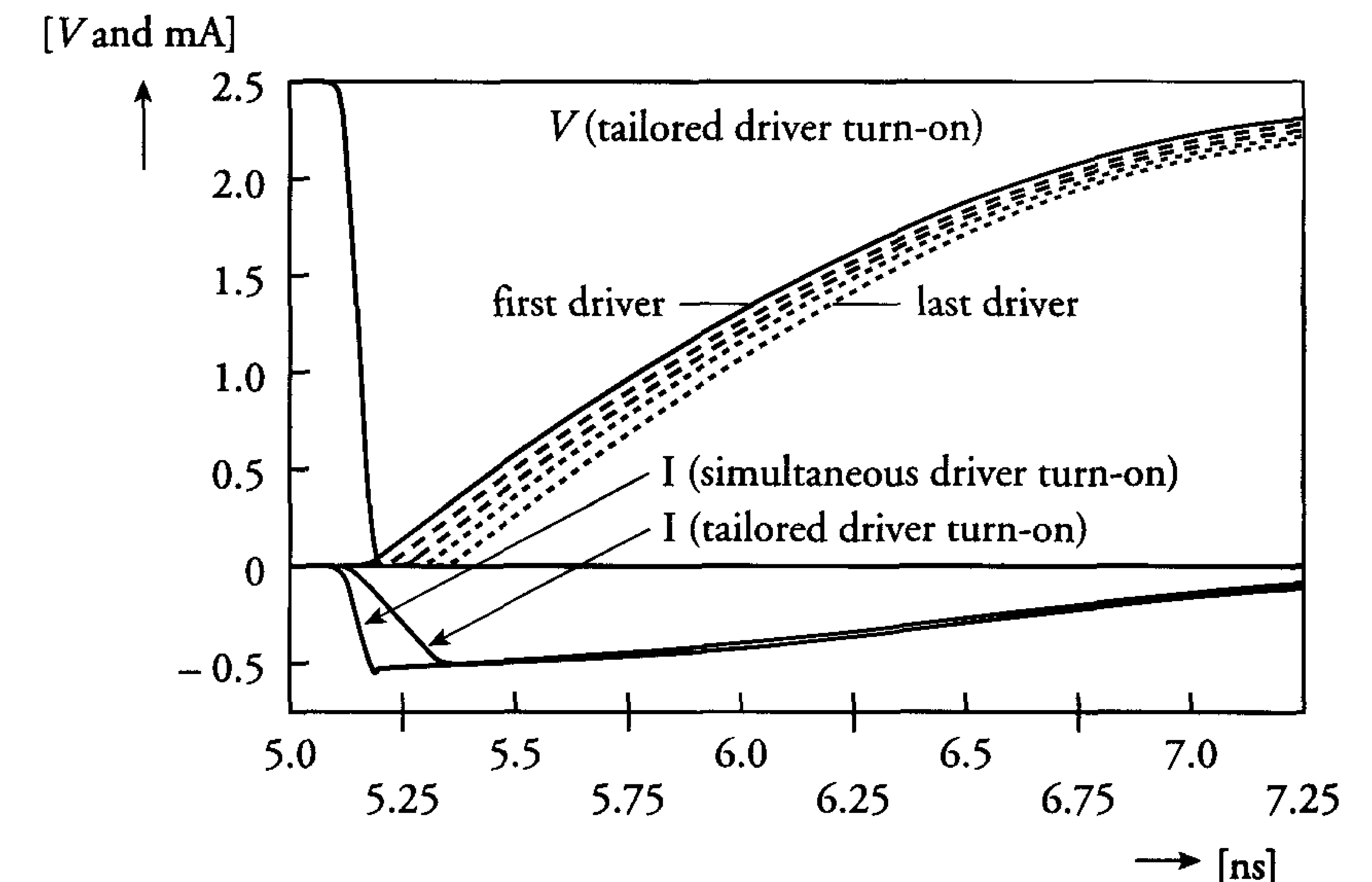


Figure 9.23: *Current and voltages in simultaneous and tailored driver switch-on networks*

Figure 9.23 also includes the current of a simultaneous driver switch-on network and of the just-described tailored driver switch-on. The result is a reduction in dI/dt of a factor of 3.5, in favour of the tailored driver switch-on network. In this example, the driver switch-on idea was used in the IC's output section. However, it can also be used when large internal buses need to be driven at high speed.

Another design solution uses output buffers, in which the dI/dt factor is limited (voltage slew rate controlled output buffers). Assume the output stage of a driver, shown in Figure 9.24, has certain W/L ratios for the p- and n-type MOS transistors.

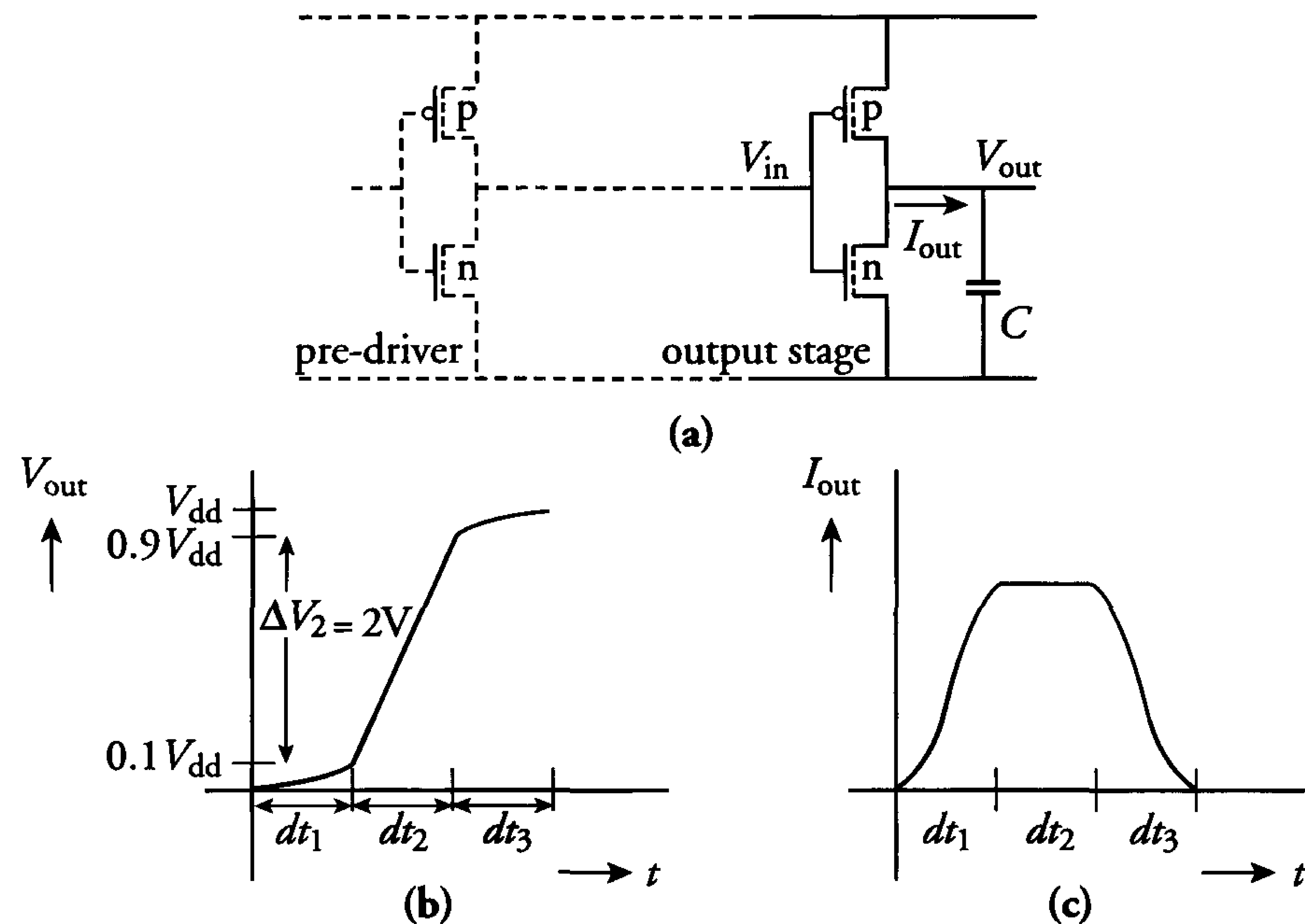


Figure 9.24: *(a) Basic output stage of an (output) driver and (b) output voltage and (c) output current*

Assume an input (V_{in}) low and output (V_{out}) high level. When the input switches to the high level, the current in the nMOS transistor increases to a maximum level. The faster the input switches, the faster the maximum level is reached and the higher the dI/dt will be. Therefore, the rising slope of the current is determined by the predriver (library designer's responsibility). As soon as there is a current, the load capacitance (C) starts to discharge. When this load capacitance is very high, the output will discharge slowly. As a result, the current will also reduce slowly to zero (small dI/dt).

Analogous to this, a small load capacitance will cause a high dI/dt. Because a designer cannot predict the different loads his chip can meet in different applications (he only specifies the maximum load), he has to limit the maximum dI/dt. For this reason, the designer must only use voltage slew rate controlled output drivers, in which the output rise and fall times are controlled via an electronic feedback loop inside the driver.

In combined analogue/digital VLSI circuits, the major part of the circuit is often of a digital nature and only a fraction is analogue. These

analogue circuits often restrict themselves to A/D and D/A converters, clock generators and clock synchronisation circuits (oscillators and PLLs), band gap references and memories (which operate as analogue circuits). Because of their analogue nature, these circuits are very susceptible to noise, while the digital circuits generate most of the (supply and substrate) noise. It is therefore advisable to isolate these analogue circuits both physically and electrically from the digital ones.

Mixed VLSI chips may contain digital circuits, as well as memories and analogue circuits. In many memories, only a few circuits are active at the same time. Such memories can be considered as large on-chip decoupling capacitances. They may contribute to reducing supply bounce, if their supply network resistance is very low, to allow a fast charge transport from the 'decoupling memory' to the location that needs the charge. Therefore, it is preferable to place such a memory close to an edge of the chip and to locate the analogue circuits behind the memories, close to the corners of an IC.

When these memories are also connected to the separated analogue supply and ground rails, the substrate bounce (originated by the digital circuit) is also locally reduced in the analogue areas close to these memories.

In summary, a robust mixed analogue/digital design is achieved when the analogue supply (V_{dd_A}) is connected to the same supply rails as the memories (only those large memories that generate very little noise) and physically located "behind" these memories, furthest away from the digital circuits. The analogue supply must be disconnected from both the digital (V_{dd_D}) and the output supply lines (V_{dd_Q}). However, the analogue circuits have the same ground as the digital part (V_{ss}), but physically disconnected from the output ground (V_{ss_Q}).

Figure 9.25 shows a block diagram with a recommended supply and ground pinning. Normally, the number of supply and ground pads on an IC will be equal. When the package has several additional pins available, they should be used as ground pins, to reduce the impedance of the ground with respect to that of the supply. This will reduce the substrate noise in mixed ICs.

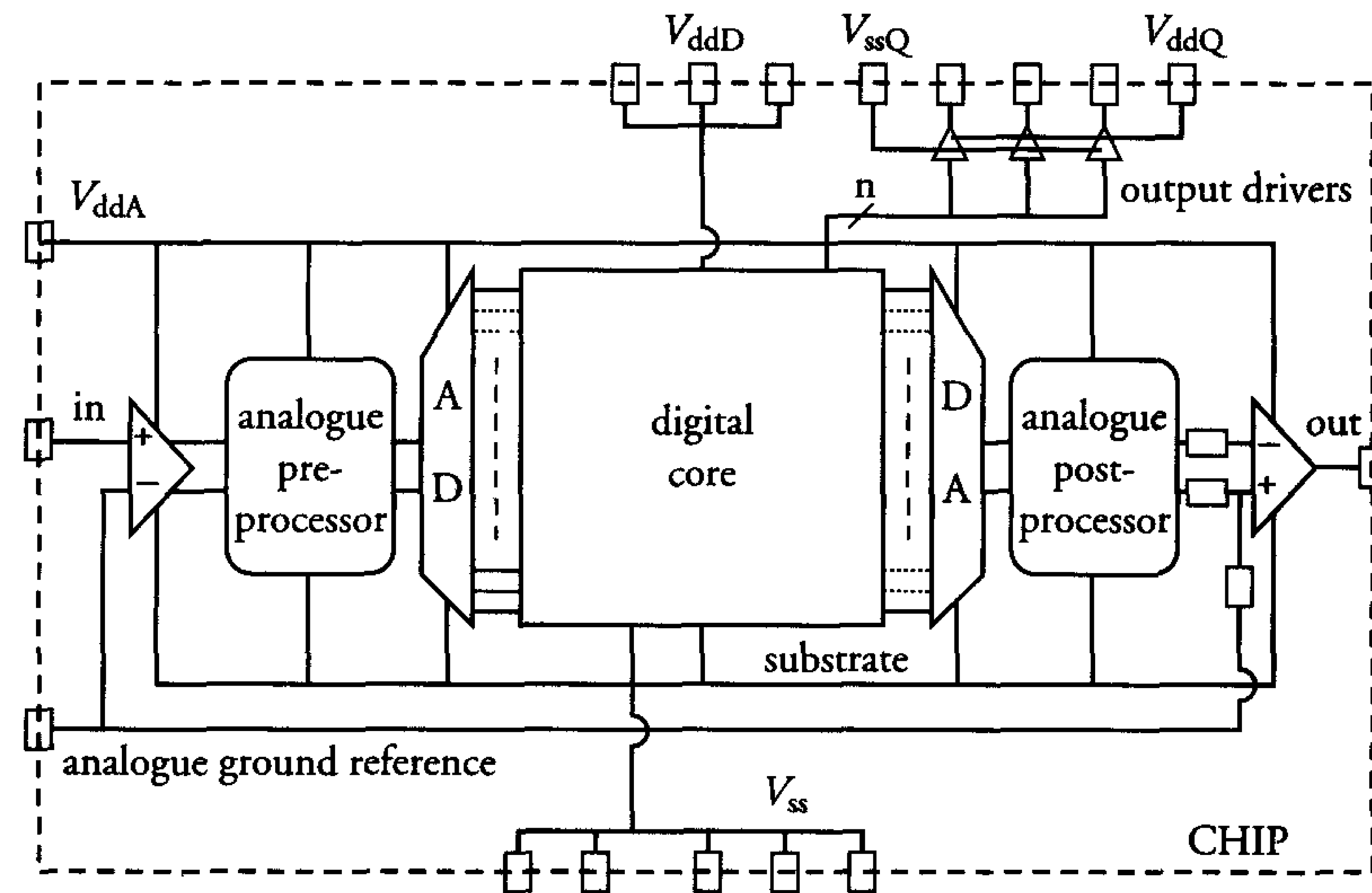


Figure 9.25: *Recommended supply and ground pinning in mixed analogue/digital ICs*

Whatever IC has to be designed, the above effects have to be considered and modelled such that they are included in the simulations, to increase the robustness of the design.

Electromagnetic compatibility (EMC)

The problem of supply and ground bounce caused by large current changes is not only restricted to on-chip circuits. High current peaks may also introduce large electromagnetic disturbances on a printed circuit board (PCB) as a result of the electromotive force, as defined in formula (9.2). Because bond pads, package and board wiring act as antennae, they can "generate" or receive an electromagnetic pulse (EMP), which can dramatically effect neighbouring electronic circuits and systems [5].

When realising electromagnetic compatible (EMC) circuits and systems, the potential occurrence of EMPs must be prevented. The use of only one or a few pins for supply and ground of complex high-performance ICs is one source of EMC problems. Even the location of these pins is very important with respect to the total value of the self-inductance. The use of three neighbouring pins for V_{dd}, for instance, results in an electromagnetic noise pulse that is twice as large as when the supply pins are equally divided over the package.

The best solution is to distribute the power and ground pins equally over the package in a sequence such as V_{dd}, V_{ss} and V_{dd}, etc., in the same way that twisted pairs are used in cables. As already discussed, outputs contain relatively large drivers with high current capabilities. Actually, each output requires a low-inductance current return path. Consequently, the best location for each output is right between a couple of V_{ss} and V_{dd} pads. This results in the smallest electromagnetic disturbances at PCB level and reduces the supply noise at chip level. However, as this is not very realistic in many designs, more outputs will be placed between one couple of supply pads. The limitation of this number is a designers' responsibility (simulation!).

Besides this, measures that reduce supply and ground bounce also improve the electromagnetic compatibility of the chip and result in a more robust and reliable operation.

Reducing supply bounce and improving EMC behaviour by using on-chip decoupling capacitances

Several methods can be used to lower the noise (dI/dt) caused by fast current fluctuations. Some of them (such as tailored driver switch-on and multiple supply and ground pins) have already been discussed in the previous sections. To reduce the self-inductance L of the bonding wire, it is also important that these supply and ground pads are positioned in the middle of the chip edges. At these positions, both the length of the package leads and of the bonding wires are often minimal.

The use of board decoupling capacitors that are placed very close to the chip reduces the power level fluctuations as a result of on-board self-inductances of the board wiring. Such decoupling capacitors are charged during the steady state. During current fluctuations, they temporarily serve as a power supply. However, on-chip current transients caused by the self-inductance of the bonding wires, package leads and on-chip supply lines cannot be reduced by putting decoupling capacitors next to the IC package. Because ICs have become larger and their complexity has grown to millions of transistors, many of the board and package level problems have been transferred to the chip level.

Over generations of ICs, the bus widths, the number of outputs and the clock load, etc. have increased dramatically. For example, microprocessor bus widths have increased from four to eight bits in the late seventies to 64 and even 128 bits today. Consequently, the total capacitance that has to be charged or discharged simultaneously has increased

proportionally. As a result of the higher resistance of the power supply net (larger chip sizes) and the increased current slew rates (dI/dt), it is no longer possible for the peak currents to be fully supplied by the power supply lines without showing large voltage bounces on these lines. During such current peaks, the necessary charge is collected locally, when available.

EXAMPLE

As an example, we take a complex video signal processor. Assume that this processor has 64 outputs. Each output is able to drive 30 pF at 2.5 ns rise and fall times. Figure 9.26 shows the circuit.

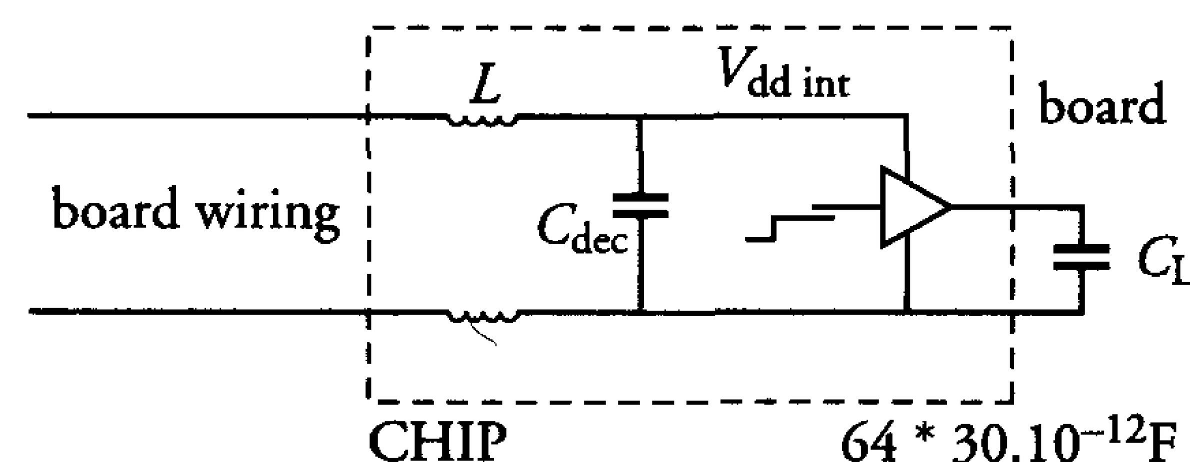


Figure 9.26: *Example of on-chip decoupling capacitance for the compensation of output current fluctuations*

Let us assume a maximum voltage dip equal to $\Delta V_{dd} = 125\,\text{mV}$. Capacitor C_{dec} is now calculated as:

$$\frac{C_L}{C_L + C_{dec}} = \frac{\Delta V_{dd}}{V_{dd}} = 1/20$$

$$C_{dec} = 19 \cdot C_L$$

In this example, it is assumed that no current is supplied by the board supply. In reality, the current is temporarily supplied by both the board supply and the on-chip decoupling capacitance. The required decoupling capacitance value may therefore reduce. A commonly-used practical value is:

$$C_{dec} = 10 \cdot C_L$$

In the example, $C_L \approx 2\,\text{nF}$ and thus:

$$C_{dec} \approx 20\,\text{nF}$$

With a gate capacitance of about $7\,\text{fF}/\mu\text{m}^2$ (oxide thickness: 5 nm), this decoupling capacitance requires an area of $3\,\text{mm}^2$, if the complete area could be used for the gate oxide.

Because of the number of contacts required to limit the channel resistance, a maximum capacitance value of about 4 to $5\,\text{fF}/\mu\text{m}^2$ can be reached (for the implementation of this decoupling capacitance, see below). The decoupling capacitance of the example will therefore occupy an area of about $5\,\text{mm}^2$. This is a relatively large area if the total chip size is less than $50\,\text{mm}^2$. However, complex ICs with that many output pins (64 in the previous example) have chip sizes of $100\,\text{mm}^2$ or more, for which this is less of a problem.

Even the switching activity inside a logic block can be so high that extra decoupling capacitance must be added to limit the supply noise. This activity depends on the kind of function the block represents.

For audio and telecommunication ICs, the activity factor (percentage of the nodes inside a logic block that switch in the same clock period) is often between 5 to 20%. However, for video ICs, this percentage is higher and can sometimes be as high as 40%. Especially this kind of logic block (with an activity factor of more than around 20%) requires additional decoupling capacitance to limit the internal supply bounce to less than 10 % of the supply voltage. This additional decoupling capacitance must be distributed all over the logic block.

When the utilisation factor (i.e. the percentage of the total logic block area that is really filled with library cells) is not close to 100%, the empty areas can be used to place a decoupling cell. Because these empty areas will have different sizes, different decoupling cells have to be available to fill up these areas. These decoupling cells can be routed in first metal only. For a five-metal layer design, they are almost completely transparent for the routing of the logic block cells and, as such, they do not increase the block area. Additional area is only required in those logic blocks that need more decoupling capacitance than can be achieved by filling the empty areas. The placement of these cells can be done by the place and route tools in a standard design flow.

Not only the logic cells may need additional decoupling. The activity inside flip-flops is even higher than that of basic logic gates (NANDs, NORs, etc.) because they have additional clock activity. A logic block may sometimes include a relatively large number of flip-flops (sometimes flip-flops occupy 50% of the total logic block area). In such a case, there is hardly any logic left that can act as decoupling capacitance. Some

libraries therefore offer two versions of flip-flop cells, one with additional decoupling capacitance inside and one without.

For the realisation of the decoupling cells, there are a few alternatives. As the gate oxide has the least thickness of any dielectric used in a MOS process, it is obvious that the gate capacitance has the highest capacitance value. However, for ESD and technology reasons, it is not allowed to connect a gate oxide capacitor right between V_{dd} (supply) and V_{ss} (ground).

A very good alternative is the cell in figure 9.27(a):

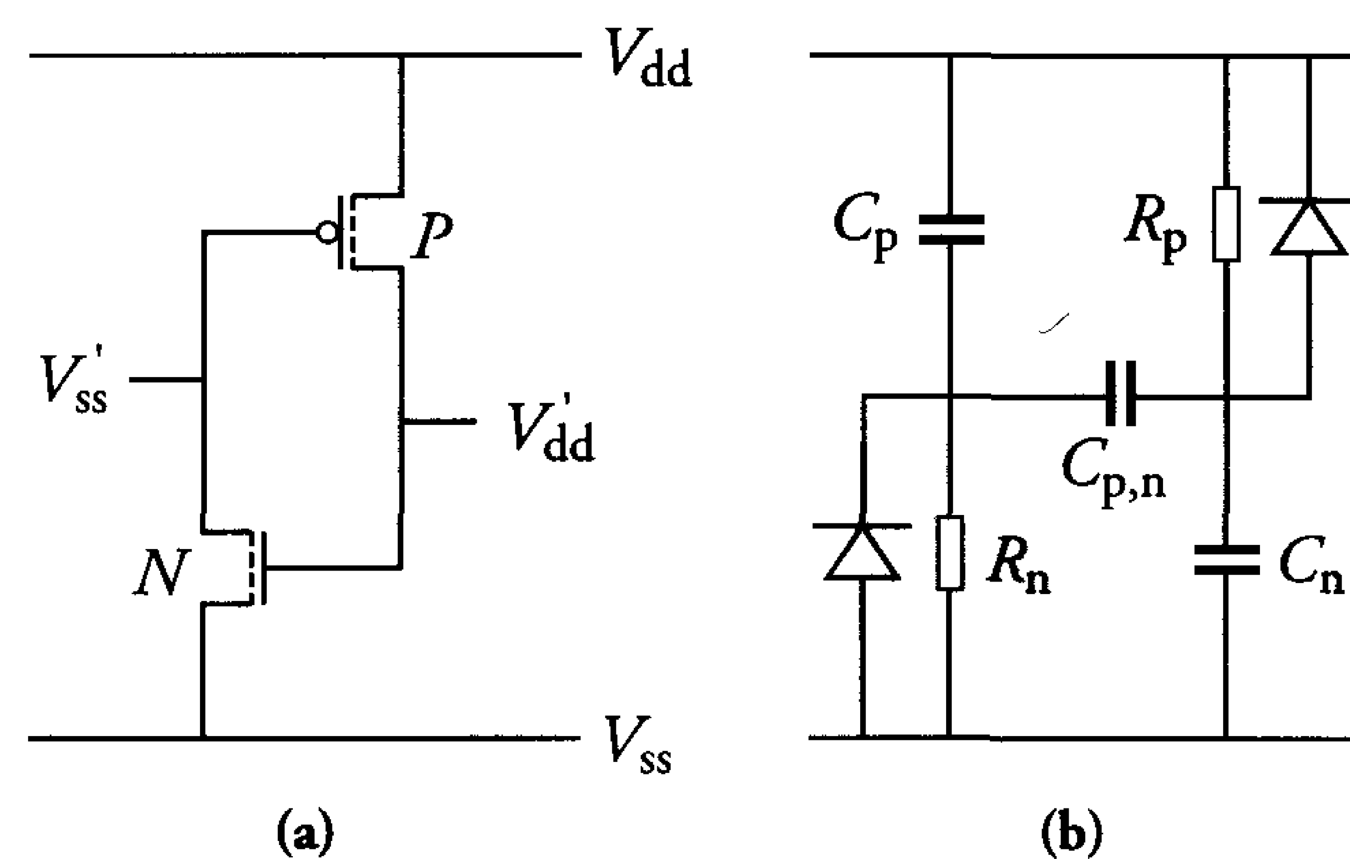


Figure 9.27: *(a) The tie-off cell used as decoupling cell and (b) its equivalent circuit*

In some cases, it is necessary to connect dummy inputs of logic gates or other cells (e.g. dummy cells in memories) to nodes which mimic V_{dd} and V_{ss}. Figure 9.27 shows an example of a cell which generates V_{dd} and ground levels; such a cell is called a tie-off cell. When the power supply is not yet switched on, all nodes are at ground level. When the power supply is switched on, the V_{ss}' node stays low (as a result of parasitic capacitances), thereby switching the pMOS transistor on. The V_{dd}' node is then be charged to V_{dd}, causing the nMOS to be switched on as well. This nMOS in turn keeps the V_{ss}' level at ground. Because of the positive feedback, the resulting state is stable. Because both transistors are on, their gate capacitors are charged and can be used for decoupling. However, the design of this cell for use as a decoupling cell requires the special attention of the designer. Figure 9.27(b) shows the equivalent circuit. The gate capacitor C_n of the nMOS transistor is

charged through the channel resistor R_p of the pMOS transistor, while C_p is charged through R_n.

It is very important to keep the RC times of this cell low. When there is a power supply dip, the charge (which is stored on the gate capacitors) must be supplied in only a very small fraction of the power dip duration. This means that both RC products (R_pC_n and R_nC_p) must be less than 200 ps. The cell must be designed by specialists and then included in the library. When implemented as a library cell and if designed correctly, this tie-off cell can offer a capacitance value of about 2 to 3 fF/μm^2 in a 0.25 μm CMOS technology.

An important issue here is that the (active) gate oxide area on the chip can dramatically increase or even double when large capacitance values are required for decoupling. This is particularly true in very high-performance ($f > 500$ MHz and/or $P > 20$ W) or mixed analogue/digital ICs.

As the gate oxide thickness is very small in a MOS transistor (currently 2 to 7 nm), there is a strong requirement on the quality of the oxide because it is responsible for the transistor performance (β and V_T) and for its reliability (oxide imperfections, pinholes, etc.), see also chapter 3. A substantial increase of the gate oxide area to implement on-chip decoupling capacitances is generally not advantageous, either in terms of yield or reliability. The designer has to take care that the amount of additional on-chip decoupling capacitance is limited to what is really required. Therefore, not all the empty areas should be filled, if this is not required from a signal integrity point of view.

In conclusion, although on-chip decoupling capacitors consume relatively large silicon areas, they are of great help in improving the signal integrity with respect to supply bounce and EMC. Figure 9.28 shows a high-performance video signal processor containing 15 nF of on-chip decoupling capacitances. The lighter areas represent these capacitances. There are many examples of ICs containing on-chip decoupling [7]. The ALPHA^{TM} processor chip of DEC contains 160 nF on-chip decoupling capacitance [8].

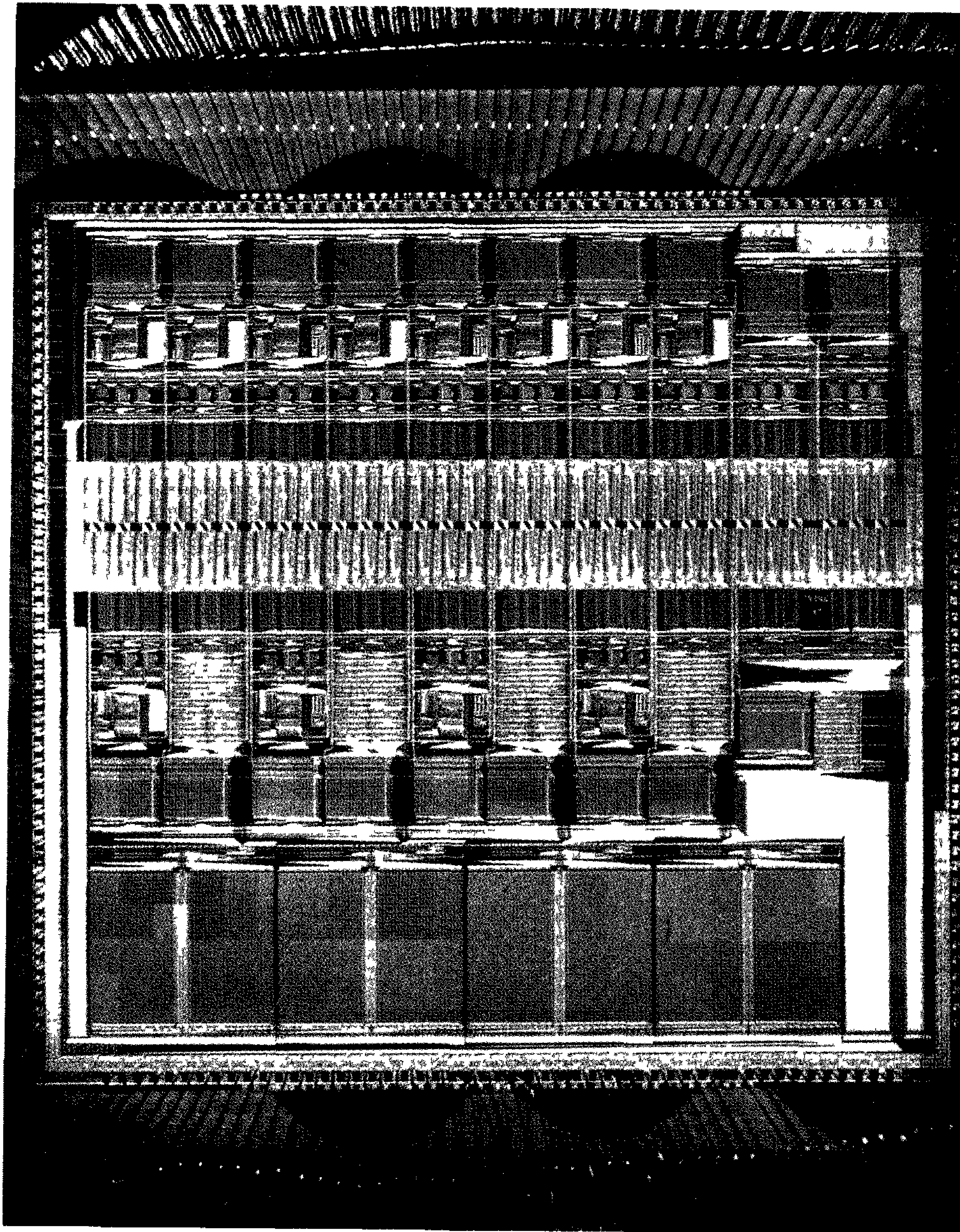


Figure 9.28: *A high-performance video signal processor containing decoupling capacitances (photo: PHILIPS)*

9.3.5 The influence of the interconnection (metallisation and dielectrics)

Although design rules shrink with each new generation of technology, the average chip area still shows some increase. As a result, logic block sizes hardly change or even tend to increase accordingly. Also, the interconnection lengths between logic gates remain at about the same order of magnitude whilst the interconnection widths and spacings decrease by a factor of about 0.7 each process generation.

This, combined with the increase of current density (and power), may lead to severe performance, reliability and signal integrity degradation. Increased line resistances lead to larger voltage drops whilst increased mutual line capacitances lead to greater cross-talk. The *RC* time of a signal along a line increases dramatically and results in intolerable line delays.

The following subsections discuss the dramatic effects of voltage drop, cross-talk and line delay if they are neglected during the design phase.

Line resistance (and related voltage drop)

During the past two decades, the first metal layer thickness has decreased from about 1 μm thickness to about 0.5 μm. Also, the width of the metal has dramatically reduced over the years. The result has been an ever-increasing resistance, which has now reached a value of about 70 m$\Omega/\Box$, which means that a 1 mm long aluminium track at a width of 0.4 μm (about 2500 $\Box$'s) represents a resistance of around 175 Ω.

The use of drivers with high current capability placed at certain distances from the supply and ground pads requires special attention with respect to the width of their supply lines. Figure 9.29 shows an example:

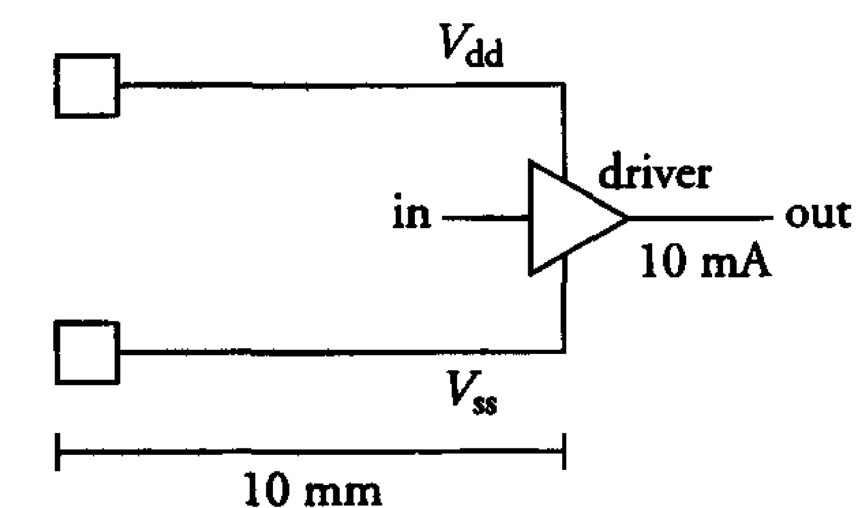


Figure 9.29: *Voltage drop on supply and ground lines during output transients*

Suppose a 10 mA driver is located at 10 mm from the nearest bond pads. If we allow a maximum voltage drop of only 200 mV along the supply lines, then the width (W) of these supply lines can be calculated from:

$$\begin{aligned} R &= \frac{\Delta V}{\Delta I} = \frac{0.2}{0.01} = 20\Omega = \frac{10 \cdot 10^{-3} \cdot 70\,\mathrm{m}\Omega/\Box}{W} \\ \Rightarrow W &= 35\,\mu\mathrm{m} \end{aligned}$$

From this example, it can be concluded that large voltage drops across the supply network of an IC can dramatically affect the speed of the circuit or can even lead to erroneous operation. Before generating the layout, the overall chip supply network must be given special attention to make sure that both average and peak currents will not lead to severe performance and/or reliability problems. Finally, line resistance will be decreased by using different metals for interconnection. In this respect, copper can give a resistance reduction of about 30 to 40%. Copper is expected to be used in processes with feature sizes equal to and beyond 0.18 μm.

Line capacitance (and related signal interference; cross-talk)

Cross-talk is an effect whereby one signal propagates partly through another one by means of capacitive coupling. On-chip inductive cross-talk, other than previously discussed, can still be neglected. There are several forms of cross-talk. Particularly in dynamic CMOS circuits, some nodes may float (no conducting path to either supply or ground) during part of the clock period. These floating (high-impedance) nodes are particularly sensitive to interference (cross-talk) from neighbouring signal tracks. Cross-talk in dynamic circuits has already been discussed in section 4.4.4. This section focuses on cross-talk between parallel signal lines in general.

As chip sizes become larger and minimum feature sizes shrink with each new process generation, the wavelengths of the signals will reach the same order of magnitude as the interconnections. When the signal rise or fall time on a 1 cm metal track goes below about 150 ps, transmission line phenomena such as reflections and characteristic impedance become increasingly important [5].
Because this is not yet the case for almost all ICs, we will restrict ourselves to noise (cross-talk) generated by the capacitive coupling between the signal lines. The combined effect of larger chip areas and higher complexity causes both an increase of the current densities and of the length

of signal wires. This prevents interconnection layer thicknesses from being scaled with the same ratio as minimum feature sizes each process generation. The side wall capacitance (and thus the coupling between two signal lines at minimum spacing) increases rapidly over the process generations. Moreover, the increase of the number of interconnection layers makes the cross-talk problem much more complex.

Figure 9.30 shows a comparison of a conventional 1 μm CMOS single metal process and the lower three metal levels of a 0.25 μm six-metal layer CMOS process. This figure clearly demonstrates the pace of scaling of the interconnections within a time frame of less than 10 years.

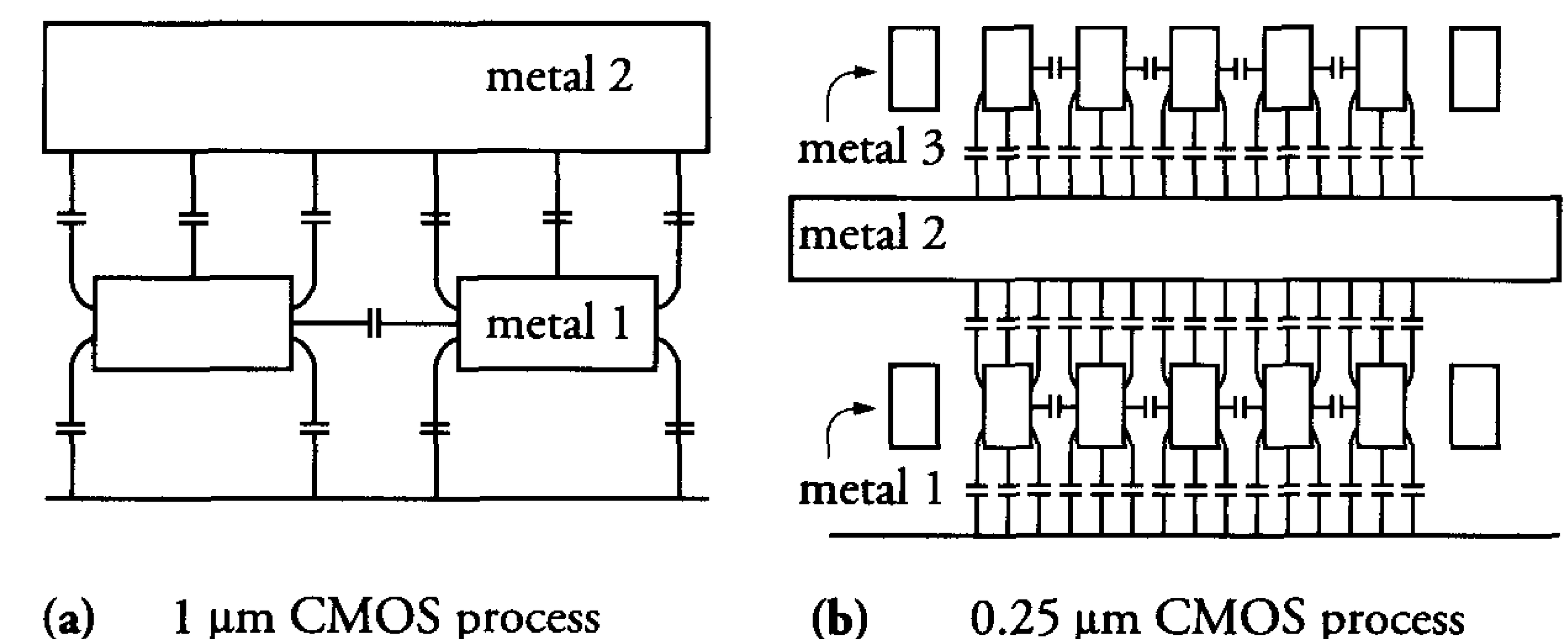


Figure 9.30: *(a) Capacitances in a single level metal 2 μm CMOS process and (b) a six-level metal 0.25 μm CMOS process (only the lower three layers are shown)*

Figure 9.31(a) contains a very simple, but representative model, which shows the cross-section of two metal tracks in the same layer. Capacitance C_{m} represents the mutual capacitance between the two tracks, while capacitance C_{ground} represents the total capacitance of track M2 to ground. The equation in the left upper part of the figure expresses the level of cross-talk.

Figure 9.31(b) shows the second metal capacitance values for different technology generations. It is clear that deep-submicron designs will exhibit severe cross-talk problems if its prevention is not part of technology development, design style and design flow.

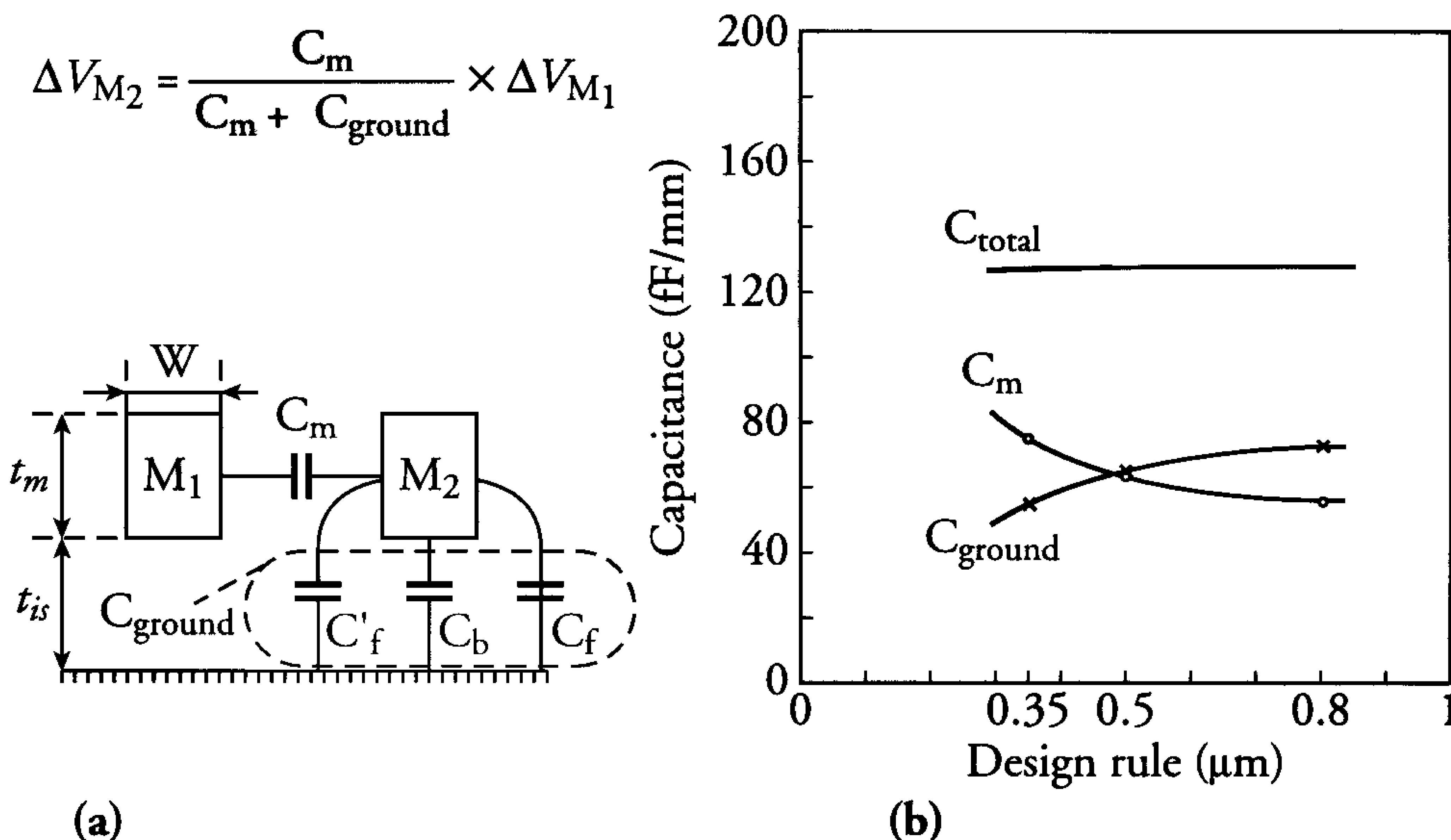


Figure 9.31: *(a) Simple capacitance model of two metal wires in the same layer (2nd metal) and (b) capacitance values for 2nd metal in different technologies*

The following example shows the severeness of the cross-talk phenomenon in a 0.25 μm process. Representative values for these capacitances per mm length are: $C_m = 80\,\text{fF/mm}$ and $C_{ground} = 40\,\text{fF/mm}$. If track M2 is floating, then a voltage swing ΔV_1 on track M1 would generate a noise pulse ΔV_2 on M2 equal to:

$$\Delta V_2 = \frac{80}{120} \cdot \Delta V_1$$

With a nominal supply voltage of 2.5 V and a maximum voltage swing ΔV_1 of 2.5 V on M1, the noise pulse on M2 (the victim wire) will be 1.6 V. If this victim wire were to have two simultaneously switching neighbours, the noise level would almost be as high as 2 V.

Because this is unacceptable, we can conclude that tri-state buses in (signal) processor ICs or precharged bit lines (in memories) are extremely susceptible to cross-talk. Attention should be paid to "floating lines", precharged lines and tri-state buses, etc., not only with respect to their logic levels (which will be corrupted by the high level of cross-talk) but also with respect to reliability.

If, for instance, one line is (pre)charged high and, after some time, its neighbour also switches high, then the first line may reach a voltage

level (in a 0.25 μm CMOS process) of 2.5 + 1.6 (cross-talk) = 4.1 V. If this voltage comes across a channel of a single transistor, hot-carrier effects will result in a change of the threshold voltage and in degradation of transistor reliability. Figure 9.32 shows an example:

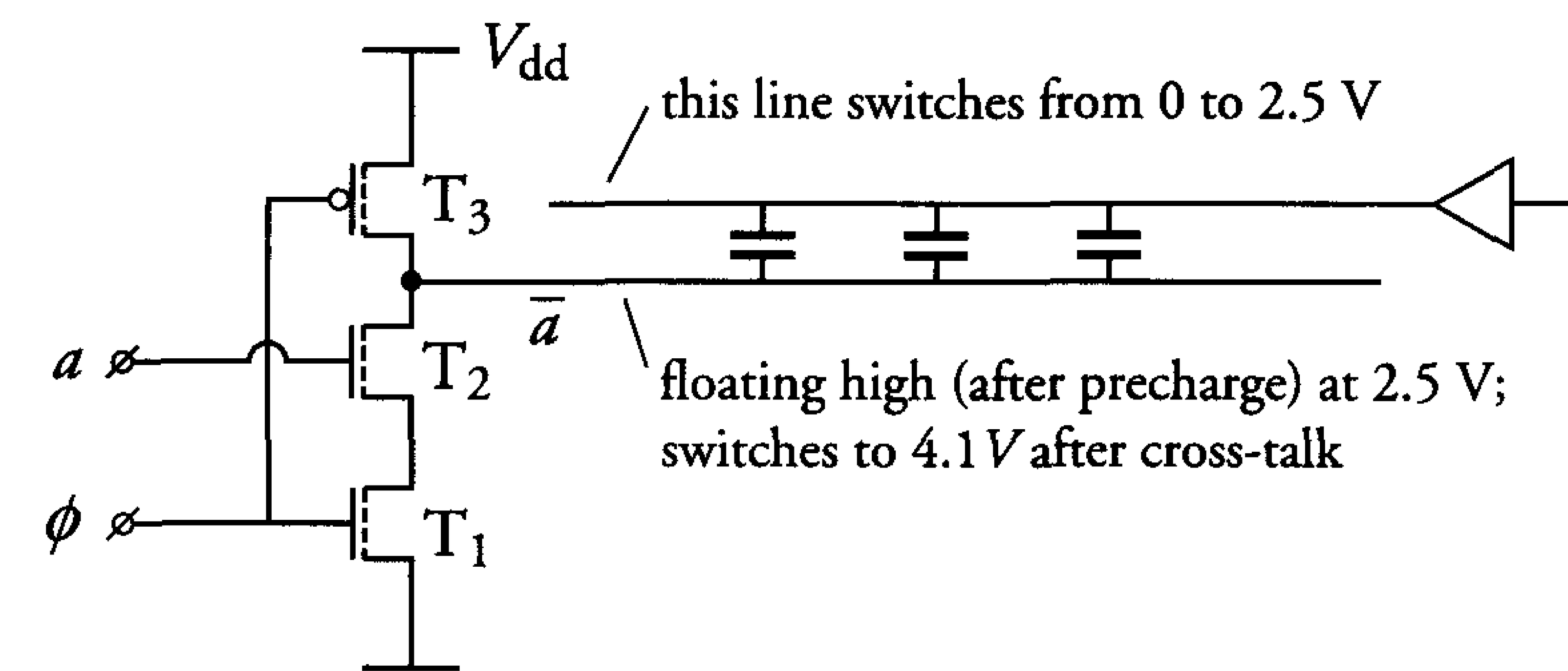


Figure 9.32: *Dynamic circuit with temporary floating output*

If input a = low, T_2 will be off and a voltage of 4.1 V may exist for a short time across the channel of transistor T_2 during the sample moment (ϕ = high). This voltage will be reduced rapidly by the pMOS transistor and by the p^+n forward biased junction of the pMOS transistor.

In many circuits, however, signal lines will never be left floating. Also, in these cases, cross-talk might corrupt signal levels if long tracks are involved.

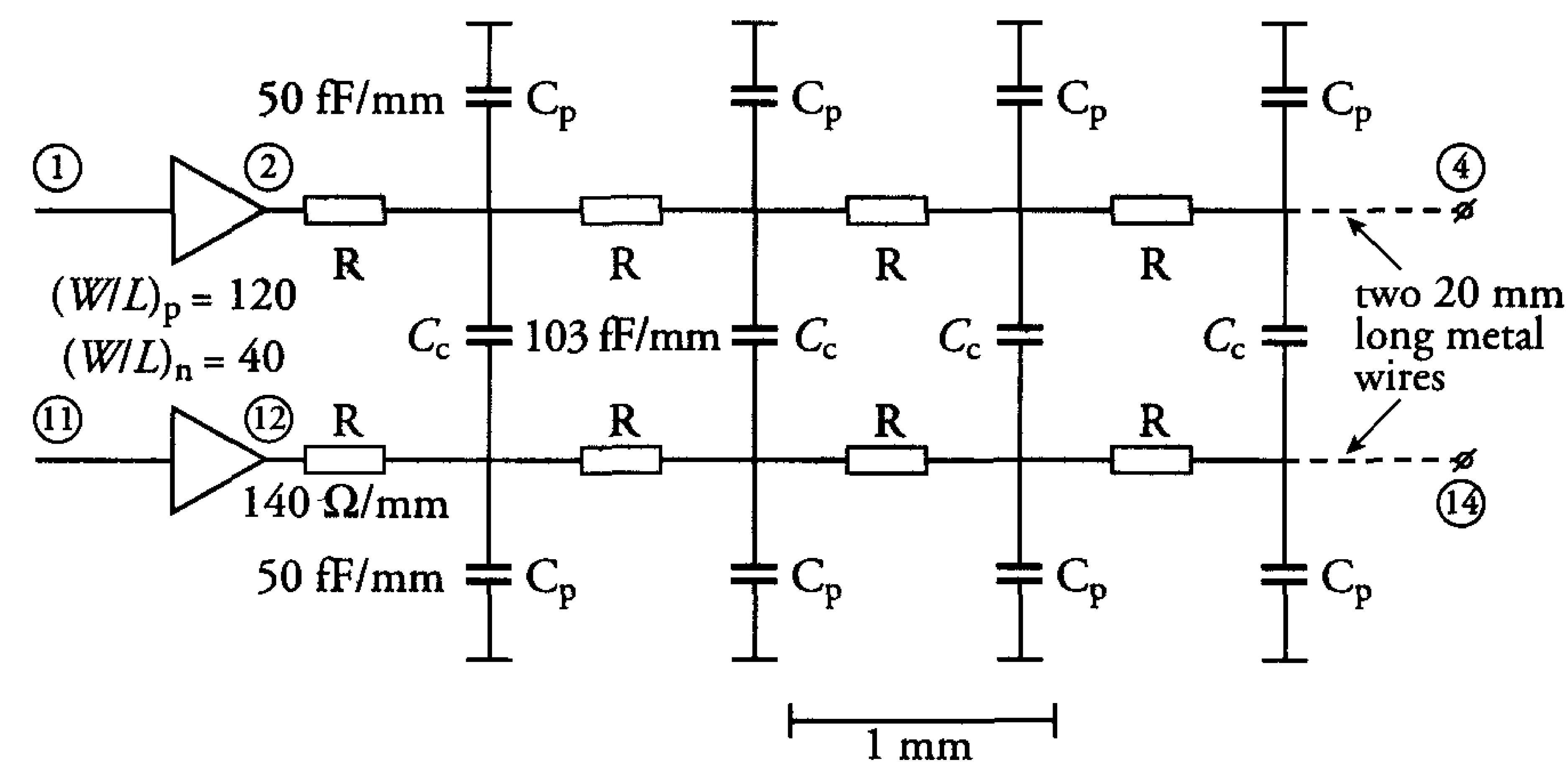


Figure 9.33: *Model for cross-talk simulation between two parallel wires*

Figure 9.33 shows a model for the simulation of the cross-talk between two 20 mm long second-metal wires at a minimum spacing in a 0.25 μm CMOS process. These wires are modelled by 20 stages, each consisting of two resistors R, one coupling capacitor C_c and two track capacitors C_p. The results of a circuit simulation of the above model, in which 1 mm wire section is represented by one resistor-capacitor stage, is shown in figure 9.34.

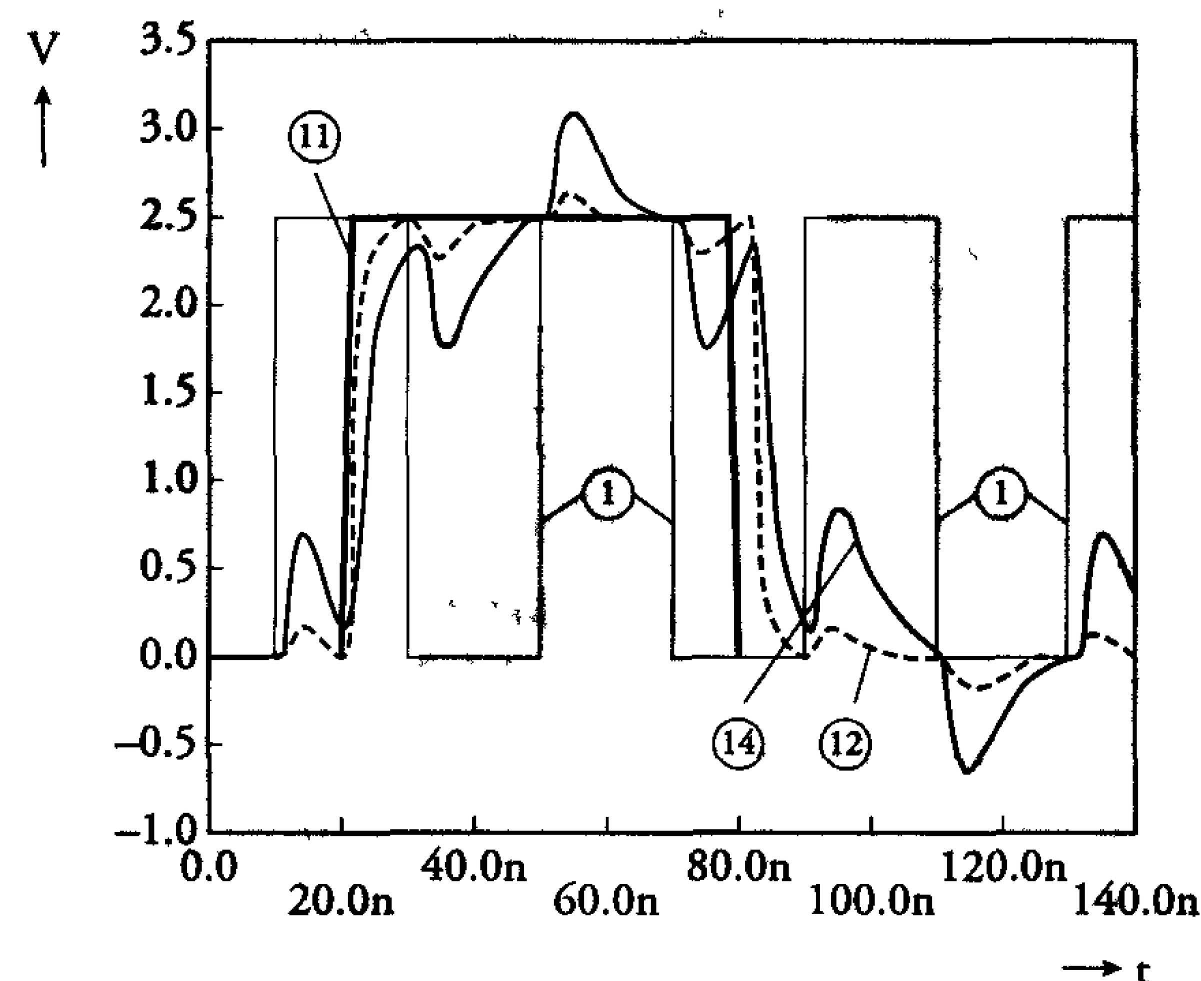


Figure 9.34: *Mutual cross-talk between two long metal lines in a 0.25 μm process*

The figure shows cross-talk noise pulses with amplitudes at wire ends of about 0.65 to 0.85 V. These values amply exceed the threshold voltage levels and the reader can imagine the effects when everything is even somewhat worse or when two neighbouring lines switch simultaneously. In this case, the in-between line may show cross-talk noise at the end of the wire that exceeds 1.3 V!

As signal wire lengths on VLSI chips of 10 to 20 mm are no exception, the cross-talk may have dramatic effects on the proper behaviour of the circuit.

In conclusion, long signal lines behave almost like tri-state buses at wire ends and are therefore very susceptible to cross-talk. Large designs in deep-submicron technologies therefore require tools to manage the cross-talk problem. Some tools have a cross-talk feature: they need characteristic technology input. This allows them to detect nodes inside logic blocks that show too much cross-talk. Another feature in this tool

is that it can re-route the critical nodes, either by placing the tracks further apart, or by placing them in different layers, to reduce the mutual capacitance [8]. Beyond 0.25 μm feature sizes, metal heights, oxide thicknesses and dielectric constants are expected to scale according to the SIA roadmap, see table 9.2.

Table 9.2: SIA interconnection roadmap: line capacitance scaling

Technology	250nm	180nm	150nm	130nm	100nm	70nm
Metal height (nm)	450	324	300	273	240	189
Minimum metal spacing (nm)	340	240	210	170	140	100
Distance to active (nm)	0.792	0.572	0.504	0.45	0.378	0.29
Effective dielectric constant	3.0-4.1	2.5-3.0	2.0-2.5	1.5-2.0	1.5-2.0	<1.5

If technology developments keep pace with this roadmap, the cross-talk will almost remain constant, as shown in figure 9.35. The figure shows the cross-talk for a uniform, low-ϵ dielectric and for layered dielectrics. It also shows the reduction in cross-talk if the minimum spacing increases by a factor of five, for example.

$$\Delta V_{victim} = \frac{2 \cdot C_{lateral}}{C_{vertical} + 2 \cdot C_{lateral}} \cdot \Delta V_{source}$$

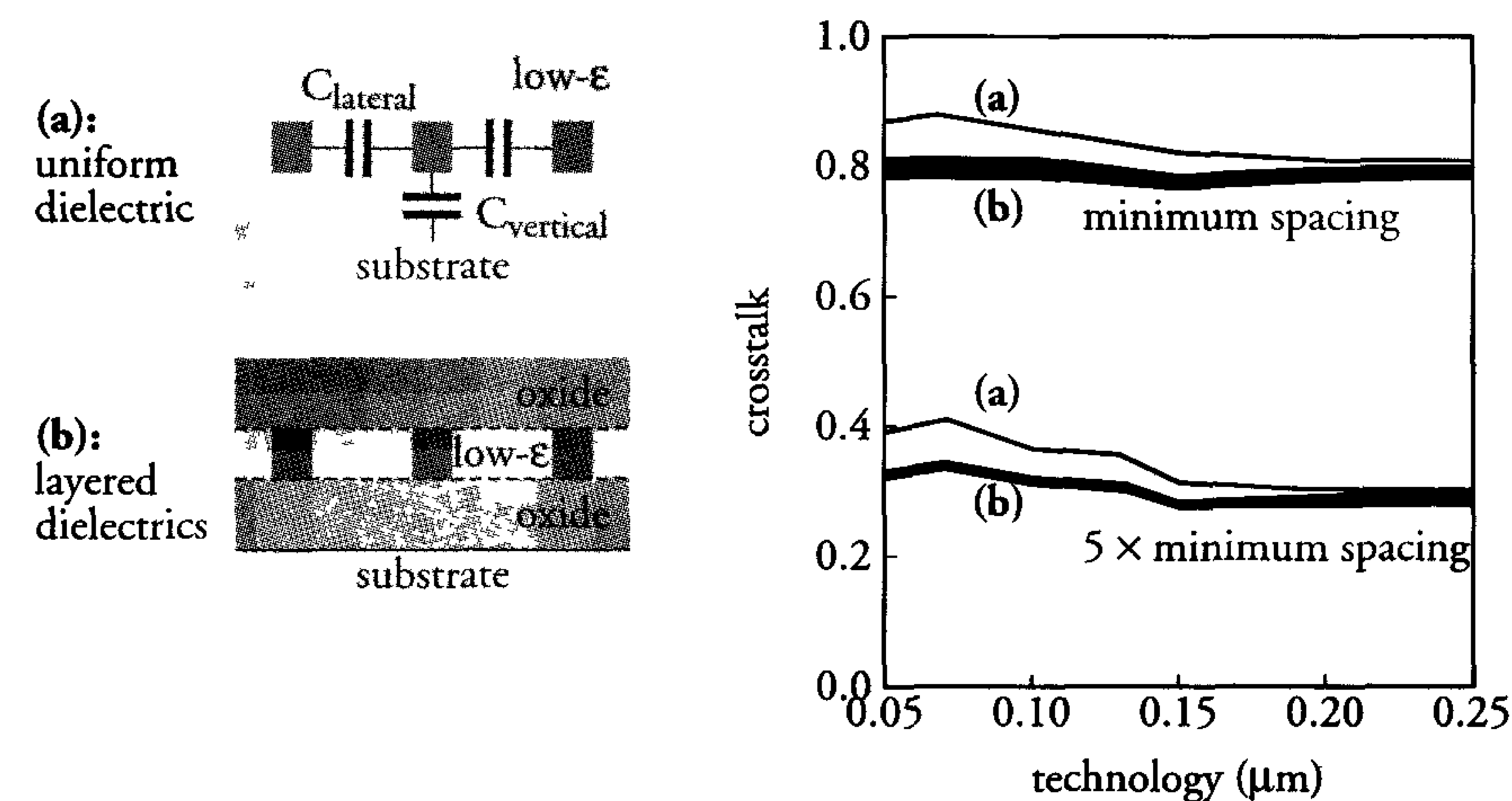


Figure 9.35: *Cross-talk trends according to capacitance values predicted in SIA roadmap*

Line delay (RC delay)

Line delay is thought to be a problem in conservative one-metal layer processes. In these processes, the crossing of data lines and clock lines was often implemented via a polysilicon bridge with a resistance of $R_{ps} \approx 40\,\Omega/\square$. Whenever the data lines were long, this could lead to relatively large delays (tens of nanoseconds) and a dramatic reduction of the performance or the malfunctioning of the chip.

Current processes offer several metal layers (Al, Cu) which have resistances of 30 to 70 m$\Omega/\square$, depending on the layer thickness. In these processes, the RC delay is not a problem for most of the signals, because they are routed over average distances and have an average load (number of inputs to which they are connected). However, there are signal lines such as clock lines, scan control lines and data buses which may run all over the chip to provide global control or communication. As discussed, such wire lengths may exceed some tens of millimetres or even hundreds in the case of clock signals.

Particularly in the case of *buses*, signal lines are completely embedded. Next to the fact that their behaviour interferes with each other, leading to cross-talk as discussed in the previous subsection, they also affect each others *signal propagation delay*. Figure 9.36(a) shows a victim line, which is embedded within two agressors. The design has been made in a 0.18 μm technology, with minimum line widths and minimum spacings between the victim line and its agressors.

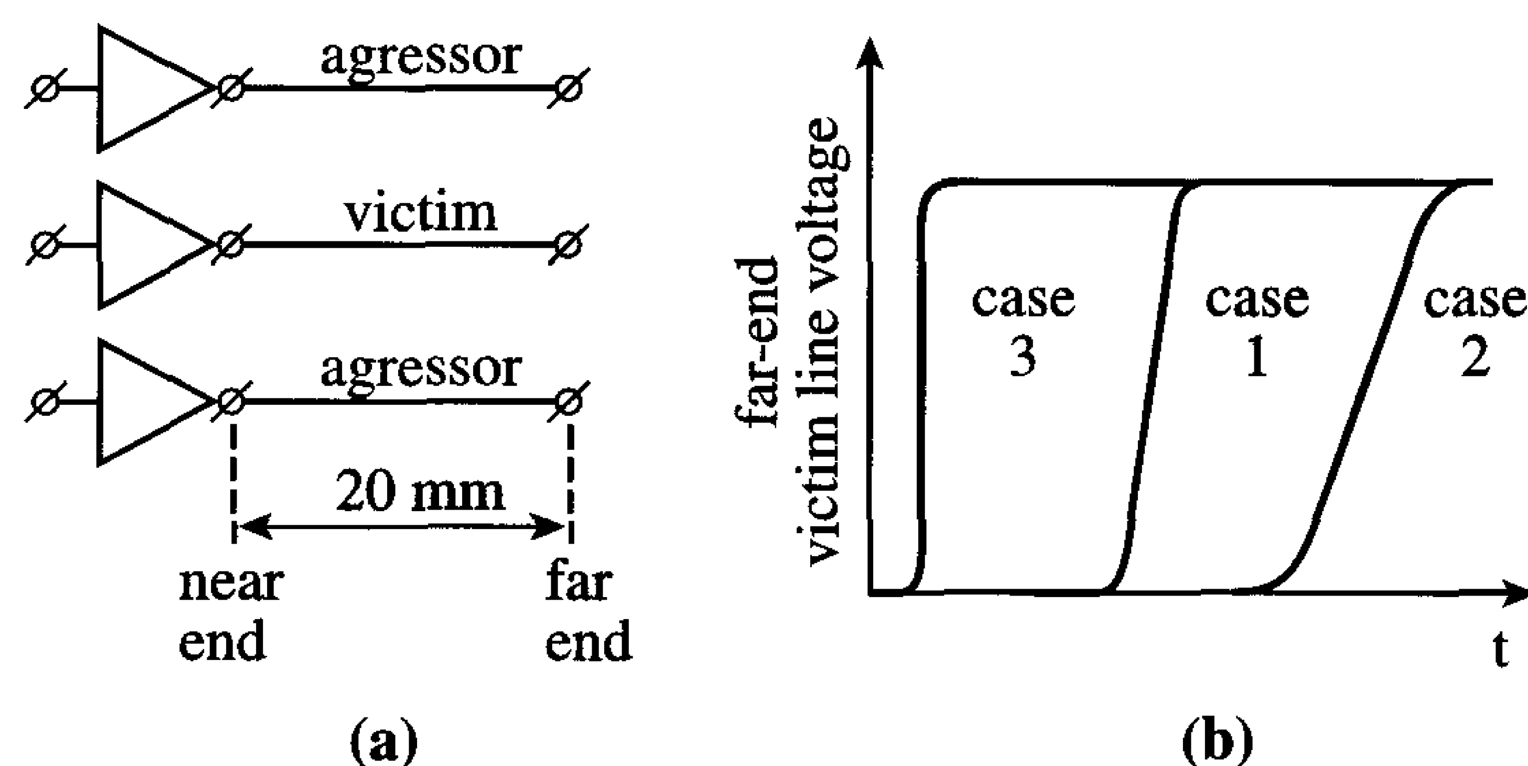


Figure 9.36: *(a) Model for signal propagation in buses (b) Far-end victim line signal for different cases*

In the following we distinguish three cases of operation:

case 1: victim switches from low to high while agressors remain low
In this case, the victim line shows a capacitance to ground and to its two neighbours, as modeled in Figure 9.33 by capacitors C_p and C_c, respectively. The far-end victim line signal is represented by case1 in Figure 9.36(b).
case 2: victim switches from low to high while agressors switch from high to low
Now the victim line shows about the same bottom capacitance but it looks like the mutual capacitance to both of its neighbours has doubled. Due to the minimum spacing between the signal lines, these mutual capacitances are much larger than the bottom capacitance. Case2 in figure 9.36(b) represents the far-end victim line signal, which shows about twice the propagation delay compared to case1.
case3: both victim and agressors switch from low to high
Since the neighbours switch in the same direction, the victim line almost only shows a bottom capacitance and therefore the total propagation delay to its far end has dramatically reduced with respect to both previous cases.

The difference in signal propagation between the worst case (opposite switching) and best case (same switching) is about a factor of ten! Since this effect is not yet included in all the design tools, it can lead to severe timing problems and it is the designers responsibility to include accurate wire models and signal propagation in the overall design simulation and verification phase. Further information on future signal propagation trends is also presented in chapter 11.

With the use of other metals for interconnection (copper) and the use of low-ϵ dielectrics from 0.18 μm technologies onwards, the purely interconnect RC delay is expected to increase according to the ITRS roadmap, as shown in figure 9.37. In summary, it is advisable that signal lines which exceed a certain critical length and load are automatically detected at the end of the design cycle. The designer is then able to check whether the corresponding RC delay is critical.

Due to the increasing interconnect propagation delay, design styles have to be changed as discussed in section 11.3.

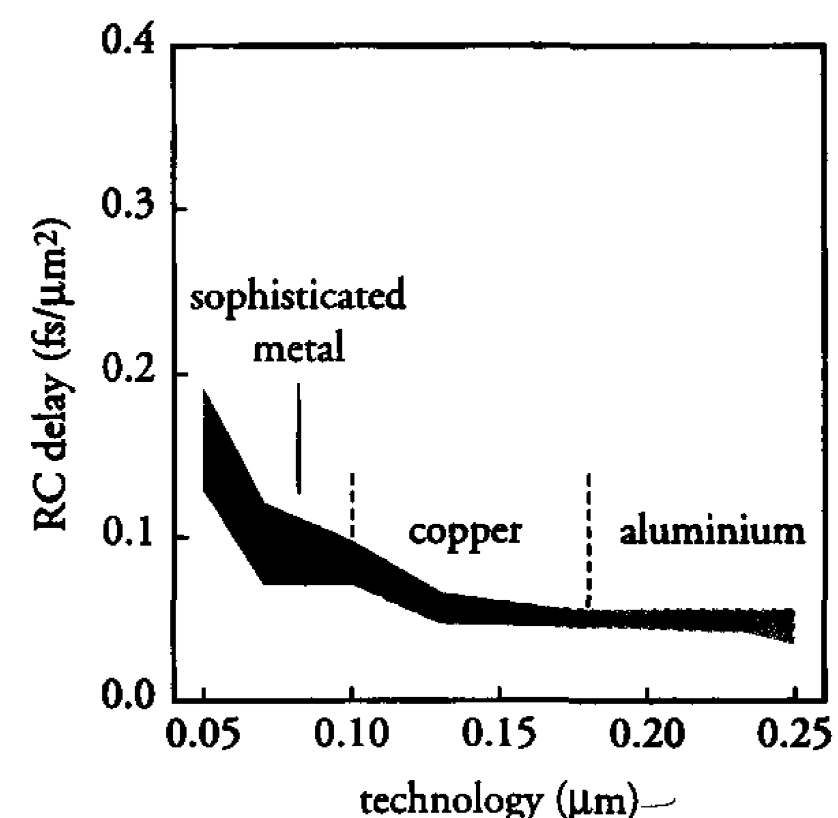


Figure 9.37: *RC delay scaling of interconnection layers according to the ITRS roadmap*

Transistor matching influence on digital circuits

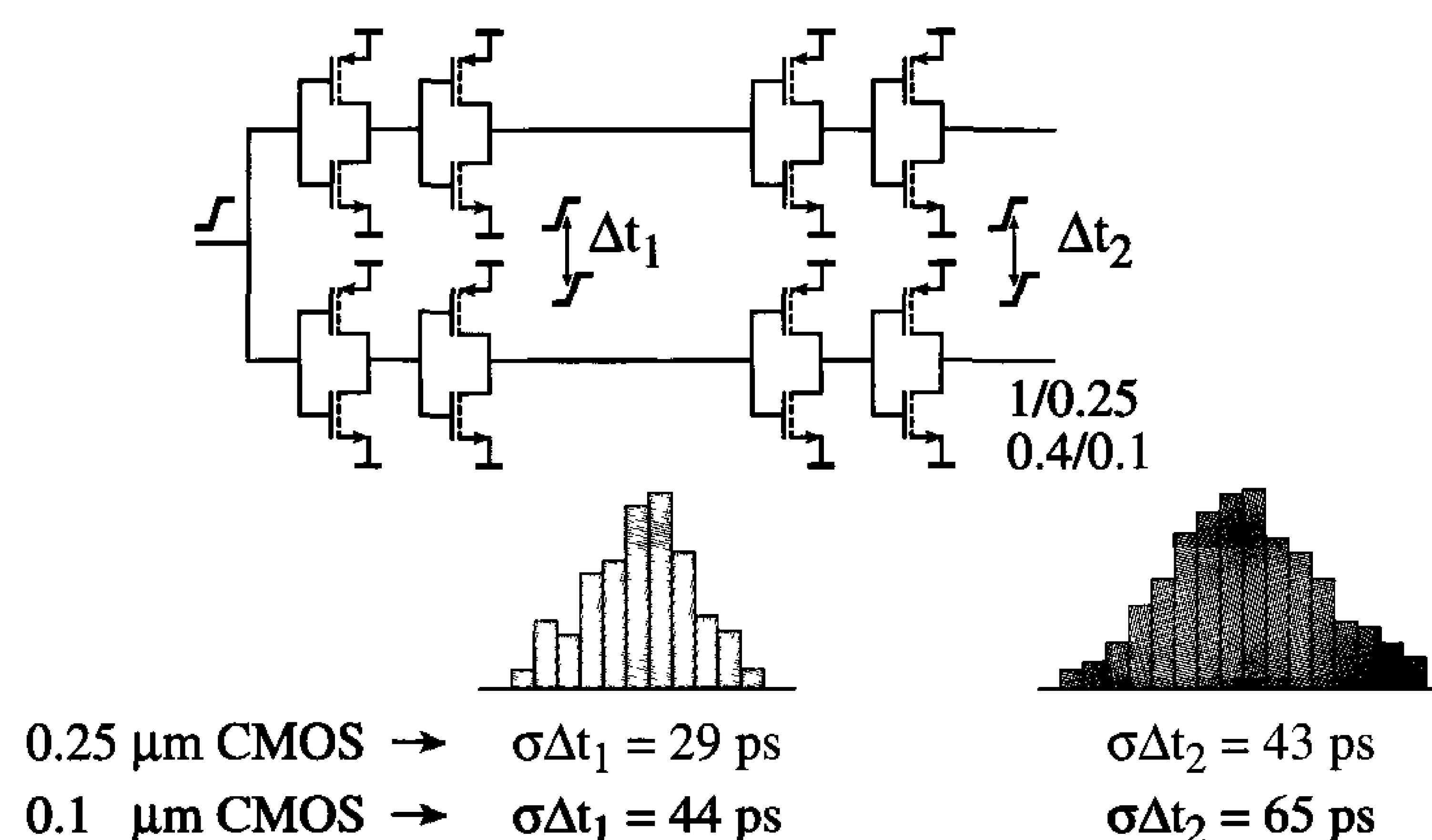


Figure 9.38: *Spread in signal arrival times due to transistor mismatch*

Matching of transistors means the extend to which two identical transistors, both in type, size and layout topology show equal device parameters, such as β and V_T. Particularly in analogue circuits (a memory is also an analogue circuit) where transistors are required to have a very high level of matching [13], the spread in V_T due to doping statistics in the channel of the MOS transistors results in inaccurate or even anomalous circuit behaviour. For minimum transistor sizes (area), this effect

increases every new IC process generation, such that both the scaling of the physical size and the operating voltage of analogue CMOS circuits lag one or two generations behind the digital CMOS circuits. Also for digital (logic) CMOS circuits, matching of transistors is becoming an important issue, resulting in different propagation delays of identical logic circuits. Figure 9.38 shows two identical inverter chains (e.g. in a clock tree), but due to the spread in V_T, they show different arrival times of the signals at their output nodes. For circuits in a 0.1 μm technology this time difference is in the order of several gate delays, depending on the depth of the logic chain. Particularly for high-speed circuits, for which timing is a critical issue, transistor matching and its modeling is of extreme importance to maintain design robustness at a sufficiently high level.

9.3.6 Design organisation

A very important issue for increasing the integrity of the database for design changes (be it in the final design stage or during a redesign) is the design organisation. There are two requirements on the database with respect to design changes:

- it should take minimum effort
- it should not introduce new errors caused by:
  - unintended modifications
  - forgetting something.

These requirements even hold for design changes after one or more years. Current VLSI chips often reuse existing blocks, such as multipliers, memories or microprocessor cores. This requires a structured and well documented database set-up and design documentation.

Ten to thirty percent of test engineers time is lost as a result of incomplete documentation of the design. What really is required is:

- Good and complete specification
- Complete (sufficient) test vectors
- Mixed signal ICs: the test engineer must understand the complete IC.

Generally, the best solution for a database set-up is a hierarchical approach, in which one has:

- Directory hierarchy = design hierarchy (easy to find your way through)
- Good version management; what object (netlist and layout, etc.) is generated from which source (VHDL description), and which one is the latest; use of make files!. In this respect, the use of a good database management system is necessary.

Another requirement, which particularly holds for custom designed (parts of) chips, is the use of (procedural) layouts with parameters (variables e.g. word size) to make a design more flexible:

- Modifications can now be performed by changing only one or more parameters (e.g. word size)
- Do not use copy/paste/edit to create a new circuit (cell) that is almost the same as the old one. The reason is that the original cell may be modified. It is then very difficult to find all derivatives and modify them individually (it is easy to forget one then).
- Use as few conditions as possible. Keep them simple:

```
if ((a ≤ b) or (c > word size) and (not (even (d))))
if (.... )
else .......
```

  and do not nest them too deeply:

```
if (.... )
  if (.... )
   if (.... )
   else
  else
else
```

Complex conditions make the code hard to understand and thus difficult to modify without introducing undesirable side effects. Instead of using complex conditions to describe irregular processes, one should use tables in which the variables (parameters) are clearly described and which are more readable.

A structured and logical database set-up reduces the chance of failures with (re)design.

9.4 Conclusions

During the last decade, the back-end of the production process (the interconnection and dielectric layers) has become a dominant factor in the performance, reliability and integrity of CMOS ICs. This chapter discusses measures that must be taken during the design to keep these parameters at a sufficiently high level. Technology measures, as presented in the SIA roadmap, should also contribute to lowering the dominance of the interconnections. Chapter 11 focuses on the scaling trends and includes both reliability and signal integrity aspects.

A robust design not only requires the integrity of the electrical and physical operation of the chip, it also includes the set-up of a very well organised database. This allows easy, correct and rapid design modifications when redesigns or different versions of the design are required.

9.5 References

[1] B. Barton, et al.,
ESSCIRC, low-power workshop 1997, Southampton

Latch-up in CMOS

[2] D.B. Estreich, R.W. Dutton,
'Modeling Latch-Up in CMOS ICs and Systems',
Proceedings of the 18th Design Automation Conference,
June 1981, pp 347-354

ESD

[3] W.D. Greason,
'Electrostatic Discharge in Electronics',
John Wiley & Sons Inc.,1993

EMC

[4] Jasper Goedbloed,
'Electromagnetic Compatibility',
Prentice Hall, 1992

[5] H.B. Bakoglu,
'Circuits, Interconnections, and Packaging for VLSI',
Addison-Wesley, 1990

[6] P.A. Chatterton and M.A. Houlden,
'EMC, Electromagnetic Theory and Practical Design',
John Wiley & Sons, 1996

Decoupling

[7] W.J. Bowhill, et al.,
'A 300 MHz 64b Quad-Issue CMOS RISC Microprocessor',
Digest: International Solid-State Circuits Conference, 1995, pp 182-183

[8] C.F. Webb, et al., 'A 400 MHz S/390 Microprocessor',
Digest: International Solid-State Circuits Conference, 1995, pp 168-169

Crosstalk

[9] M.R. Choudhury, et al.,
'A 300 MHz CMOS microprocessor with Multi-Media',
Digest: International Solid-State Circuits Conference, 1995, pp 170-171

Clocking and Timing

[10] Kerry Bernstein, et al., 'High-Speed CMOS Design Styles',
Kluwer Academic Publishers, 1999

[11] Stefan Rusu,
'Circuit Design Challenges for Integrated Systems',
Workshop on Integrated Systems, European Solid-State Circuits Conference, September, 1999

[12] H.A. Collins and R.E. Nikel,
'DDR-SDRAM high-speed, source-synchronous interfaces',
EDN, September 2, 1999

Transistor Matching

[13] Maarten Vertregt, 'Embedded Analog Technology',
IEDM short course on "System-On-a-Chip Technology, December 5, 1999

9.6 Exercises

1. Explain why the internal chip latch-up sensitivity will decrease every new process generation.
2. What are the main causes of supply noise inside a VLSI chip?
3. Explain why the power supply lines to a large driver circuit (e.g. clock driver or output driver) should be wider than the output signal track.
4. Why would a single-phase clock distribution network be preferred above a two-phase clock distribution network?
5. What is generally the best place to position the clock drivers and why?
6. What are the main causes of clock skew and what are the measures to reduce it?
7. Explain how the back-end of the manufacturing process is dominating the IC behaviour.
8. Mention several reasons for increasing dI/dt.
9. What is the impact of an increased dI/dt on the signal integrity?
10. Explain why the use of a good database management system is required during the design of a VLSI chip.

Chapter 10

Testing, debugging, yield and packaging

10.1 Introduction

Although this is almost the final chapter in this book, it does not mean that the topics discussed here are less important than those of the previous chapters.

Testing, debugging, yield and packaging have a substantial influence on the ultimate costs of a chip. Relatively short discussions of these topics are therefore included in this chapter.

An integrated circuit can fall victim to a large variety of failure mechanisms. Ideally, the related problems are detected early in the manufacturing process. However, some only show up during the final tests, or even worse, they might not be identified before the chip is soldered on a customer's board.

An overview of advanced failure analysis tools is presented in this chapter as well, to allow an early failure detection and a fast identification of the failure mechanism during first-silicon debugging, to prevent customer returns.

The engineering and evaluation of first silicon until it is considered to be "error free" happens to be a tough job. Programmable processors, for example, may be used in an almost unlimited number of different applications. It is almost impossible to guarantee even "fifth-time-right" silicon for these kinds of ICs.

Even when a failure is detected during the testing of first silicon, it might take a considerable time before the cause of failure is located and

proven. This is because complex ICs contain up to several millions of transistors and up to several hundreds of I/O pins. It is therefore very complex to locate an internal failure via a limited number of external pins (I/O). Moreover, because of the increased number of interconnection (metal) layers, physical probing of signals has almost become impossible. Design for debugging should therefore be adopted as a general design approach, to ease the detection of design bugs and other failure mechanisms during the engineering phase of first silicon.

Finally, this chapter is concluded by a presentation of failure analysis methods that support the detection of failures and their diagnosis.

10.2 Testing

Testing is done to bridge the gap between customer requirements and the quality of the design in combination with the manufacturing process. Testing thus helps to increase the quality of an IC. The yield is determined by testing and can be influenced by the complexity of the test. However, a simple test may lead to a higher yield but can lead to more *customer returns* ("escapes"). The yield for large and complex ICs can be relatively low and can dominate the ultimate costs.

Usually, pre-tests, also known as *e-sort* (early sort), are performed directly on a wafer, while the final tests are performed on the packaged die. There is often a lot of overlap between the pre-tests and the final tests. As a result of the associated additional costs, the number of redundant tests must be limited. During pre-test, the individual ICs are tested on the wafer by probing the bond pads of the chip. Figure 10.1 shows a photograph of an IC under test. The *probes* on the *probe card* used can be seen in the figure. A *probe station* brings these small needles into contact with the IC's bond pads. A *test computer* provides pre-determined *stimuli* for the IC and compares actual output signals to expected responses. The stimuli should ensure that a large percentage of possible faults will result in discrepancies. This percentage is called *fault coverage* with respect to the applied fault model and, unfortunately, is almost always less than 100%.

The fault coverage, however, is always related to the fault model used (stuck-at, bridging, stuck-open, transition (only once), gate delay and path delay). Redundancy can be another reason for reduced fault coverage. The test stimuli and response signals are transferred through a connector that provides a bi-directional link between the probes and the

test computer. The test computer, *Automatic Test Equipment* (ATE), can also be used to automatically step from one circuit to another so that a number of ICs can be tested in rapid sequence.

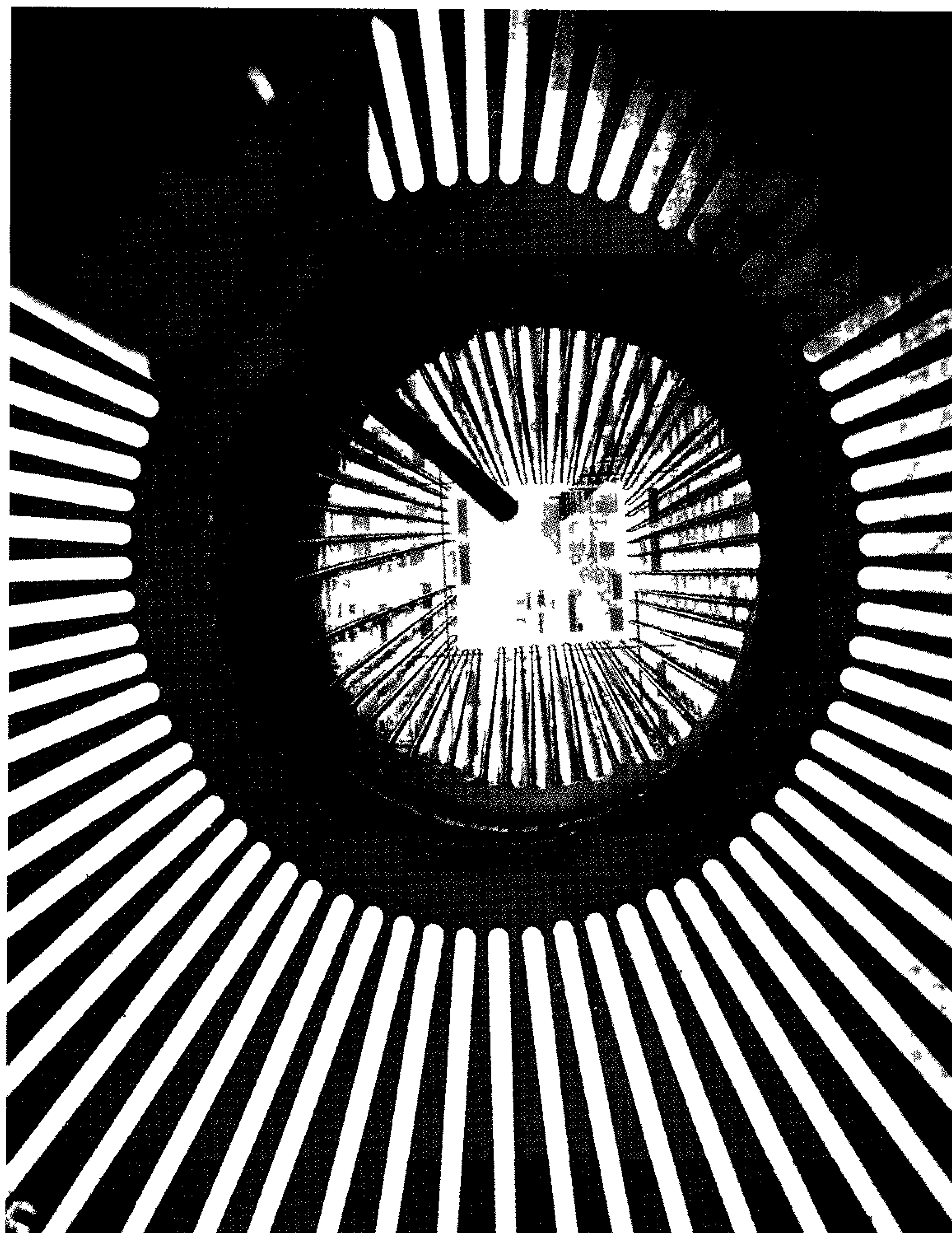


Figure 10.1: *Probe card with probes and a chip in a test environment (Photo: PHILIPS)*

The quest for high bit or gate densities consumes much design effort aimed at the realisation of a maximum amount of electronics on a minimum area. However, designers must ensure that their circuits are *testable*. For VLSI circuits, an increase in testability may, for instance, result in a chip area 'sacrifice' of 10% and a 50% reduction in test costs.

It was relatively easy to manually determine test stimuli (vectors) for complete SSI (small-scale integration) circuits. For VLSI circuits, however, this is impracticable and has led to the development of computer programs that generate test vectors.

Many test problems are related to dynamic effects such as *cross-talk*, *charge sharing*, critical timing and noise. Most automatically generated tests detect the *'stuck-at-one'* and *'stuck-at-zero'* faults, which cause circuit nodes to remain at '0' and '1', respectively. Deep-submicron technologies, however, will result in lower supply voltages and, consequently, reduced *noise margins*. This will produce faults that are much more difficult to classify than the traditional *stuck-at* faults.

Initially, the problems associated with large sets of test vectors were solved by applying faster test computers. However, the costs of generating the test vectors increased exponentially with chip complexity. It became evident that the generation of a complete test that affords 100 % fault coverage for even an LSI circuit posed a significant problem. Methods to approach this fundamental problem have therefore been sought. A detailed description of the various methods for improving the IC testability is beyond the scope of this book. However, several actions that designers can take are mentioned below.

- Make the design I_{ddq} testable.
  During the eighties, the testing of ICs was based on *stuck-at fault* models, which could detect failures at logic gates and flip-flops when their outputs were short circuited to V_{dd} or ground (stuck-at-one or stuck-at-zero, respectively). However, with these simple models, it was not possible to cover most process-oriented defects. With shrinking feature sizes, and thinner gate oxides, the most commonly occurring defects are gate oxide short circuits and bridging defects. Besides these defects, power supply short circuits and punch-through failures can also be detected by I_{ddq} *testing*.

  In a normal static CMOS logic gate, either the pMOS pull-up network is conducting, keeping the output at high level (logic "1"), or the nMOS pull-down network is conducting, keeping the output at low level (logic "0"). In the steady state, no current usually flows through such a logic gate, except for a negligibly small subthreshold leakage current in the switched-off network in the logic gate. In a normal situation, the magnitude of this leakage current is very low and should be in the order of 10 nA or less. At such a level of background current, larger steady-state currents, caused

by different process-defect mechanisms, can easily be detected by measuring, as these currents are several orders of magnitude higher than the leakage current.

For example, common gate oxide defects may result in current values in the order of micro-amperes to several milli-amperes, depending on the size of the defect and on the size of the transistor involved. A drain-source bridging defect can easily cause steady-state currents up to several milli-amperes as well.
During the measurement of the *steady-state current*, the chip has to be put in the steady-state mode. In many CMOS ICs, this state can be achieved by just switching off the clock. However, the chip often has to be put in a special I_{ddq} test mode before switching off the clock. In this way, defects are detected by the level of the supply current during the steady state. This is called I_{ddq} testing. I_{ddq} test pattern generation is only needed to put the chip or different parts of the chip in a certain mode (controllability). Observability need not be supported, as results of the test are measured via I_{ddq} currents. Because the current needs to settle during the measurement, I_{ddq} testing is a relatively slow process.

Especially in circuits that contain non-static CMOS circuits, such as PLLs, A/D, D/A and other analogue circuits, floating nodes (e.g. tri-state buses), dynamic and pseudo-nMOS circuits need additional attention during the design to make the total chip I_{ddq} testable.

At which I_{ddq} level the chip should be considered defect depends on many things. The number of gates is one important parameter, while the level of the threshold voltages of the nMOS and pMOS transistors is also dominant in determining the critical I_{ddq} level. Because of scaling, the threshold voltage is reduced every process generation, to maintain or increase the speed of each new generation of ICs. As shown in the previous chapter, the sub-threshold current in a transistor increases by about a factor twelve at each threshold voltage reduction of 100 mV. In the near future, I_{ddq} testability will therefore only be possible if the leakage current remains relatively low, or if the chip can be put in a low-leakage mode during I_{ddq} testing (e.g. modulation of the threshold voltage by applying a back-bias voltage during I_{ddq} test). It is claimed that I_{ddq} will not be feasible for large deep-submicron designs. More information on I_{ddq} testing can be found in [3].

- Subdivide the chip into separately testable functional blocks.
  The fault model and test pattern generation must be adapted to realistic defects. This means that the different functional blocks, such as random logic, SRAM, DRAM, flash, ROM and buses etc., need different and separate tests.

- Add self-test logic to suitable parts of the circuit.
  The costs of testing will dramatically increase as a result of the increase in the speed of the circuits, the reduction of the voltages (smaller noise margins) and the increase in the number of bond pads. The cost of a tester will increase from 1 million to 10 million US$ in less than a decade. *Built-in Self Test (BIST)* techniques are currently used in several (embedded) memories. To reduce the cost of overall chip testing, BIST techniques must also be included in the design of digital and analogue blocks.

- Improve testability by implementing additional test hardware.
  This is done by facilitating the serial connection of all the registers and single flip-flops in each logic functional block. Through such a *scan chain*, data can be scanned in serially to each set of input bits of each logic block. Then, that logic block can be put into the normal functional mode and the result of this operation can be put into the output flip-flops or registers. Next, the chip is again put into the scan mode and the result can be scanned out through the serial scan chain, see also chapter 9. Overall accessibility is ensured because each flip-flop in each logic block is included in such a scan chain.

- Include Boundary Scan Test for enhanced system testability.
  Advances in IC and packaging complexity lead to such densely integrated modules that overall system accessibility is reduced. Also, the need for faster time to market requires flexible and fast in-system testability. In 1990, a breakthrough in system test methods was made with the standardisation of the so-called *Boundary Scan Test (BST; IEEE 1149.1, JTAG)* method. BST reduces the overall test costs and simplifies board and system level testing.

  Although BST increases chip and board costs (additional area dedicated to *design-for-testability* circuits), this is recovered by the advantages mentioned in this section. BST also supports system production efficiency and in-field serviceability. With BST, interconnection failures during the assembly of ICs and in between ICs

on a board, such as the open circuits, short circuits and stuck-at faults, can be detected. In the BST approach, a boundary cell, which contains a flip-flop, is positioned between every pin to core connection. Each cell is also connected to its two neighbours, see figure 10.2. In the BST test mode, these cells form a scan register, which is able to serially scan in and scan out test data.

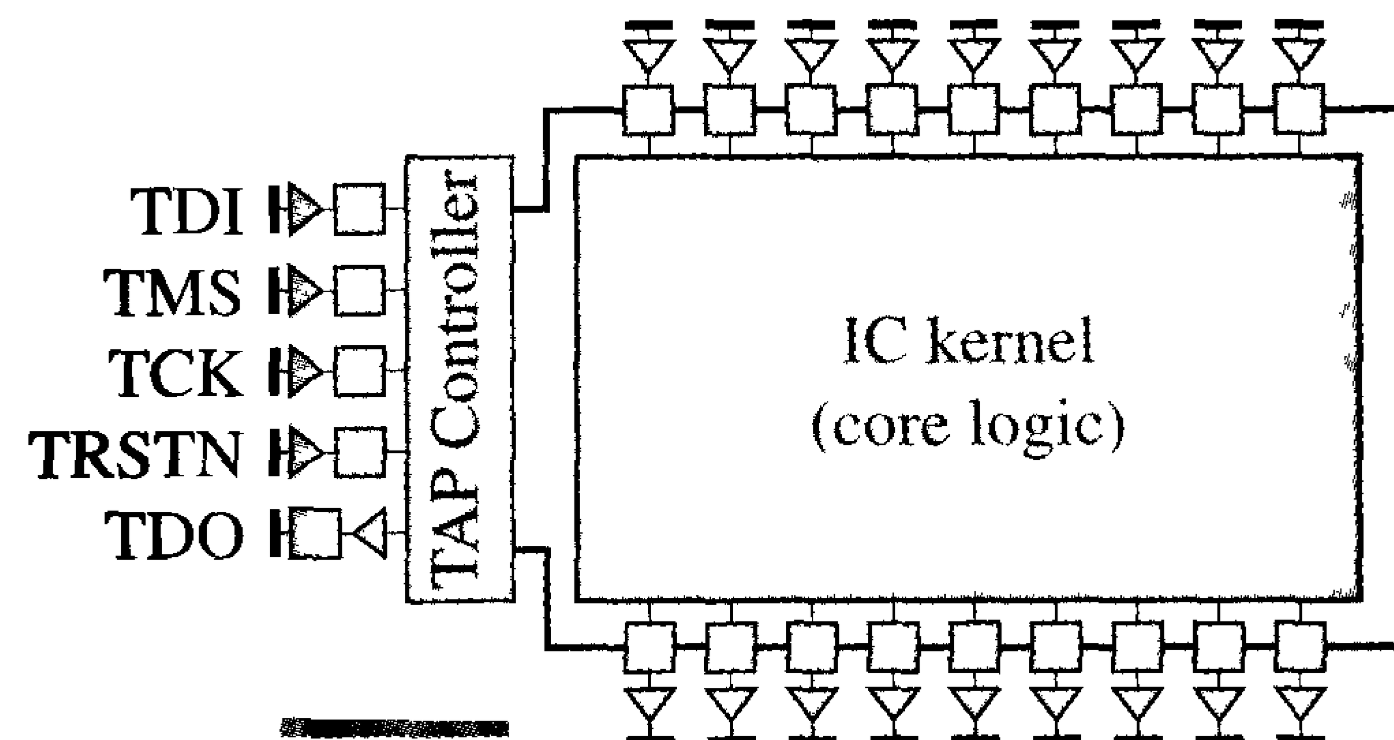


Figure 10.2: *The Boundary Scan Test approach*

Independently of the cores, such a scan chain can drive and monitor the pin connection of each chip in the system. A test clock and an additional test pin control the BST test mode of the system. BST supports three basic tests: interconnection tests between BST chips, IC core tests and function monitoring during normal circuit operation.

Because of the BST standard, ICs from different vendors supporting BST can be placed on the same board in a system to allow overall system testability. BST replaces the conventional 'bed of nails' test technique, which has become impractical. Ideally, all components on a board are equipped with BST. However, even if some components do not have BST, there are still substantial benefits. IEEE 1149.1 mandates a few instructions to support board level interconnection testing, but is open for private instructions. Many companies make dozens of such private instructions, e.g. for IC production testing, silicon debugging, emulation and application debugging, etc.

Methods for testability improvement are meant for computer testing of ICs. Prior to the computer test phase, however, design problems may appear during *IC characterisation* and *IC engineering*. On-chip *waveform measurements* are essential when timing errors, noise margin problems or other non-stuck-at errors are suspected. These measurements facilitate functional checking of different IC parts and local verification of timing specifications.

Mechanical probing, *e-beam testing* and *electro-optic sampling* are common techniques for on-chip waveform measurements. These techniques are discussed in section 10.7.

10.3 Yield

The current diameter of wafers used in modern IC production is mostly 8 to 12 inches. The size of an IC determines the number of dies per wafer. Most IC sizes range between 25 mm^2 and 200 mm^2 and their number per wafer therefore ranges from a few hundred to a few thousand. The ultimate price of an IC is determined by the number of *Functionally Good Dies per Wafer* (FGDW). This number is not only dependent on the number of dies per wafer but also on the *yield* of the process used. Quite a lot of dies on a wafer do not meet their specified requirements during testing. An additional number of dies is lost during packaging. The yield observed during wafer probing depends on the quality of the manufacturing and on the sensitivity of the design to process-induced defects. The production of deep-submicron ICs places very high demands on the factory building, the production environment and the chemicals. Disturbances in the production environment may be attributed to the following parameters:

- *Temperature*: Fluctuations in temperature may cause the projected image of the mask on the wafer to exceed the required tolerances.
- *Humidity*: High humidity results in a poor bond between the photoresist layer and wafer. This may result in underetching during the subsequent processing step (delamination).
- *Vibrations*: Vibrations that occur during a photolithographic step may lead to inaccurate pattern images on the wafer and result in open or short circuits.

- *Light*: The photolithographic process is sensitive to UV light. Light filters are therefore used to protect wafers during photolithographic steps. The photolithographic environment is often called the *'yellow room'* because of the specially coated lamps used in it.

- *Process induced or dust particles*: Particles that contaminate the wafer during a processing step may damage the actual layer or disturb a photolithographic step. This can eventually lead to incorrect circuit performance. For this reason, manufacturing areas are currently qualified by the class of their *clean room(s)*. Modern advanced clean rooms are of *class one*. This means that, on average, each cubic foot ($\approx$ 28 litres) of air contains no more than one dust particle with a diameter greater than 0.1μm. In contrast, a cubic foot of open air contains 10^9 to 10^{10} dust particles that are at least 0.1μm in diameter. The standard applied in *conventional clean rooms* required a class one room to have no more than one dust particle with a diameter greater than 0.5μm per cubic foot. This was because smaller particles could not be detected. A conventional class one clean room is comparable to class 100 in the currently-used classification. Modern factories, however, have class 10-100 for the overall environment, but have class 1 only at the place of handling (add-on *SMIF* (Standard Manufacturing InterFace) *environment*).

  During handling, the class of the environment increases as a result of contamination. When the SMIF environment is integrated in the factory, it can have class 0.01. The throughput of wafers is much higher than with an add-on SMIF environment, because the handling can be done *without* robots. Figure 10.3 shows a photograph of a SMIF environment. Clean room operators need to wear special suits to maintain high quality standards of the clean room with respect to contamination.

- *Electrostatic charge*: Electrostatic charge attracts small dust particles. Very high charge accumulation may occur at a low humidity. This can lead to a discharge which damages the electronic circuits on ICs.

- *The purity of the chemicals*: The chemicals used must be extremely pure to guarantee the high grade of reproducibility and reliability required for ICs.

Figure 10.3: *In a clean room, SMIF pods locally improve the class of the environment (photo: PHILIPS)*

The above parameters, the complexity of the process and the size of an IC determine the yield. Disturbances anywhere during wafer processing may cause defects. The yield Y may be expressed as follows:

$$Y = M \cdot e^{-D_0 A} \tag{10.1}$$

where Y represents the pre-test yield, D_0 the defect density in diffusion and the product defect susceptibility and A the chip area.

The factor M is a yield model parameter which includes the *Area Usage Factor* (AUF), parametric yield loss and clustering effects. The AUF depends on the production equipment used, such as clamp ring positions, stepper wafer layout definition, stepper alignment marker areas or other drop-in structures (if applicable).

The parametric yield is determined by the match of the product design and process window. *Design for manufacturability* will have a positive influence on the yield. Especially in the early phase in the development, yield loss is dominated by parametric/systematic issues. Such

defects are the result of structural failure mechanisms, which may be caused either by physical process defects or by an incorrect or process sensitive design, and are relatively easy to find.

Most non-uniformly distributed defects originate from 'critical' processing steps. Particularly the steps that involve masks with very dense patterns are considered to be potentially critical. These masks include those used to define patterns in thin oxide regions, polysilicon layers and in metal layers.

The factor M, which is area independent, does not include the unusable wafer area close to the wafer edge. The usable wafer area (see figure 10.4) is defined by the total area occupied by complete dies, with the exclusion of a circular edge area (with a width of several millimetres) and a bottom flat side. Current wafers (8" wafers and larger) no longer contain a flat side, but only a notch. The total number of dies within this usable area is called *Potential Good Dies per Wafer* ($PGDW$).

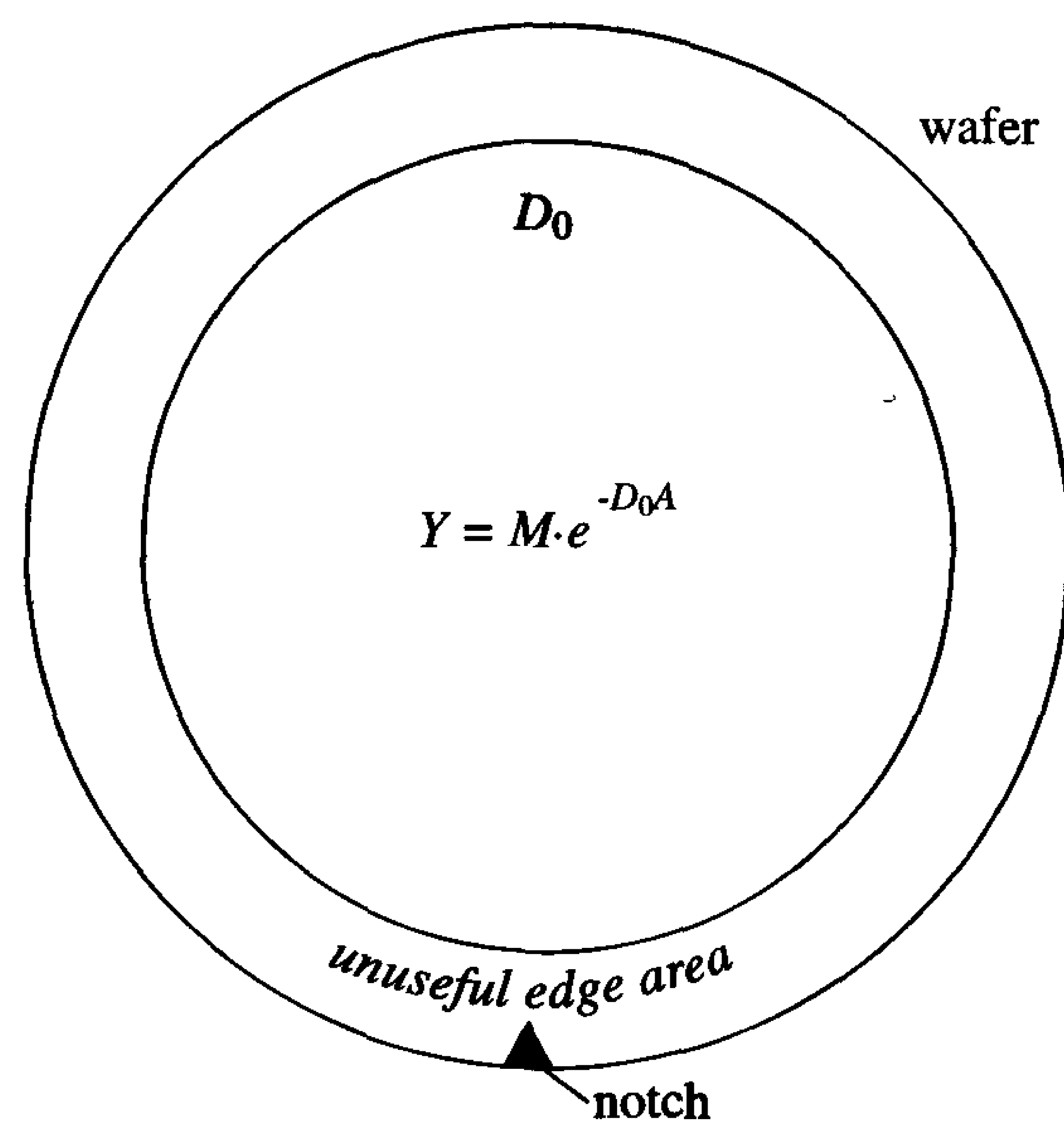


Figure 10.4: *Useful wafer area for PGDW*

The defect density D_0 in equation (10.1) represents the density of defects causing uniformly distributed failures. These are uncorrelated and randomly distributed over the wafer. Examples include dust particles which may affect each process step.

The number of *Functionally Good Dies per Wafer* ($FGDW$) is:

$$FGDW = PGDW \cdot Y \tag{10.2}$$

Particularly in the early phase of process development, M will be relatively low and D_0 will be relatively high. Figure 10.5 shows an example of the yield Y according to equation (10.1) as a function of the die area A for two cases for a 0.25 μm CMOS process. Case 1 shows the situation during an early development stage of a new process, when $M = 0.6$ and $D_0 = 2$ [defects/cm^2]. Case 2 may represent the situation after a year ($M = 0.85$ and $D_0 = 0.5$ [defects/cm^2]). For more mature processes, typical values for $M = 0.97$ and $D_0 = 0.25$ [defects/cm^2] (case 3).

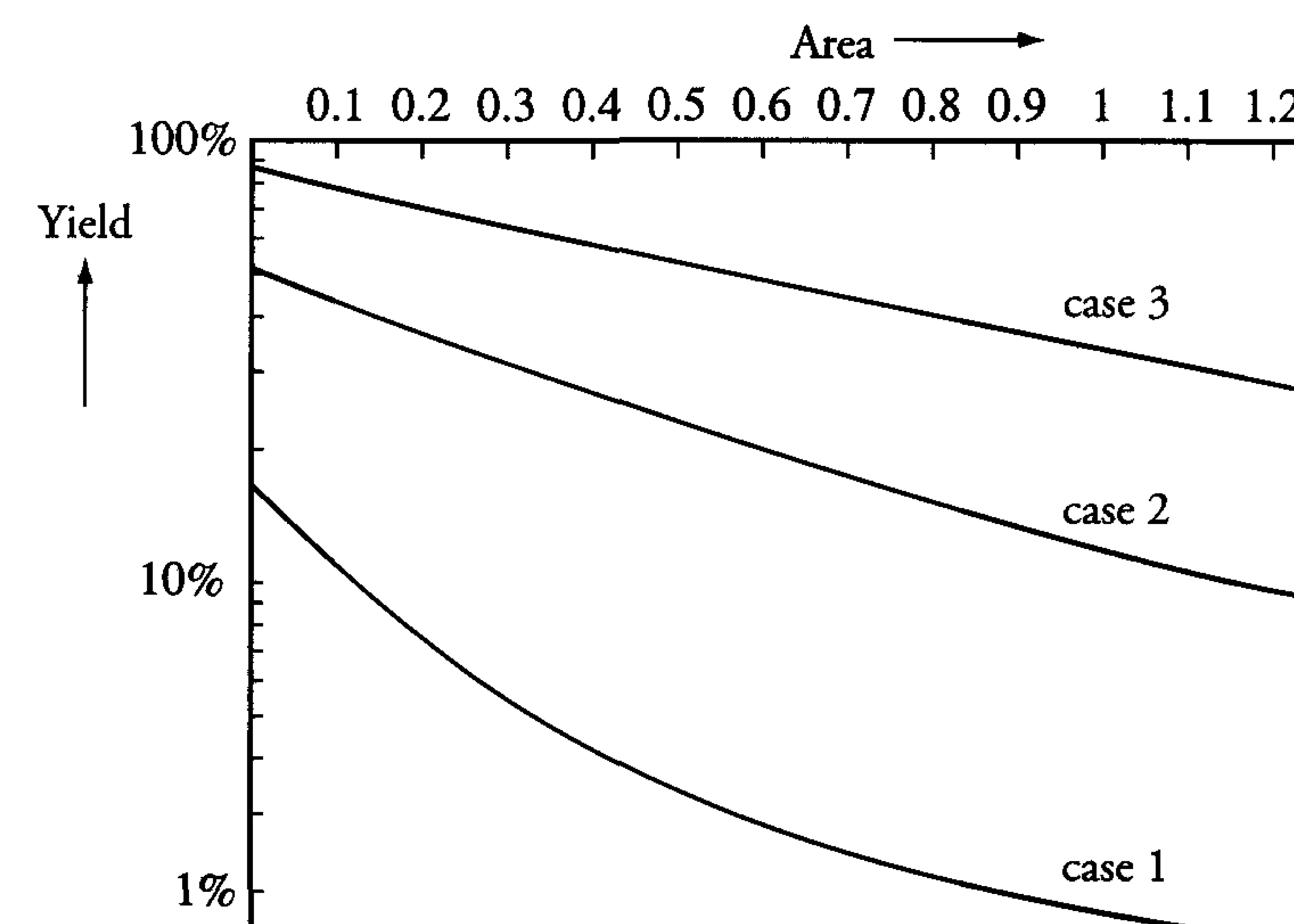


Figure 10.5: *Yield curves at different stages of process maturity*

For the purpose of yield control, *Process Control Modules* (PCMs) are included on wafers. Usually, a wafer contains about five PCMs reasonably distributed over its surface area. A PCM often contains transistors of various sizes (W, L) for the electrical characterisation of parameters such as β and V_T. PCMs also usually contain relatively large structures that facilitate the measurement of possible short circuits, for example, between two layers. Often, more than a hundred parameters can be measured on a PCM.

During the introduction of a new process, the PCMs on all wafers are often measured. When a process becomes mature, the PCMs of about

six randomly chosen wafers in a batch of 25 or 50 wafers are measured. The measurement results are used as an early feedback to control the process.

Finally, when the correct dies are packaged, the final tests are done, which, besides functional, structural and reliability tests, also check the connections between package and die. These final tests, in combination with the pre-test (wafer test), must limit the number of customer returns to a minimum.

10.4 Packaging

10.4.1 Introduction

The dies on each wafer are tested on a probe station, where an ink spot is deposited on each die that fails the test. The dies are then separated by means of a *diamond saw* or a *diamond tipped scribe*. For this purpose, *scribe lanes* of 60 μm to 200 μm width are designed around each die. Most wafers are currently completely sawn through, to separate the individual chips. Sometimes, the wafers are only partially sawn (i.e. not completely sawn through), in which case the wafer is placed on a soft carrier. A roller then applies pressure to the wafer and breaks it into the individual dies. Only the dies without an ink spot, i.e. the good dies, are packaged.

Currently, ICs may contain tens to hundreds of millions of transistors. With such high integration densities, the IC package has become increasingly important in determining not only the size of the component, but also its overall performance and price. Higher lead count, smaller pitch, minimum footprint area and reduced component volume all contribute to a more compact circuit assembly. As the package directly affects factors such as heat dissipation and frequency dependency, choosing the right package is essential in optimising IC performance.

The development of the IC package is a dynamic technology. Applications that were unattainable only a few years ago are now common place thanks to advances in package design. From mobile telecommunications and satellite broadcasting to aerospace and automotive applications, each imposes its own individual demands on the electronic package.

10.4.2 Package categories

Packages can be classified into board mounting methods, construction form and power handling capability. The packages in these "power" categories offer a high thermal dissipation, enabling IC usage in some of the most demanding application areas. Three major categories of packages can be distinguished:

- *Through-hole packages*, whereby the pins fit into plated-through holes in a PCB.
- *Surface mount dual/quad packages*, whereby the pins are soldered on top of the PCB surface.
- *Surface mount area array packages*, which have an array of pins or balls that are soldered on a PCB surface as well.

  Apart from the Pin-Grid Array (PGA) packages, these area array packages belong to the category of surface mount packages, which are gaining an increasingly fast popularity and are expected to rapidly gain market share. Figure 10.6 shows the relative, current and expected market shares of each of these categories.

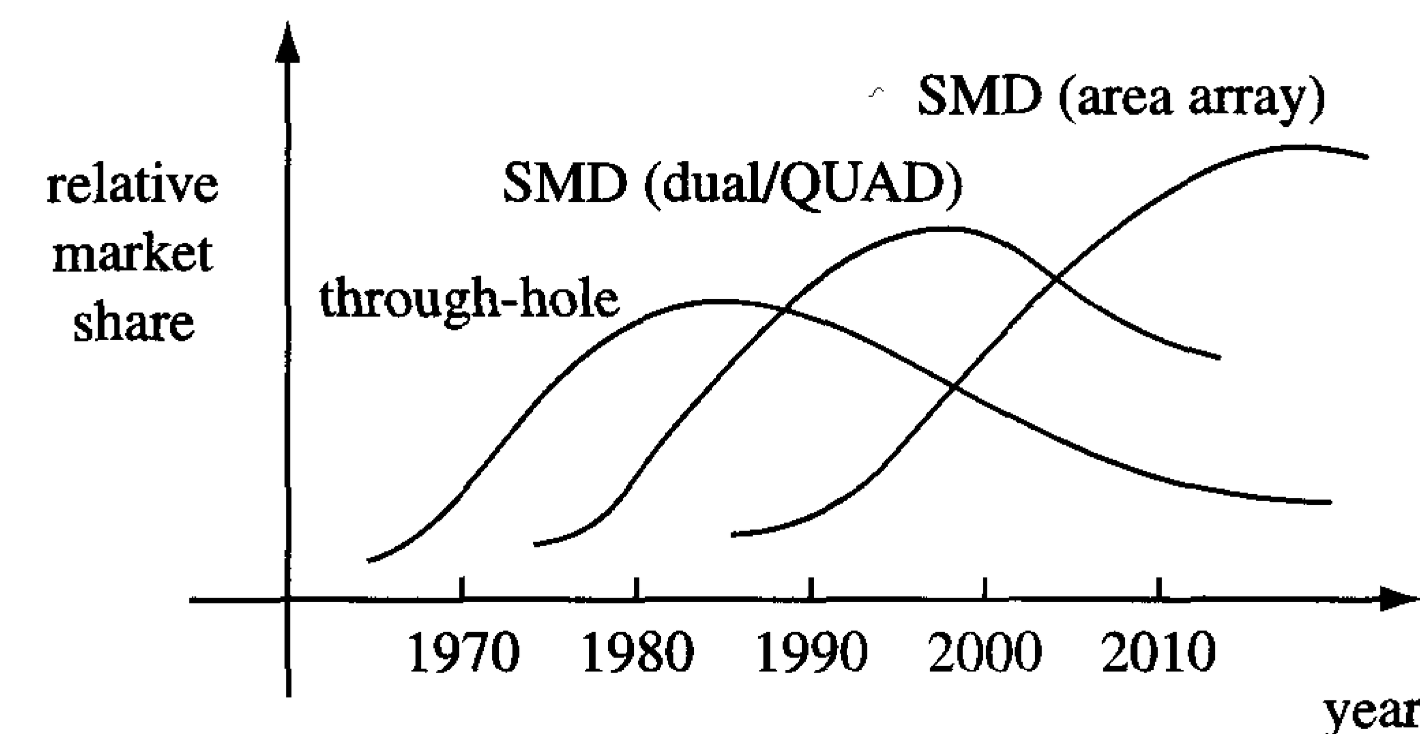


Figure 10.6: *Relative, current and expected market shares*

The major part of through-hole packages consists of dual-in-line (DIL) packages. Although the relative market share has decreased over the years and will probably continue to do so, the absolute number of DIL packages will remain approximately constant for many years to come. Figure 10.7 shows an overview of the major packages [27].

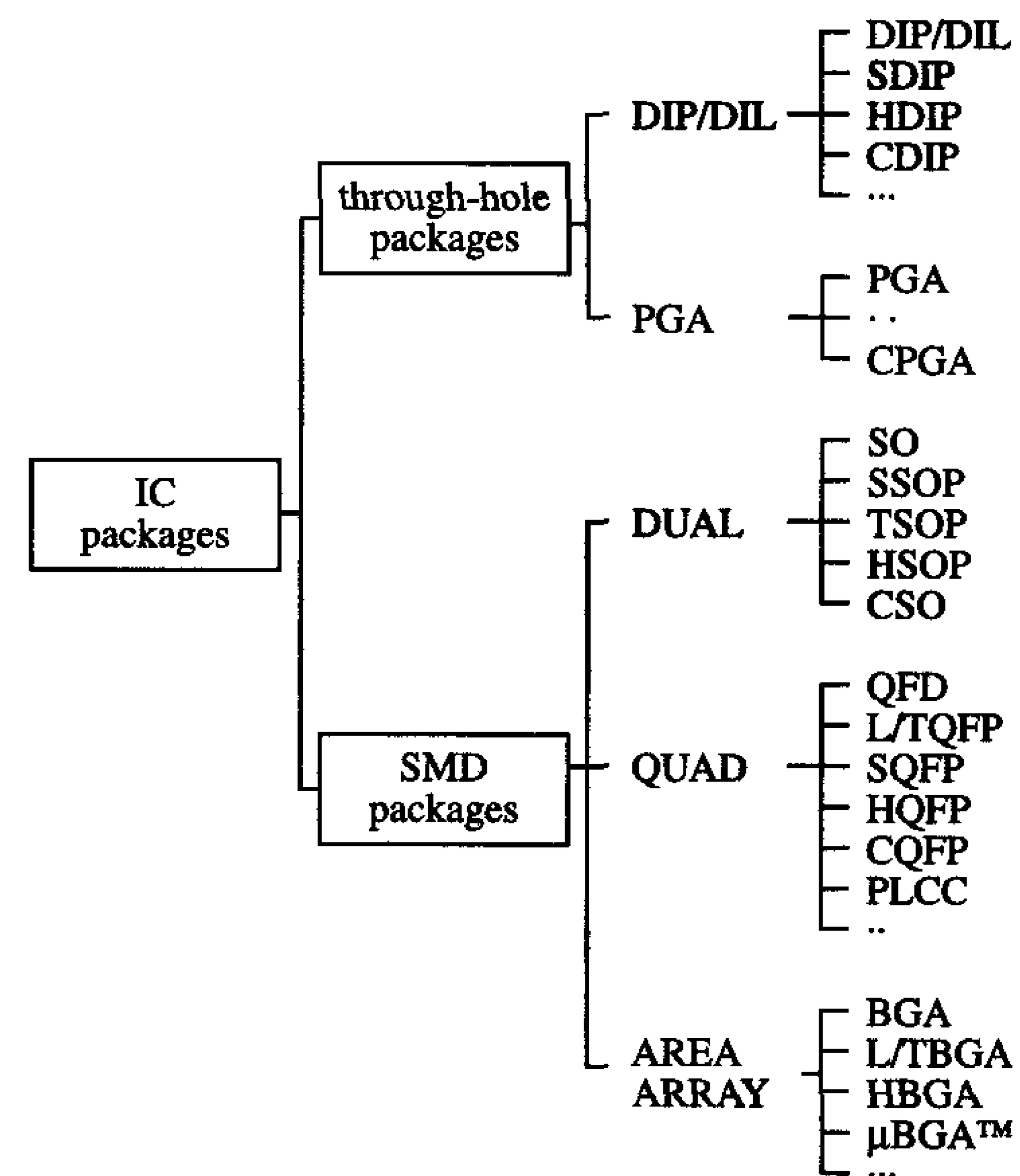


Figure 10.7: *Classification of the major packages*

Different versions have been developed for each of the package categories: thin (T), low-profile (L), shrink (S), heat dissipating (H), power (P.), for different applications. Figure 10.8 shows the IC package trends.

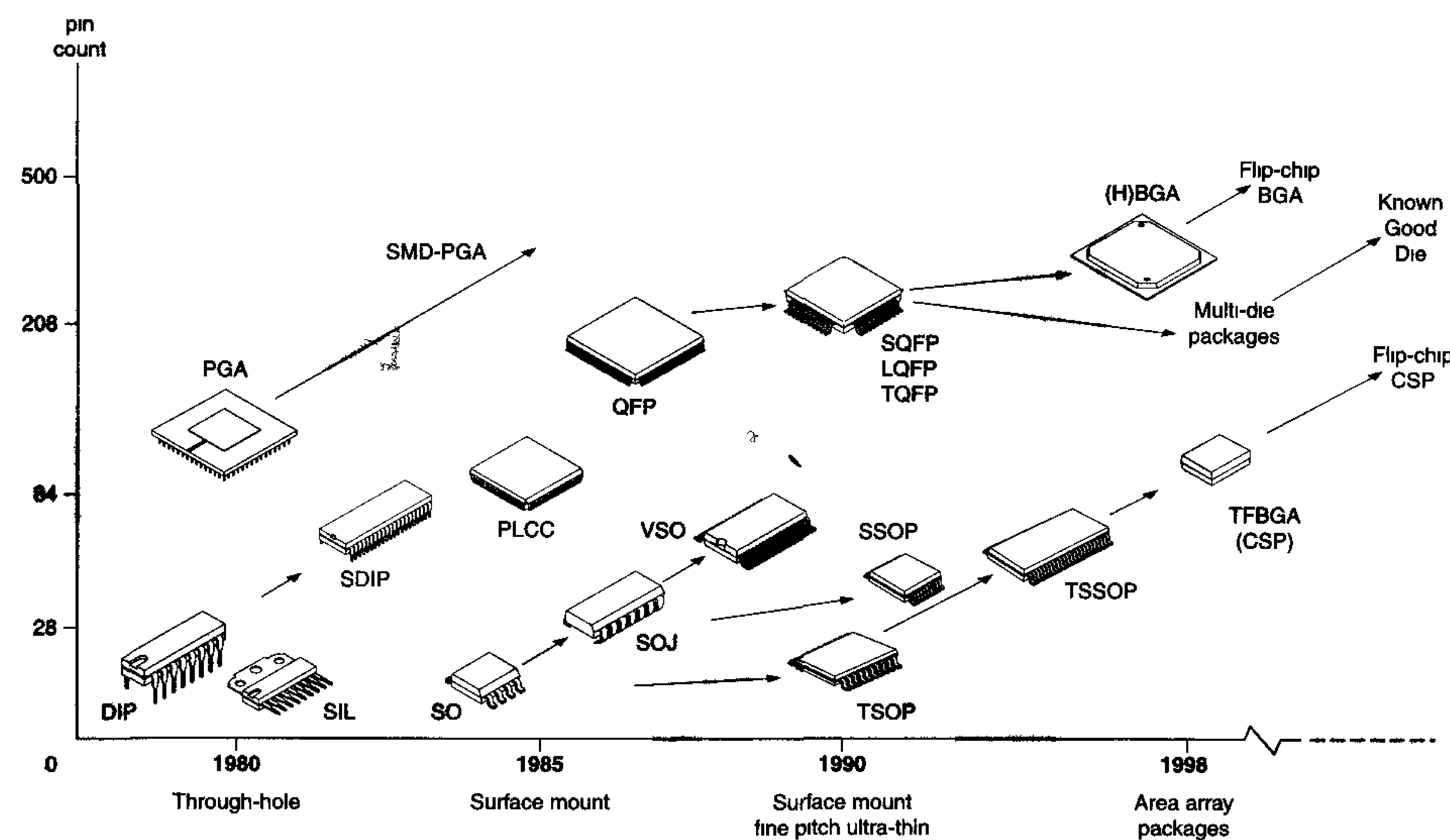


Figure 10.8: *World-wide package trends*

Figures 10.9 and 10.10 show several through-hole and surface mount packages, respectively.

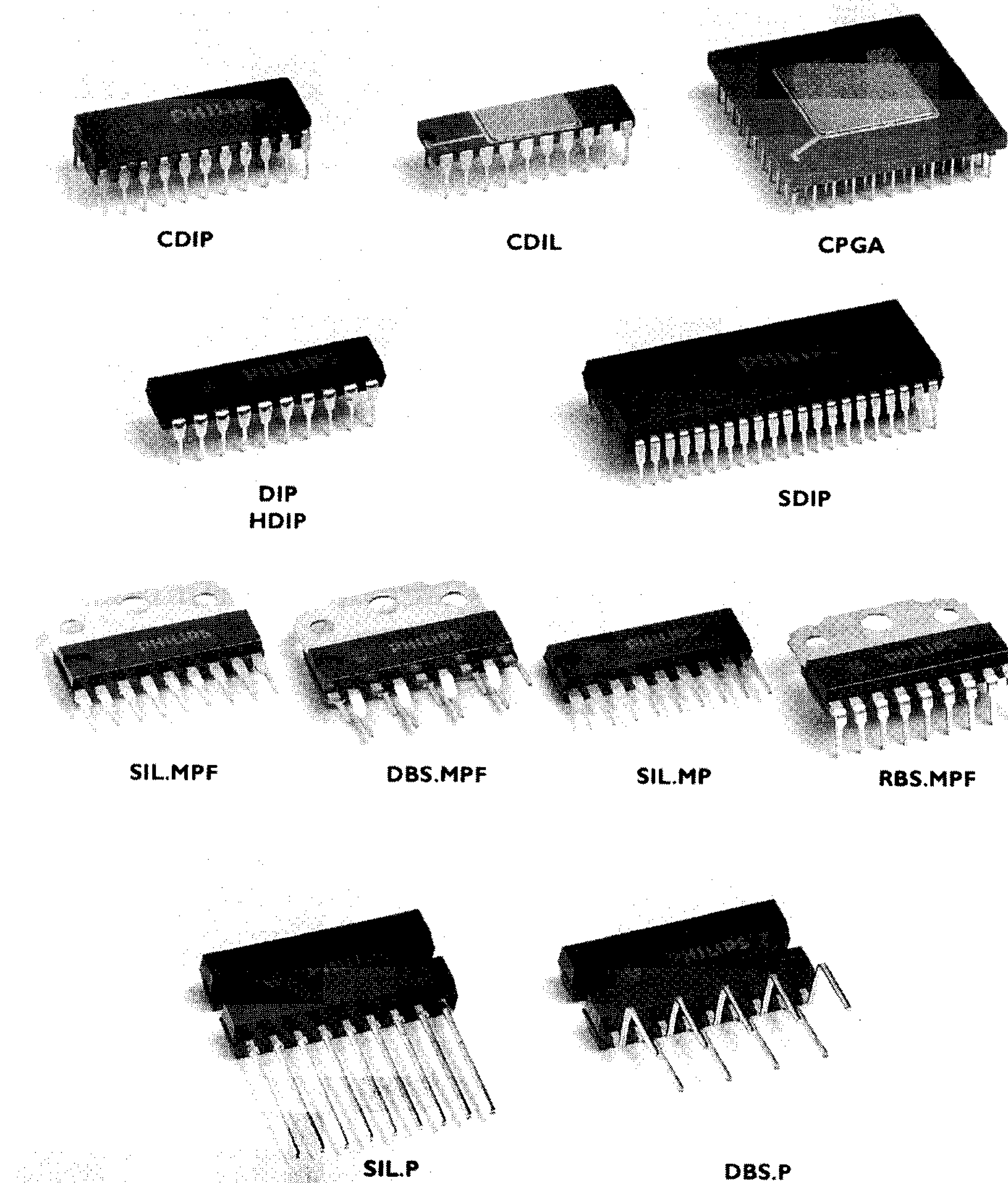


Figure 10.9: *Through-hole packages (Photo: PHILIPS)*

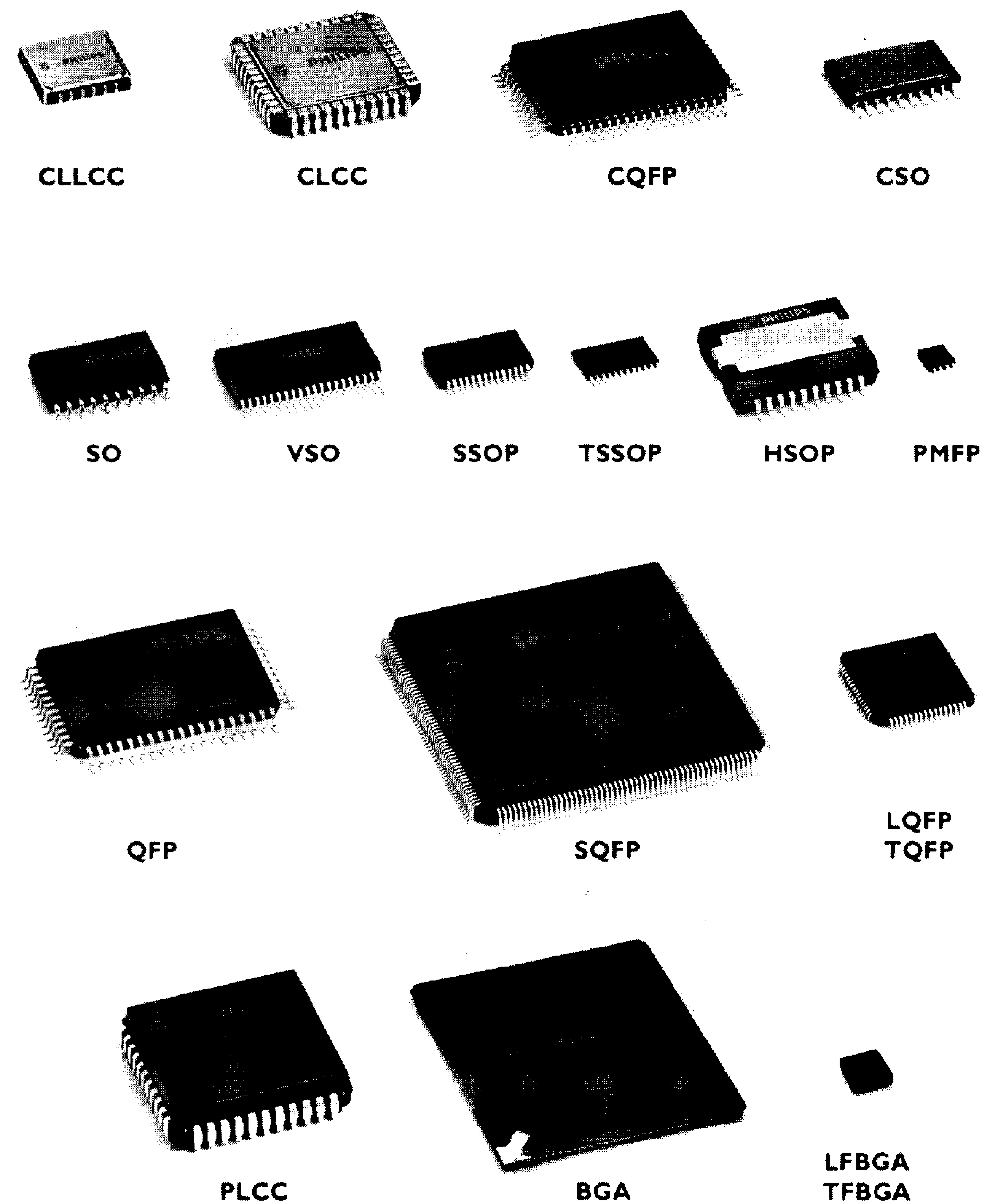


Figure 10.10: *Surface mount packages (Photo: PHILIPS)*

Another classification of ICs can be presented by how the electrical connections between a chip's bond pads and its corresponding package or board connections are made.

10.4.3 Die attachment and bonding techniques

The three common techniques currently used to create these connections are *Wire Bonding* (WB), *Tape Automated Bonding* (TAB) and *Flip-Chip bonding* (FC). Brief descriptions of these techniques follow and the number of connections possible in each of them are shown in figure 10.11.

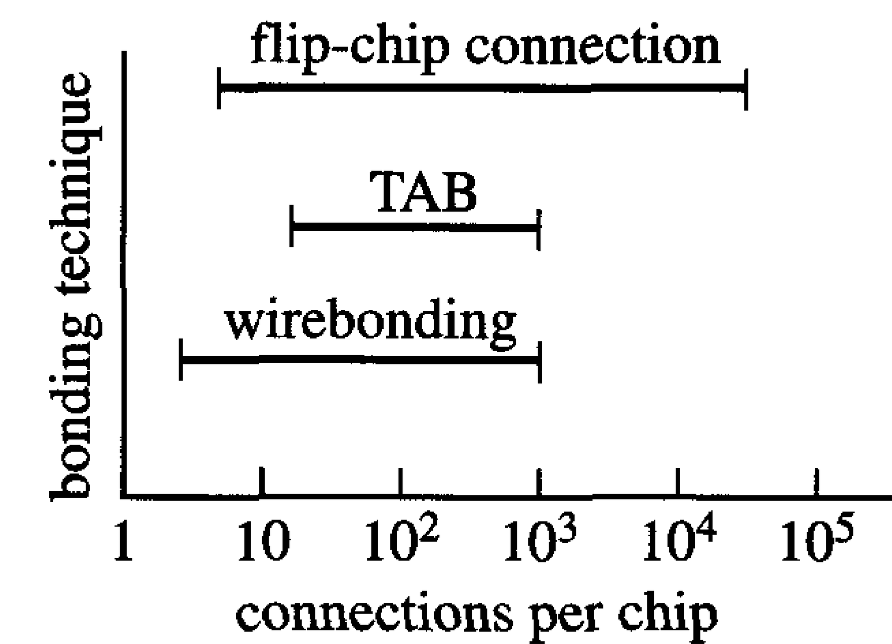


Figure 10.11: *Number of possible connections for the three common bonding techniques*

- *Wire bonding*: The connections in this technique are real wires. Wire bonding is applied in the majority of different packages. The underside of the die is first fixed in the package cavity. For this purpose, a mixture of epoxy and a metal (aluminium, silver or gold) is used to ensure a low electrical and thermal resistance between the die and the package. The wires are then bonded one at a time to the die and the package. Figure 10.12 shows an example of a wire bonded chip.

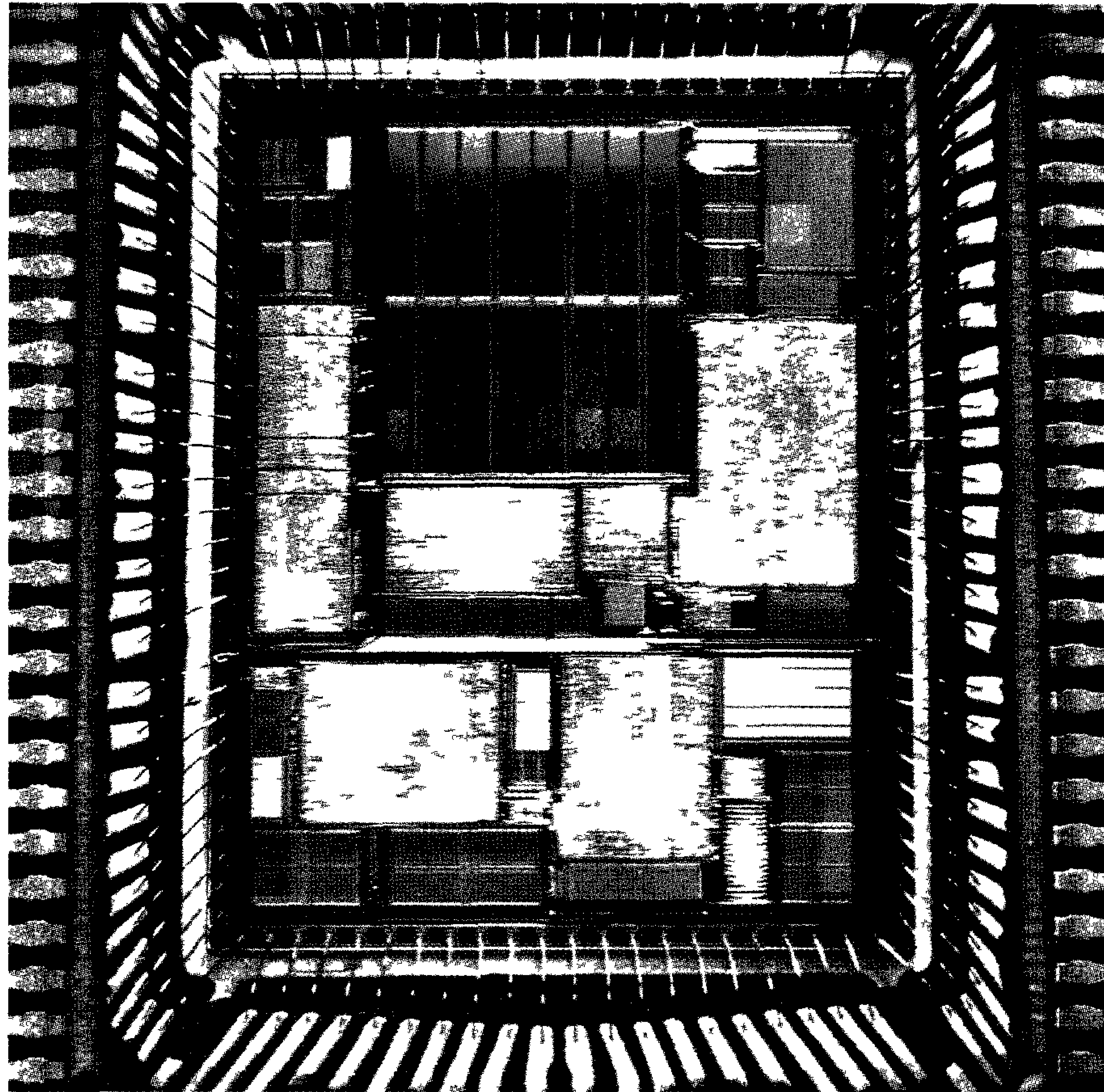


Figure 10.12: *Example of wire bonding (Photo: PHILIPS)*

IC *reliability* is strongly influenced by the quality of the bonding wires. The material, diameter, length and spacing of the wires are important factors. Most applications require a diameter of 25 μm to 40 μm, but power devices can require diameters of up to 150 μm. Very high current variations in high-speed VLSI circuits can cause an increased voltage drop ΔV across the bonding wires. This is because of the inductance (L) of the wires and is expressed as follows:

$$\Delta V = \mathrm{L}\frac{\mathrm{d}i}{\mathrm{d}t}$$

The above voltage drop may become critical during fast circuit operation unless suitable design measures are taken. This topic is addressed in chapter 9. The following two bonding techniques involve direct bonding of the bare die instead of wire bonding.

- *Tape Automated Bonding (TAB)*: This highly automated technique is suitable for packaging large volumes of ICs with small or large numbers of pins. TAB is explained with the aid of figure 10.13, which shows the sequence of events associated with the application of TAB in a PLCC packaging process. Gold 'bumps' are formed on the die's bond pads or on the connection points of an *inner lead frame.* Inner lead frames are copper and are generally mounted on a polymide tape or film. The chip and the lead frame are brought together and a *thermocompression step* is used to convert the gold bumps to actual interconnections between the chip and lead frame. This process is called *inner lead bonding.* The chip and its inner leads are subsequently punched out of the tape and connected to a PLCC *outer lead frame* by means of *outer lead bonding.* The chip and its leads are then encapsulated and the packaged IC is punched out of the outer lead frame.

- *Flip-chip bonding*: This is a direct chip-attach technology, which accommodates dies that may have several hundred bond pads placed anywhere on their top surfaces. Solder balls are deposited on the die bond pads, usually when they are still on the wafer, and at corresponding locations on a board substrate. The upside-down die (or 'flip-chip') is then aligned with the substrate. The temperature is subsequently increased and the solder melts to simultaneously create all connections. The resulting very short connections (low self-inductance (L) values) and the high package density are among the advantages of the flip-chip bonding technique. The complexity and the inability to visually inspect the connections are among its disadvantages. Figure 10.14 shows three substrates that each contain four flip-chips. Flip-chip bonding is also often referred to as controlled-collapse chip connection (C-4).

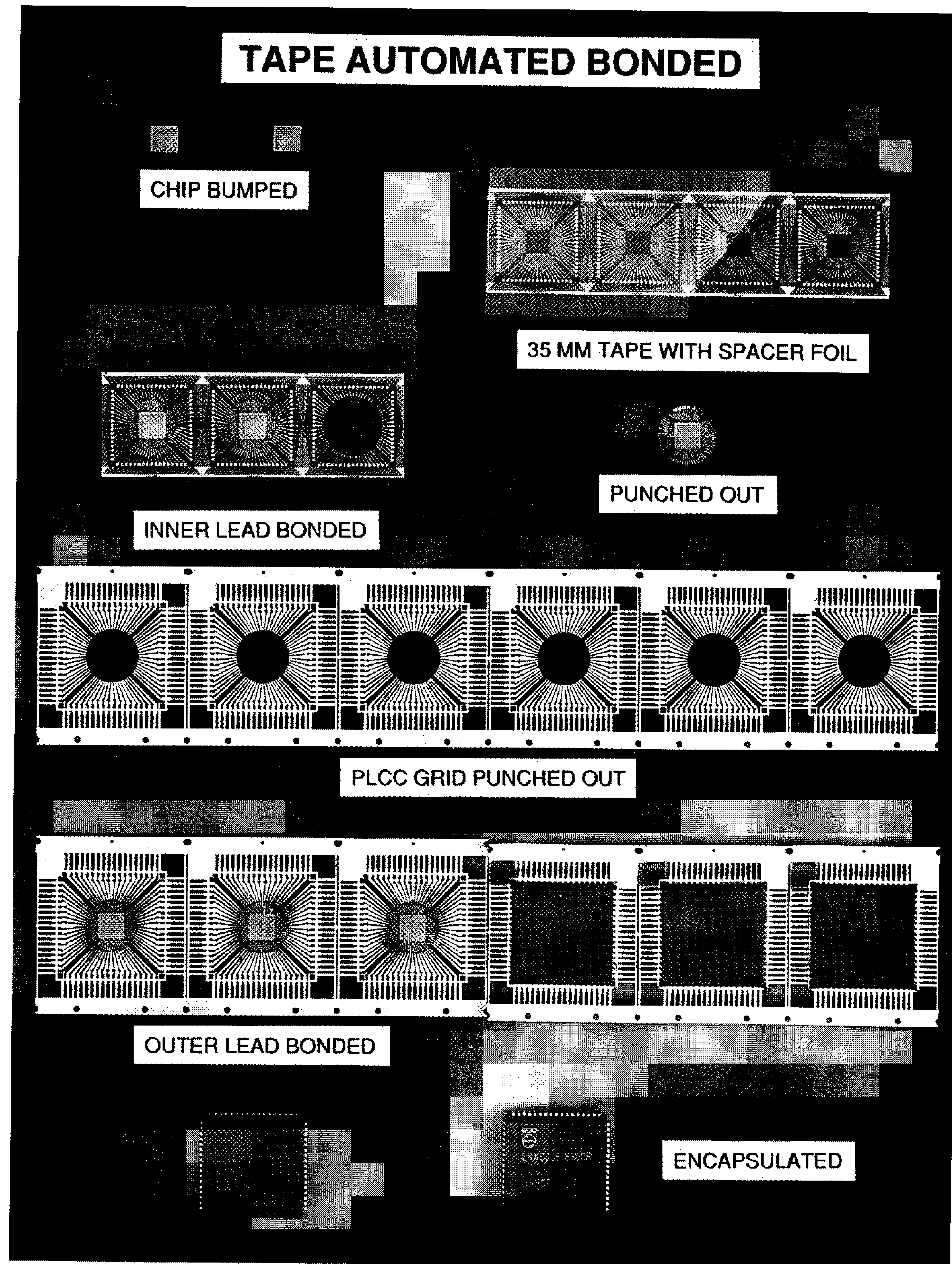


Figure 10.13: *The application of TAB in a PLCC packaging process (photo: PHILIPS)*

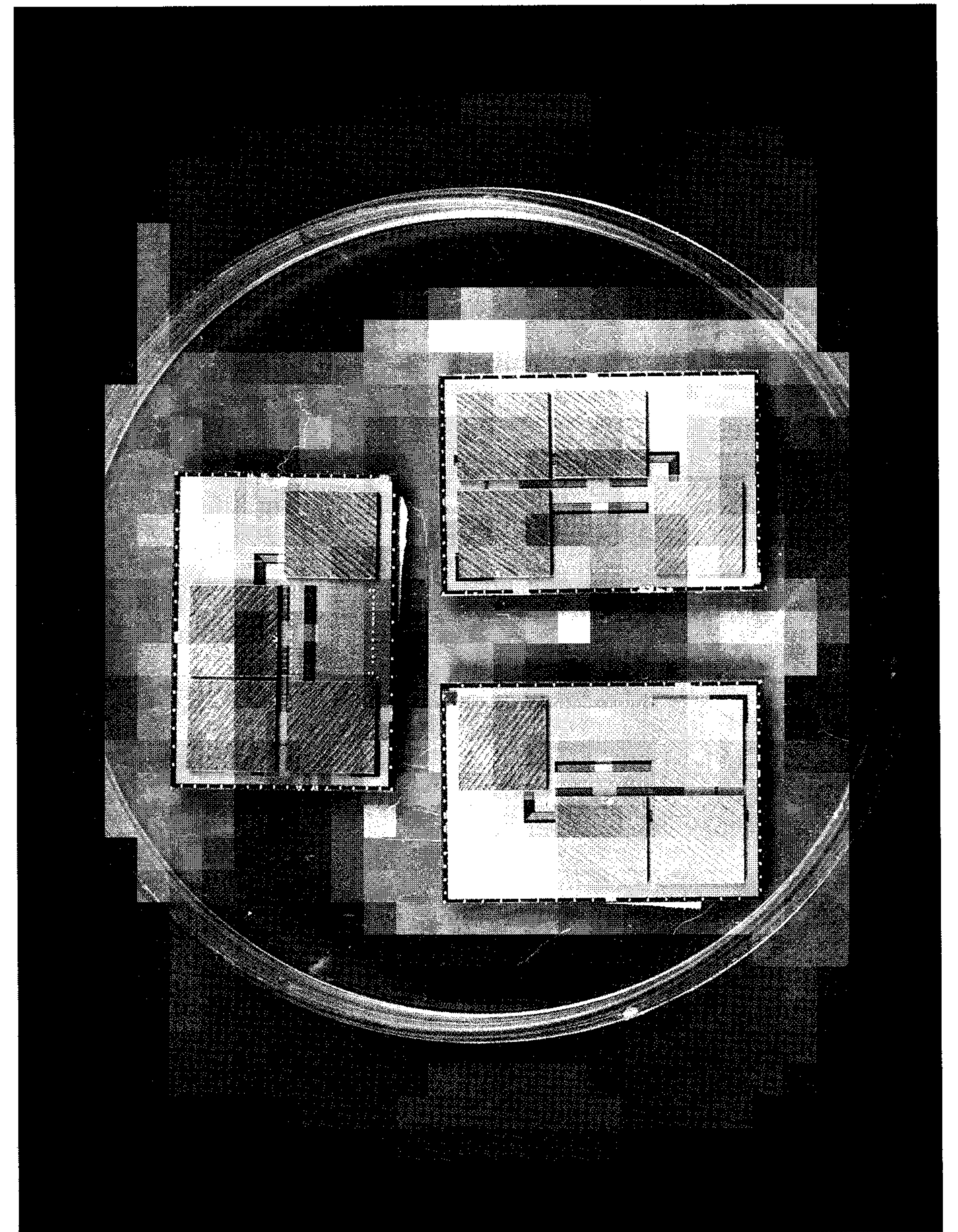


Figure 10.14: *Substrates that each contain four flip-chips (photo: PHILIPS)*

Besides the area and pin count, two other important parameters characterise an IC package: its *thermal resistance* and the *self-inductance* of its pins.

For a given *power dissipation*, $R_{\text{therm.}}$ determines the difference between the ambient temperature and the temperature of a packaged die. Thermal resistance from junction (chip) to ambient is expressed as follows:

$$R_{\text{therm.}} = \frac{\Delta Temp}{P} \quad [^{\circ}\text{K/watt}] \tag{10.3}$$

where P represents the power dissipation of an IC and
$\Delta Temp$ represents the temperature difference between the chip and its environment.

In practice, the maximum IC power dissipation is used to select a package with a suitable thermal resistance. An IC that dissipates 500 mW in a ceramic package with a thermal resistance of 15 °C/W is considered as an example.

If we neglect the thermal resistance from chip to package, then the temperature of this IC will be about 7.5°C higher than the ambient temperature. Complex, high-performance ICs may dissipate tens of Watts, which cannot be sunk by a normal chip package alone. Even specific high-power packages like the ceramic or metal versions in a package family have a too high thermal resistance. For such high power ICs, special heat sinks can be attached to the power packages to increase the total cooling area and reduce the thermal resistance. Several incorporate a fan on top of the package to increase the air flow for cooling of the chip. The Intel Pentium™ is an example of this.

The physical and electrical properties (area, self-inductances (L) and capacitances) of the packages can contribute significantly to the performance and signal integrity of a design. Generally, a small value of L is advantageous for both signal integrity and speed. Because of the existence of inductances, fast current changes ($\mathrm{d}I/\mathrm{d}t$) may cause relatively large voltage changes (ΔV), see also chapter 9, according to:

$$\Delta V = L \cdot \frac{\mathrm{d}I}{\mathrm{d}t} \tag{10.4}$$

Developments in package technology have led to a reduction in the values for L. These values vary per pin with the different lengths of the bonding wires and leads (about 10 nH/cm) and with the different bonding structures. A dual in-line has the largest lead inductance (2-50 nH),

because of the absence of a ground plane (usually) and because of its longer lead lengths. TAB (Taped Automated Bonding) usually results in smaller inductance values (0.5-10 nH), because of the presence of a ground plane and shorter and thicker leads. Packages that use multilayer ceramic substrates with power and ground planes, like PGAs (Pin-Grid Array), also have small inductance values for their pins. When the chip is directly attached, via TAB or flip-chip technology, then one level of packaging (bonding and package leads) is eliminated, resulting in very low inductance values. As an example: flip-chip technology, in which the chip with its solder bump connections is directly attached to the substrate, has the lowest inductance values (0.15-1 nH).

In the packaging of ICs, we distinguish several levels of interconnections:

- First level of interconnection: chip to package connection.
- Second level of interconnection: package to PCB connection.
- Third level of interconnection: PCB wires.
- Fourth level of interconnection: PCB to system (back-planes) connection.

It should be clear that not only the package-to-board connections must be optimised for high-performance ICs, but that a well-developed board and system interconnection structure is also required. The use of flip-chip technologies reduces the chip to carrier interconnections. The use of Surface Mount Technology (SMT) reduces the board interconnections, see figure 10.15.

More about bonding techniques can be found in [29] and [30].

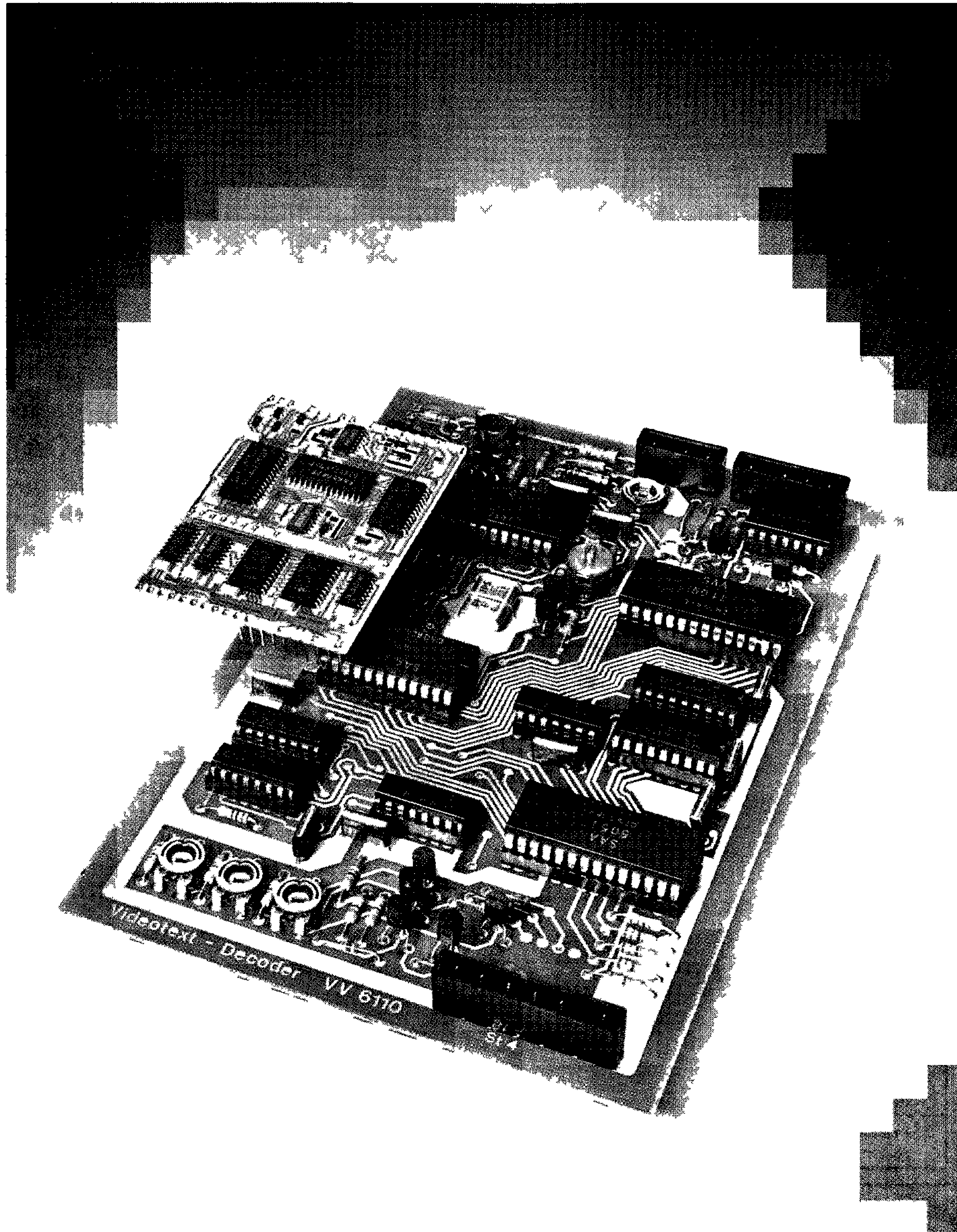


Figure 10.15: *Area reduction afforded by SMT, compared with conventional through-hole packages (Photo PHILIPS)*

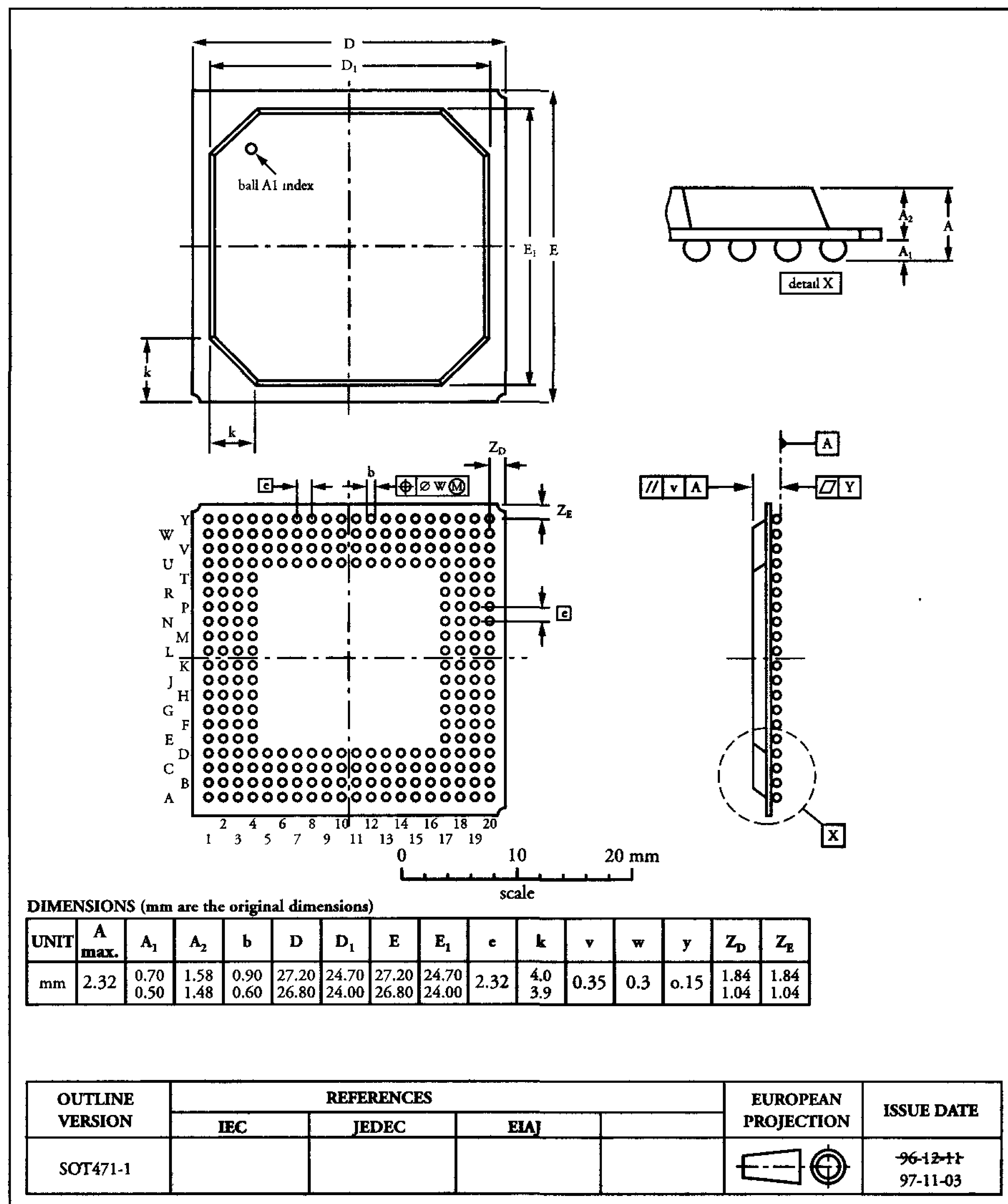
10.4.4 Advances in IC packaging technology

During the eighties, SMD technology became very popular in consumer electronics. Because of the continuous drive for increased system complexity and higher performance, several promising developments in packaging are gaining in market share: Ball-Grid Arrays (BGAs), Chip Scale Packages (CSPs) and Multi-Chip Modules (MCMs).

Ball-Grid Array (BGA) packages

According to the SIA roadmap in table 11.3 the number of pads on a chip will increase dramatically. This is because multi-million gate designs will require a lot of external connections to get both data and power supplied to and from the chip. The increased computing power requires a higher bandwidth of the chip to board interface. This is why the area array packages are rapidly gaining popularity. Analogous to an area array CSP (Chip Scale Package), the *Ball-Grid Array* (BGA) package connects peripheral IC pads to an array of solder bumps on the package bottom. The BGA package allows a very high pin count chip to be attached to a board with surface mount technology. The number of pins varies from less than 200 for chip-scale BGAs (μBGA) to about a thousand or more for the Super Ball-Grid Array (SBGA). Figure 10.16 shows a BGA256 package outline drawing [27]. Figure 10.17 shows a photograph of a BGA.

Compared to through-hole packages, BGAs show similar performance increase as flip-chip packages as a result of the direct attachment of the solder bumps to the board. The array of solder bumps not only supports a high pin count for increased signal band width but also serves to sink a relatively large part of the heat dissipation from the chip to the board.



BGA256: plastic ball grid array package; 256 balls; body 27 × 27 × 1.55 mm SOT471-1

DIMENSIONS (mm are the original dimensions)

UNIT	A max.	A_1	A_2	b	D	D_1	E	E_1	e	k	v	w	y	Z_D	Z_E
mm	2.32	0.70 0.50	1.58 1.48	0.90 0.60	27.20 26.80	24.70 24.00	27.20 26.80	24.70 24.00	2.32	4.0 3.9	0.35	0.3	o.15	1.84 1.04	1.84 1.04

OUTLINE VERSION	REFERENCES				EUROPEAN PROJECTION	ISSUE DATE
	IEC	JEDEC	EIAJ			
SOT471-1						~~96-12-11~~ 97-11-03

Figure 10.16: *Outline drawing of a BGA256 plastic ball-grid array*

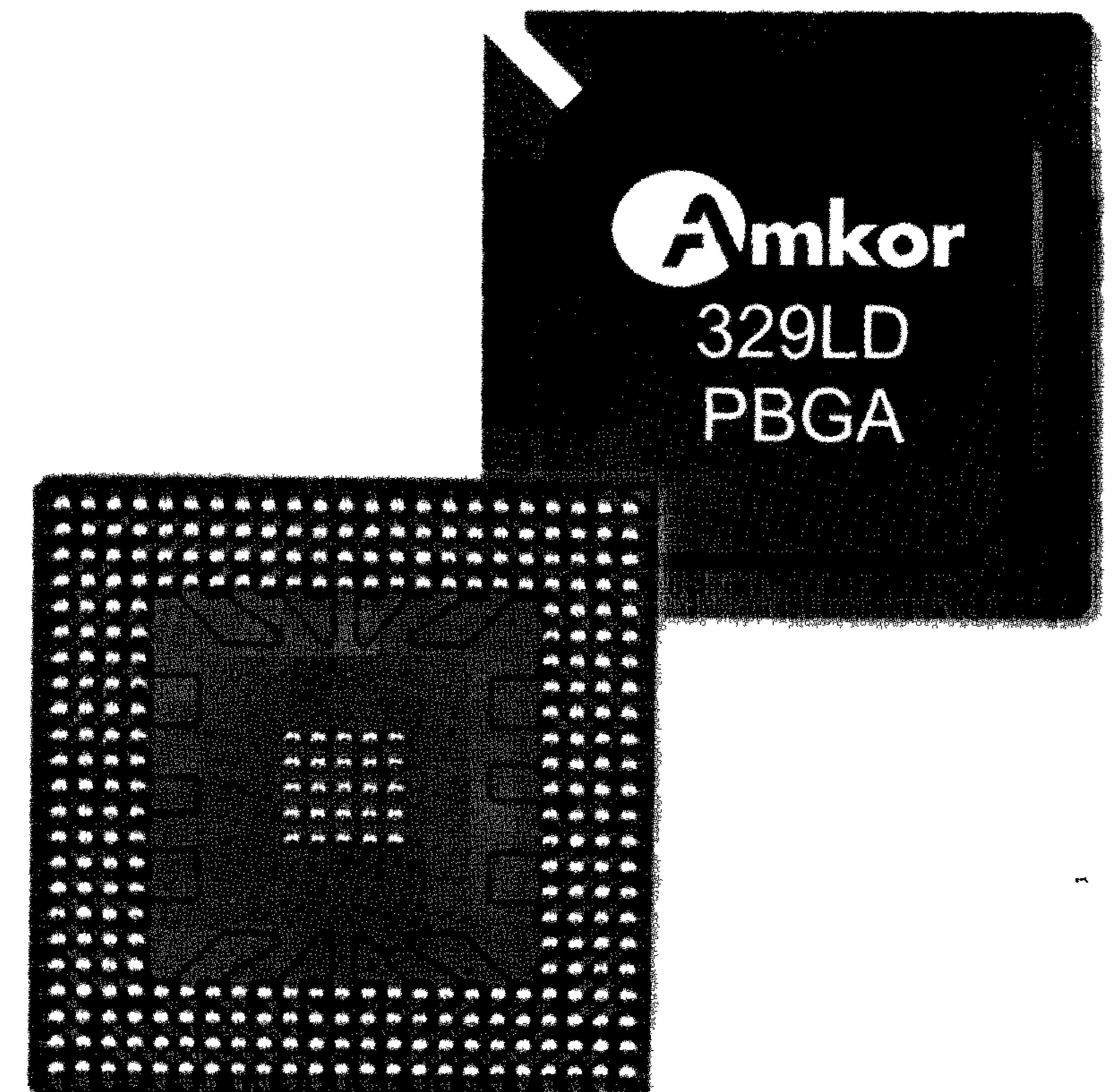


Figure 10.17: *Photograph of a BGA (Photo: Amkor)*

Chip Scale Packages (CSPs)

A *Chip Scale Package* (CSP) is defined to be a package whose sizes are only about 20% larger than the die itself. The required board area is comparable to flip-chip, but it offers the advantage of a mechanical protection because the die is encapsulated. CSP technologies require additional preparation of the dies with solder bumps, which is usually performed at the wafer level [26]. At the moment, this process is not yet widely available.

Most flip-chip ICs contain area bumps located at a relatively large pitch. Because the flip-chip is a bare die attachment, the same footprint is required on the substrate to which it must be attached. Therefore, a reduction of the pitch of the bumps increases the complexity of the flip-chip attachment process. CSP can be performed at the wafer level or per individual die by using thin film-like technologies to redistribute and reroute the peripheral bond pads to an area array on top of the die, see figure 10.18.

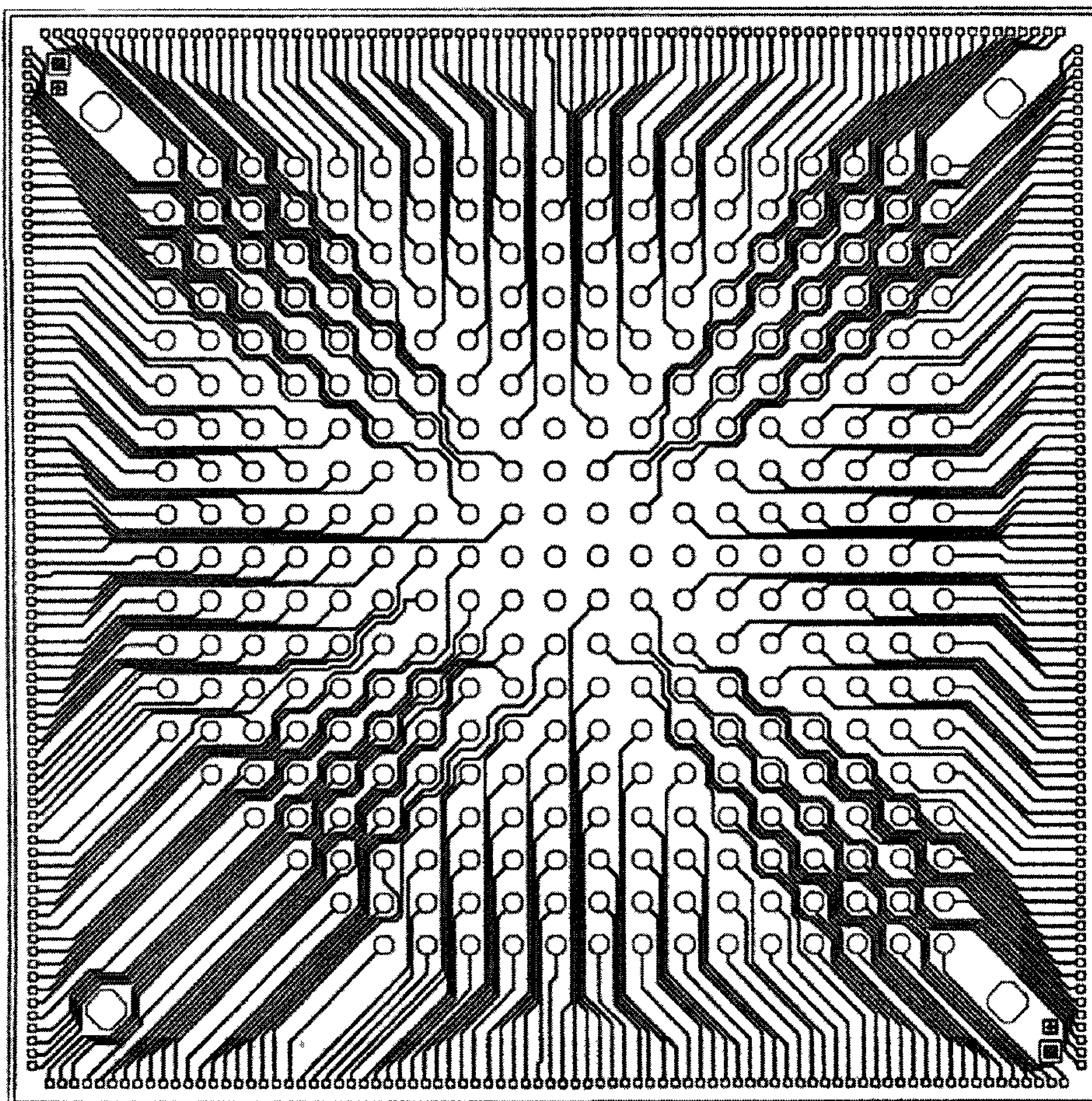


Figure 10.18: *Pad redistribution; from peripheral pads to an area array of pads (source: Technical University of Berlin)*

After that, bumps can be formed on the new set of bond pads. During a further encapsulation step, the top of the bumps are left exposed.

Next, an outer bump is deposited on each inner bump. These outer bumps serve to attach the chip to the printed circuit board. In this flip-chip attachment step, the die is turned upside down and positioned such that its solder bump pattern matches the pads on the substrate. The final attachment to the circuit board requires a heat and/or pressure step. CSP advantages include:

- ease of handling
- ease of (standardised) test and burn-in procedures
- assembly process compatible with existing SMT
- multiple sources available
- small size and low inductance pads.

CSPs are not expected to erode the SO/TSOP or BGA package designs [21]. However, they do provide an entry in the expansion of the market for bare die and known good die because of the test and burn-in capabilities at package level.

Commercially available CSP ICs include: mini BGA, μBGA, fine pitch BGA, flip-chip BGA and a lot of other chip size packages. Memories were the first commercial CSP products and they still count for the largest application for chip-scale packages. Portable consumer products are currently turning to CSP because of their performance and area advantages.

Multi-Chip Modules (MCM)

An *MCM* is a package that offers room to closely-spaced ICs on a single substrate. To increase signal performance and minimise signal propagation delay across the interconnections, bare chip attachment technologies (such as flip-chip) are often applied to connect the different internal MCM dies. The increasing system requirements for an overall small size and weight, a high performance and reliability will drive bonding techniques from wire bonding to TAB and flip-chip, and package technology from single-chip to MCM packaging.

To the outside world, an MCM looks like a single chip and can be treated as such. Several MCM technologies have been presented over the years [22]: MCM-C (co-fired), MCM-D (deposited) and MCM-L (laminated). In MCM-C, a thick-film technology (which is normally

used in hybrid circuits on ceramic substrates) is applied. In MCM-D, the metal tracks are formed by a multilayer structure of thin or thick film metal on dielectrics. Organic printed circuit laminates are used as substrates in the MCM-L technology. While MCM-L is used in low-end, fast time to market products, MCM-Ds are capable of supporting many (> 100) chips on one large substrate [22].

Cost effectiveness is the reason for merging the different MCM technologies into hybrids. The use of MCM technologies within standard IC packages offers low-cost solutions and a lower customer acceptance threshold. For a customer, an MCM should resemble a standard package as much as possible. Standard packages in which MCM technologies can be applied include QFPs and BGAs. Figure 10.19 shows an MCM-PBGA used in a consumer product.

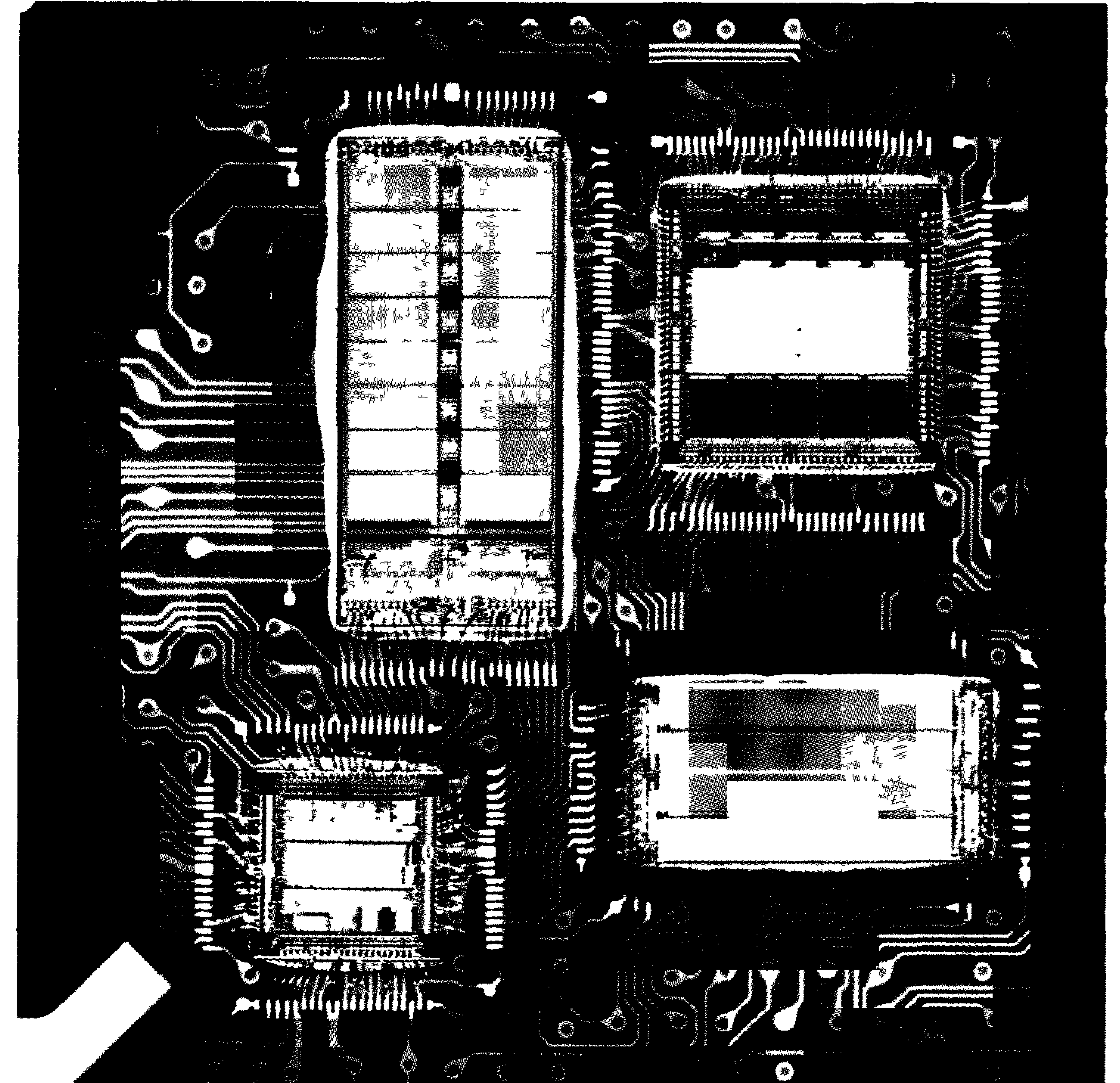
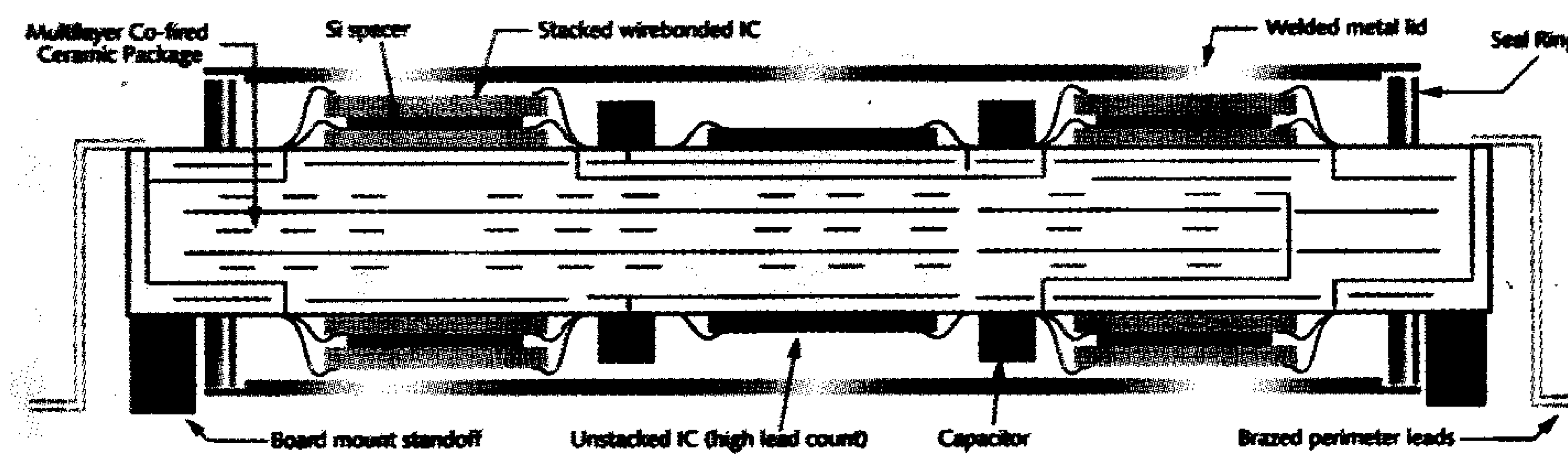
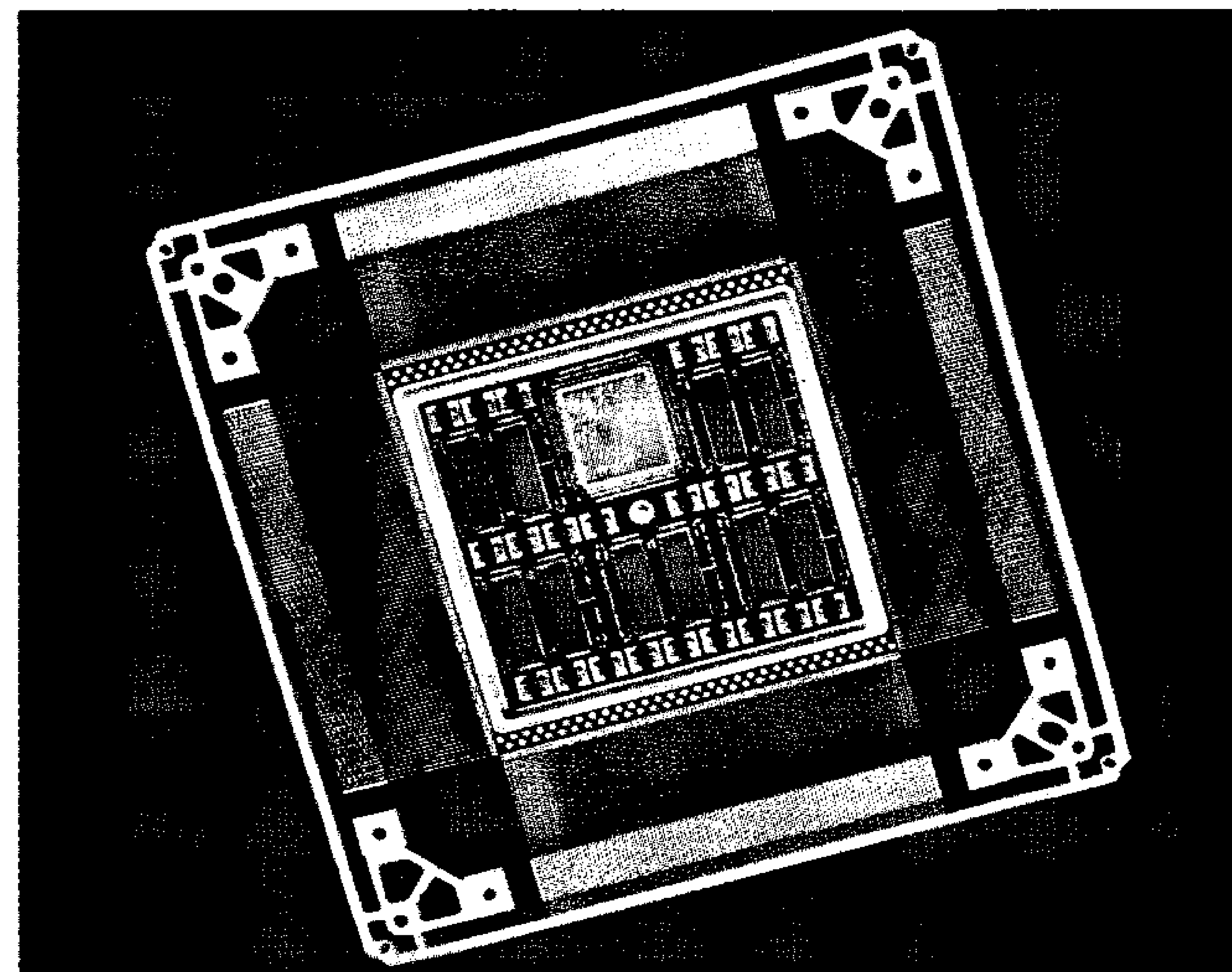


Figure 10.19: *Example of an MCM-PBGA used in a consumer product (Source: PHILIPS/Amkor-Anan)*

Most current generations of MCMs are single sided. However, advanced double-sided MCM technologies with stacked dies are being developed for dedicated applications. Figure 10.20 shows a cross-section of such a concept for a spaceborne application. It also shows the top side of an assembled memory MCM and the bottom side of an assembled 32-bit processor MCM. Both MCMs are double-sided QFPs with matching lead frames, seal rings, lids and dimensions [23].





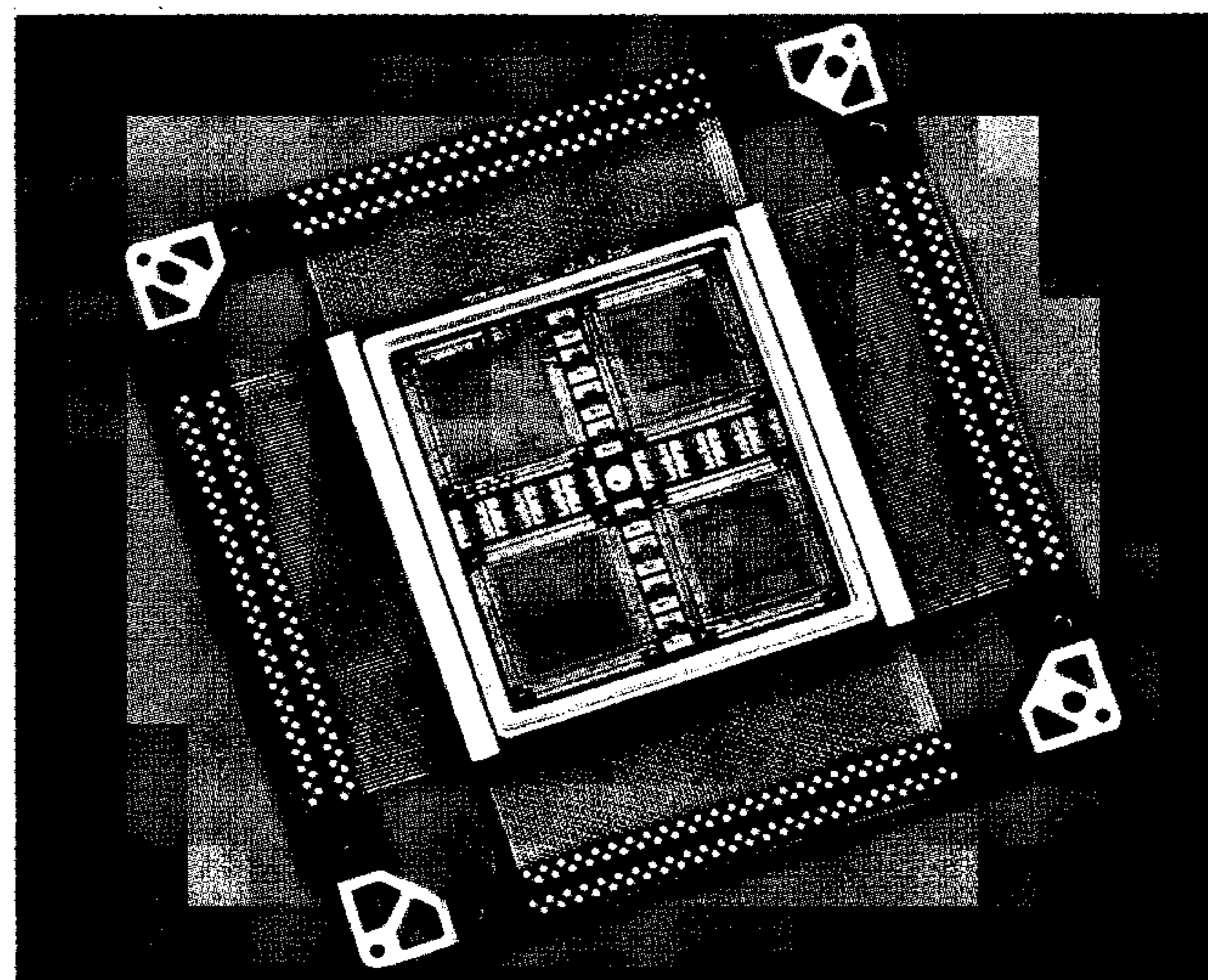


Figure 10.20: *Cross-section of a double sided MCM substrate with top and bottom side MCM ICs (Photo: Honeywell)*

In summary, an MCM is a packaging technology that fully supports chip-level performance through different levels of short interconnections and satisfies system and reliability requirements.

Testing is one of the large issues in MCM technology. Bare dies have been sold for decades to hybrid IC manufacturers at a price reduction of 25 to 50%, compared to packaged devices. These dies were not even tested. A prerequisite for a high degree of MCM production yield is to start the assembly process with completely tested fully functional ICs in bare die form, which are "known-good" die.

The ideal definition of *Known-Good Die (KGD)* will state that a high (> 99,9%) confidence level is required for ICs to meet their specifications, to be free of defects, and to remain so after such tests such as burn-in and environmental stress screening, during the assembly process and during lifetime field exposure [24]. Removing a bad component means a large loss because one bad die can ruin a very expensive module. The problem of KGDs is increasing seriously as the ICs and MCMs become more complex [25]. Figure 10.21 shows how fast the MCM yield will reduce when the average yield of the individual chip is not close to 100%.

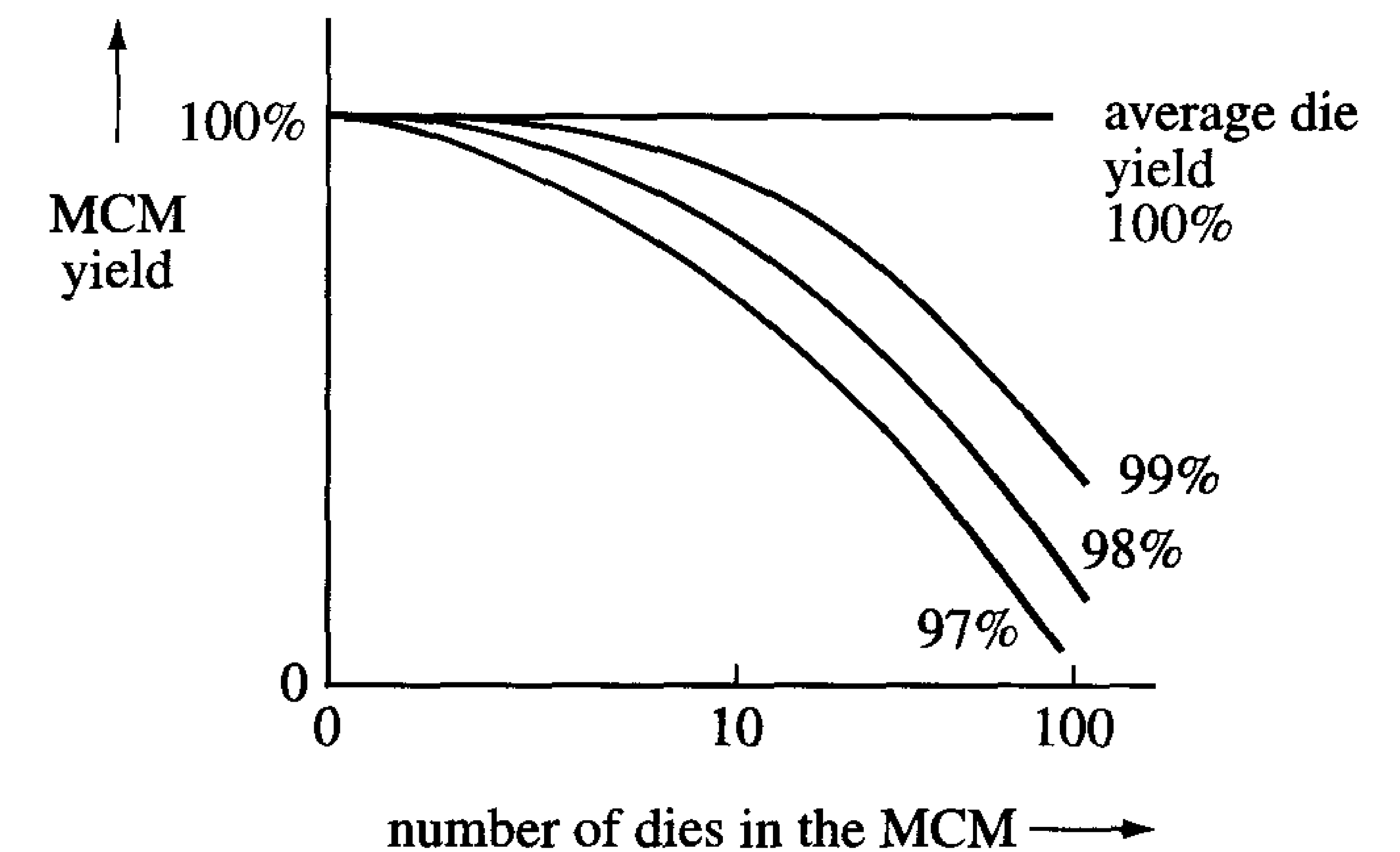


Figure 10.21: *MCM yield curves as a function of the average yield of the individual dies*

Most of the problems, however, turn out to be related to ESD and overstress rather than problems inherent in the dies [25]. Known-good assembly is therefore as much of a prerequisite for a high MCM production yield as the known-good die level.

10.5 Quality and reliability of packaged dies

Various quality and reliability tests are applied to packaged ICs before they are approved for sale or for application in high volume production. Many of these tests are standardised. An insight into the background to these tests and their implementations is provided below.

10.5.1 Quality

Vulnerability to *electrostatic discharge* and sensitivity to *latch-up* are two important quality criteria on which chips are tested.

Parasitic bipolar devices in all CMOS chips form a *thyristor* between its supply and ground terminals. Activation of this thyristor results in latch-up. The result is a dramatic increase in current consumption and a chip malfunction. A chip's latch-up sensitivity can be tested by sequentially applying a voltage of one-and-a-half times the maximum specified voltage to each pin, while limiting the available current to, for example, 500 mA. The actual current consumption is observed for signs of latch-up.

Since ESD and latch-up sensitivity can be influenced by the design, these topics are discussed in detail in chapter 9. In addition, chapter 9 describes technological and design measures which can be taken to reduce the chances of failure in the associated tests. ESD tests and the related quality requirements are also discussed in that chapter.

10.5.2 Reliability

The increasing complexity of ICs means that their reliability has a considerable effect on the reliability of electronic products in which they are applied. Reliability is therefore an important property of an IC and receives considerable attention from IC manufacturers. Related tests subject an IC in active and non-active states to various *stress conditions*. This facilitates rapid evaluation of the IC's sensitivity to external factors such as temperature changes and humidity. The most important *reliability tests* are as follows:

- *Electrical endurance test*: This test exposes an IC to a high temperature (125 °C to 150 °C), while its supply voltage is maintained at the specified maximum or higher value (3 V to 4 V for an IC with a nominal supply voltage of 2.5 V). Constant and varying signals

are applied during the test, which may last for 1000 hours. The electrical endurance test reveals the following:

  - *Infant Mortality*, i.e. faults which are likely to arise in the early months of an IC's normal application;
  - *Early Failure Rate*, i.e. faults which are likely to arise after half a year;
  - *Intrinsic Failure Rate*, i.e. the probability of a fault occurring during a specified number of years;
  - *Wearout*, i.e. effects of prolonged use on the product.

  Faults that are observed during the electrical endurance test can usually be traced to errors in the manufacturing process which preceded IC packaging.

- *Temperature-cycle test*: This test emulates practical temperature changes by exposing the product to rapid and extreme temperature variation cycles. The minimum temperature in each cycle is between −55 °C and −65 °C. The maximum temperature is 150 °C. The number of cycles used is typically five hundred. The test is carried out in an inert gas and/or an inert liquid. The main purpose of the temperature-cycle test is to check the robustness of the package and the robustness of the connections between the package and its die. The test should reveal possible incompatibilities between the temperature expansion coefficients of the various parts of an IC, e.g. the die, the lead frame and the package material.

- *Humidity test*: This test exposes an IC to a relative humidity of 85 % in a hot environment (85 °C to 135 °C). The test reveals the effects of corrosion on the package and provides an indication of the quality of the *scratch-protection layer*. Usually, the corrosion process is accelerated by applying different voltages to consecutive pairs of pins, with 0 V on one pin and 2.5 V (for a 0.25 μm CMOS chip) on the other. Most humidity tests last 1000 hours.

The required specifications of an IC depend on its application field, envisaged customer, status and supplier. It can therefore take a number of years before the quality and reliability of a new IC in a new manufacturing process reaches an acceptable level.

10.6 Potential first silicon problems

When first silicon, either on a wafer or mounted in a package, is subjected to the first tests, one or even all tests might fail. Passing a test means that everything must be correct: the technology must be within specification, the tester operation must be correct, the test software (vectors and timing) must be right, connections between tester and chip (interface and probe card) must be proper and, finally, the design must be right. Therefore, passing a test means the logical AND of correct processing, correct tester and interface operation, correct software and, finally, correct design.

Especially in the beginning of the engineering phase of first silicon, problems may occur with the tester, its interface or the test software. Also, problems may arise from marginal processing or marginal design.

The following subsections discuss each of the different categories of failure causes.

10.6.1 Problems with testing

Very complex ICs contain millions of transistors and can have several hundreds of bond pads. It is therefore a tough job to locate the failure somewhere in the chip, when, for instance, one output signal fails. The relation between an error inside the circuit and the output pins is very vague. Dedicated advanced testing techniques are already included in the design to support testing. Because not all functional blocks have (direct) access to output pins, they will be part of a scan chain. In many cases, these scan chains run (and are tested) at lower frequencies. A potential problem is that such blocks are found to operate correctly on the tester (at a lower frequency) but may show failures when the chip is put in the application (board; speed check). Therefore, the chip should run at the same speed during scan test as in the application.

Test data, such as test vectors, that compare data or output data are also subjected to failures. Testing of complex high-performance VLSI chips means that a lot of different test vectors must be offered to the chip at the right time. Normally, the test response is compared with the "comparison data", most of which is generated during the simulation of the silicon at the verification phase of the design. To reduce the number of test pins and test time, large parts of the chip are tested via scan chains.

A reduction of the number of test vectors is often achieved by the implementation of Multiple Input Signature Registers (MISRs), which allow compression of data over a number of clock cycles. The final data is then scanned out. Because such tests are not functional tests, they may not yet have been simulated thoroughly during the design phase, leading to incorrect test pattern generation or incorrect comparison data. Moreover, when a bit failure occurs in compressed test data (signature), it is very difficult to locate the cause of the failure. This requires a lot of simulation. Data compression techniques during testing must only be used if other techniques are not satisfactory.

Other causes of test errors are timing errors. Sometimes, the switch from a functional test to a scan test or vice versa may take more time on the chip for the multiplexers to adopt the new state. Waits must then be included in the test programs to properly test the chip. Even set-up and hold times for input pins or the amount of load that the tester offers to a chip output pin must be thoroughly verified. In some cases, even the tester hardware might show problems.

An important, but not yet discussed, source of initial test failures is the probe card, which is used in the initial test phase during failure analysis on the wafer instead of on packaged dies. In such a test environment, limited ground bounce can only be achieved by taking several measures. These measures are all related to preventing or limiting current slew rates (dI/dt). Placing decoupling capacitances close to the supply pads is one measure. Another measure is to prevent large (ground) current loops. This can be achieved by using star grounds instead of serial grounds, see figure 10.22.

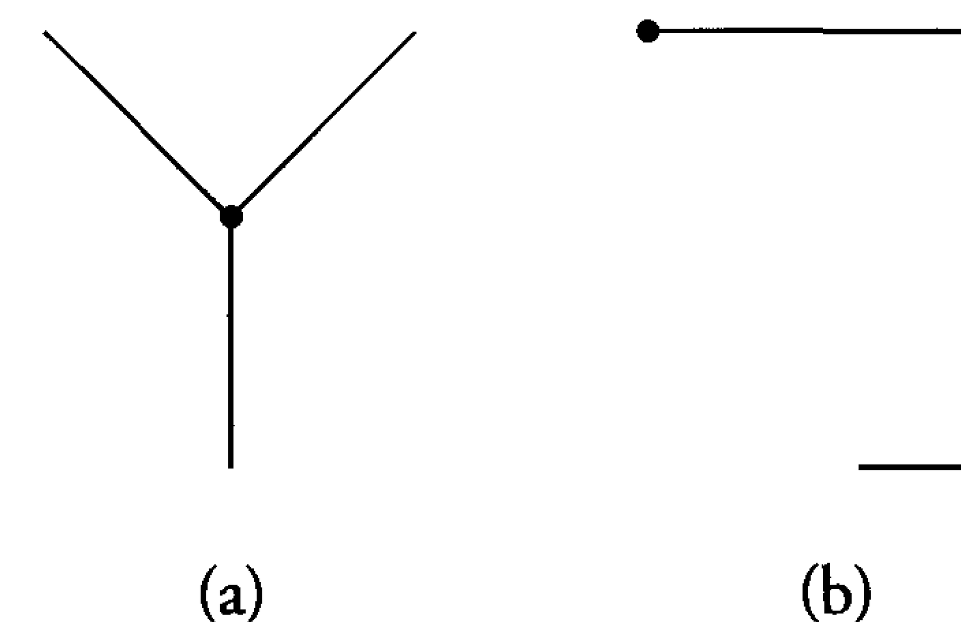


Figure 10.22: *Limiting large (ground) current loops by using (a) star grounds instead of (b) serial grounds*

Especially outputs can generate large current slew rates. The measurement of V_{o_L} and V_{o_H}, for instance, will often be done sequentially instead of testing simultaneously for all outputs. In conclusion, failures may arise during the development of the tests, during the development of the test boards and during the testing itself. Passing these test phases carefully can save a lot of time and frustration during the evaluation of first silicon.

10.6.2 Problems caused by marginal or out-of-specification processing

Each batch of wafers is processed under different environmental conditions: dust, temperature, humidity, implanter energy, etching time and doping levels, etc. This means that dies from different batches may show different electrical behaviour. The number of dust particles, for example, is one of the dominating factors that determines the yield, see section 10.3. In the following, we describe the influence of the most important technology parameters on the electrical behaviour of the chip.

Gate oxide thickness

The gate oxide thickness is the smallest dimension in the manufacture of MOS devices. It controls the gain factor β and the threshold voltage V_T, and it can also affect the IC's reliability.

When the gate oxide is thin, β will be high and an increased current capability of the transistors will be the result. In some circuit blocks, especially in memories, signals have to arrive in a certain sequence and they therefore propagate through different delay paths. However, when transistors become faster, the difference in delay paths may change, or may even become negative. This may cause a race, resulting in malfunctioning of the circuit.

Too thin gate oxide may also lead to pinholes, which are oxide imperfections where the oxide thickness is locally reduced. Sometimes, the oxide thickness at such a pinhole may be so thin that the voltage across it might cause carriers to tunnel through this oxide. The resulting leakage current increases slowly over time and eventually, as a result of this oxide breakdown mechanism, the chip no longer functions correctly. This process might take an hour, a week, a month or even a year. The sooner it is detected, the better. However, detection after shipping the device to customers will be disastrous. Therefore, a bad gate oxide reduces reliability and can often be detected by means of I_{ddq} testing.

Polysilicon width

The distance between the source and drain of a transistor (called the channel length) is determined by the width of the polysilicon, forming the gate of the transistor. The wider the polysilicon, the larger the transistor channel lengths will be and the slower the transistor becomes. Signals that propagate through a combination of metal tracks and transistors will show different timing diagrams when polysilicon widths are wider than expected. This may lead to timing problems as a result of slowly operating circuits. On the other hand, narrow polysilicon leads to fast transistors. This may again result in timing problems such as races.

Threshold voltage

A change in threshold voltage can have different effects on the electrical behaviour of the chip.

A high threshold voltage, caused by a different channel dope, a thicker gate oxide or a larger body factor (K-factor) results in slower operation of the transistors. Especially a high body effect may lead to problems in pass-transistor logic and latches that use pass transistors. This high body factor may cause these circuits to operate significantly slower.

In contrast, a low threshold voltage results in somewhat faster circuits. Sub-threshold currents, which increase by a factor of about 12 to 15 every 100 mV decrease of the threshold voltage, may cause larger standby currents. This is an especially important consideration in low-power battery-driven applications.

For TTL compatible inputs of MOS ICs, the threshold voltage is also very important. The transistor current drive capability is proportional to $(V_{gs} - V_T)^2$. Because the TTL high level is specified as 2 V, a change of 100 mV in the threshold voltage (e.g. from 0.5 V to 0.4 V) causes a change in current of more than 10%. Also, with a change in V_T, the switching levels of TTL input circuits may vary dramatically, such that input timing may also become critical. At lower threshold voltages, the low noise margin at the input may also become critical.

Substrate and/or n-well dope

All together the n^+ diffusions of an nMOS transistor, the p^- substrate, the p^+ diffusions of the pMOS transistors and the n-well form parasitic thyristors.

When the p^- substrate is pulled to more than a junction voltage ($\approx$ 0.8 V) above the n^+ diffusion, such a thyristor might switch on, see also section 9.2. Because of the positive feedback in such a thyristor, it operates like a latch and the current may increase to relatively large values. This effect is called latch-up and can only be eliminated when the power supply is switched off.

Low substrate dope allows the thyristor to switch on much earlier and makes the circuit more susceptible to latch-up. The doping levels of substrate and n-well also determine the threshold voltages of the nMOS and pMOS transistors, respectively, as well as the thickness of the depletion layers across their source and drain junctions. The latter, in turn, determines the parasitic junction capacitances.

10.6.3 Problems caused by marginal design

Currently, verification software for integrated circuits has evolved to mature tools that are part of every design flow. Especially the verification on Register Transfer Level (RTL) and logic level (gate level) offers the potential of designing chips in which no logic error can occur. These tools almost guarantee that everything on the chip is connected correctly according to the specification. It is therefore important to first verify the specification, either by simulation or by emulation. Sometimes, in an application, the chip does not perform the function it was meant to execute. In many cases, it later appeared that the specification was insufficiently verified.

A hardware failure in very complex (programmable) chips can sometimes only be detected during very dedicated application tests. The number of different applications (and thus programs) of such chips is almost unlimited and extremely hard to simulate within an acceptable time.

Currently, most ASICs are designed in a mature process via a mature design flow and run at medium clock frequencies. First-time-right ASICs therefore should be the rule rather than the exception. However, modern technologies (0.35 μm CMOS and below) offer small feature sizes and thus the ability to integrate millions of transistor on one single chip. This, combined with the trend of increasing chip area, challenges the designer with many potential electronic problems that are not yet (or only partly) dealt with by the tools. Chapter 9 focuses on the underlying physical effects and on the measures that a designer can take to maintain the IC's reliability and signal integrity at a sufficiently high level.

10.7 First-silicon debug and failure analysis

10.7.1 Introduction

Current VLSI chips may contain millions of transistors, but only several hundreds of I/O pins. This means that only a few logic blocks have direct access to output pins. These blocks can be measured at full functional speed.

Without a direct access to the output pins, the other blocks must be accessed through a scan chain and tested as such. In many cases, these scan tests run at a lower speed. This might lead to problems that show up only when the blocks are used in the real application because all the circuit is then running at full speed.

Logical (design) errors are easy to locate, both in scan test or in full functional test. However, timing errors are much more complex to identify.

When failures show up during the engineering phase of an IC, it is important to know their source: whether it is logical, short circuit, latch-up or timing, etc. I_{ddq} testing is a means to quickly detect leakage currents and floating nodes, etc.

For circuits that can be tested at full functional speed, Shmoo plots can be drawn to find the source of the failure. Afterwards, different failure analysis techniques can be applied to locate the failure: picoprobing, liquid crystal, light emission and electron beam (E-beam). There are several other techniques that support these analysis tools and allow a quick repair of only a few samples.

10.7.2 I_{ddq} testing

When a synchronous chip has been completely designed in static CMOS, hardly any current should flow when the clock is switched off. The only current that flows is the sub-threshold and other small leakage currents. However, in some cases, local higher-amplitude currents can flow. The cause of such currents is discussed in section 10.2.

With I_{ddq} testing, all currents that reach a certain level can be easily and quickly detected. I_{ddq} measurements are done after the chip has been brought to different states. Well-chosen I_{ddq} test vectors (may be normal - AMSAL - test vectors) can give a good overall coverage. A coverage of 98% with only ten test vectors can sometimes be achieved. Because the chip is put through a sequence of functional and I_{ddq} test

modes alternately, large current peaks are interchanged with quiescent currents of some nano-amperes. This requires large dynamics of the tester current meters. Usually, special I_{ddq} monitors have to be developed and put on the load board to support a proper I_{ddq} current sensing. I_{ddq} testing is therefore a good means of locating certain defects or unusual behaviour which cause increased current levels during steady state. More on I_{ddq} testing can be found in section 10.2.

10.7.3 Diagnosis via Shmoo plots

When a complete chip or part of a chip can be functionally tested, and an insight about operating margins with respect to the specification is required, then a Shmoo plot can be measured and plotted.

A Shmoo plot shows the operating area of the software, the tester, the interface between tester and chip, and the chip itself, with respect to different parameters. When a Shmoo plot is not according to expectation (specification), the failure does not necessarily need to be in the IC design. It can also be in the technology, tester software or interface or the tester itself.

Figure 10.23 shows a Shmoo plot of V_{dd} versus frequency:

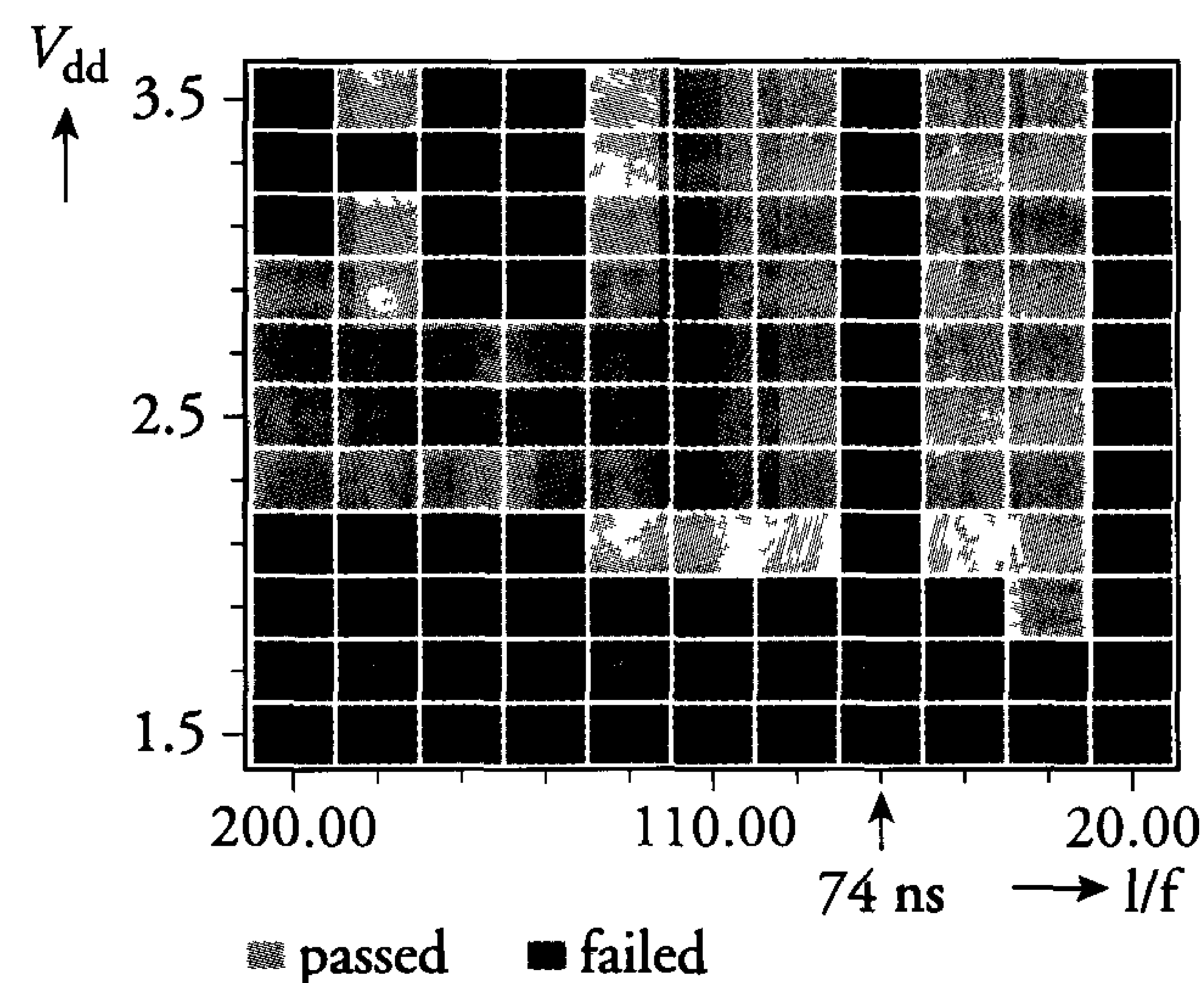


Figure 10.23: *Example of* V_{dd} *versus frequency Shmoo plot*

A Shmoo plot, which shows the operating area of a chip is, in fact, a quality measure. It shows whether the chip is marginal with respect to its specification. Measurements of Shmoo plots can be repeated at different temperatures to see how the margins shift.

The grey areas in figure 10.23 are areas in which the chip operates correctly. It is very peculiar that the chip does not function at 74 ns period time, independent of the supply voltage. It took a couple of days before it appeared that the tester had a problem in generating test vectors at that frequency.

When the environment has proven to be correct, then, with too small operating areas of the chip, several different Shmoo plots must be measured to find dependencies: supply voltage, frequency, set-up time, temperature and I/O levels, etc.:

- If delay paths between flip-flops are too long:
  → frequency versus supply voltage Shmoo plot: lower frequency → better operation and higher voltage → faster circuits
  Conclusion: ⇒ use frequency versus supply voltage Shmoo plot at a fixed temperature.

- if races occur:
  → supply voltage versus temperature Shmoo plot: higher voltage → faster circuits and higher temperature → slower circuits
  Conclusion: use supply voltage versus temperature Shmoo plot at a fixed frequency.

- In case of wrong sizes of circuits (latches and buffers, etc.), these are supply dependent

- A Shmoo plot diagnosis may take a lot of time. Once a diagnosis has been made, it must always be verified by other techniques (such as probing). This is shown by the following example:

EXAMPLE:
A certain signal processor contained two separated V_{dd} supply connections: V_{dd_1} and V_{dd_2}. Figure 10.24 shows the Shmoo plot of the operating area of the memory on that chip:

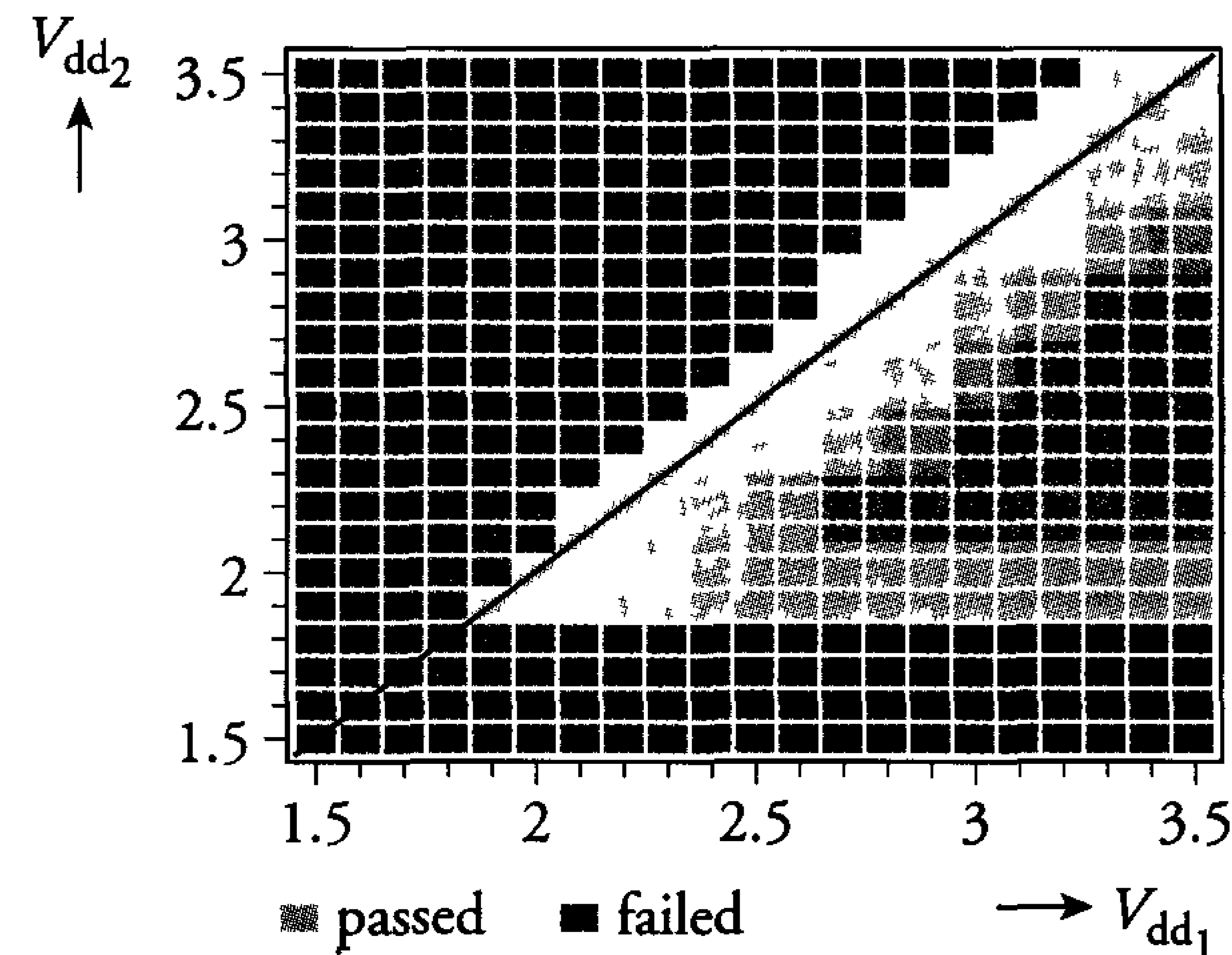


Figure 10.24: *Example of a very critical Shmoo plot*

Although these supplies, V_{dd_1} and V_{dd_2}, should be externally connected together, the operating margin is only very limited and, with a small change of the process, the chip would no longer function. At first, the input registers of the memory were suspected. Because the ϕ and $\overline{\phi}$ clocks are generated by a clock buffer which is supplied by V_{dd_1} and the latches (figure 10.25) of the input register were supplied by V_{dd_2}, these latches could be the cause of the problem.

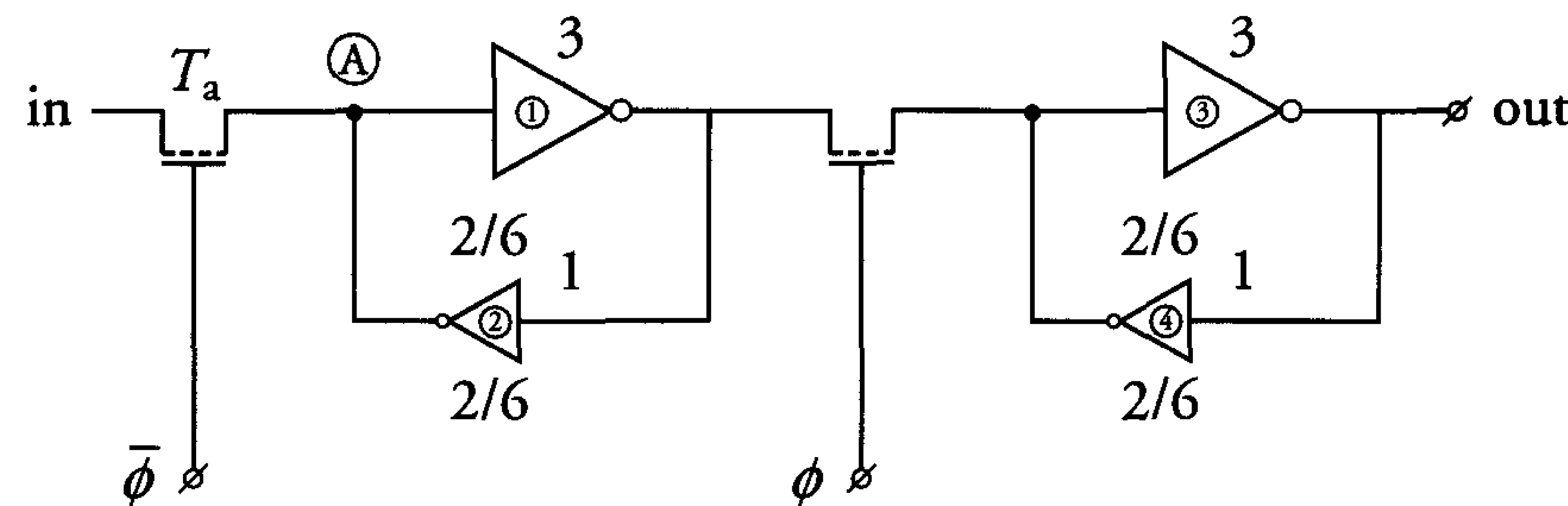


Figure 10.25: *Circuit for potential cause of the problem in the video signal processor*

Inverters 1 and 3 have a switching point equal to about $V_{dd_2}/2$, because the pMOS transistor width is three times the nMOS width (so $\beta_n \approx \beta_p$). Clocks ϕ and $\overline{\phi}$ are supplied via power supply V_{dd_1}. When $\overline{\phi}$ is high (V_{dd_1}), the voltage on node A will not be higher than: $V_{dd_1} - V_{T_a}$.

Because of the back-bias effect, V_{T_a} will be relatively high. Therefore, if $V_{dd_1} - V_{T_a} < V_{dd_2}/2$, the flip-flop will fail to switch to a logic "1". The results of a circuit simulation are shown in figure 10.26:

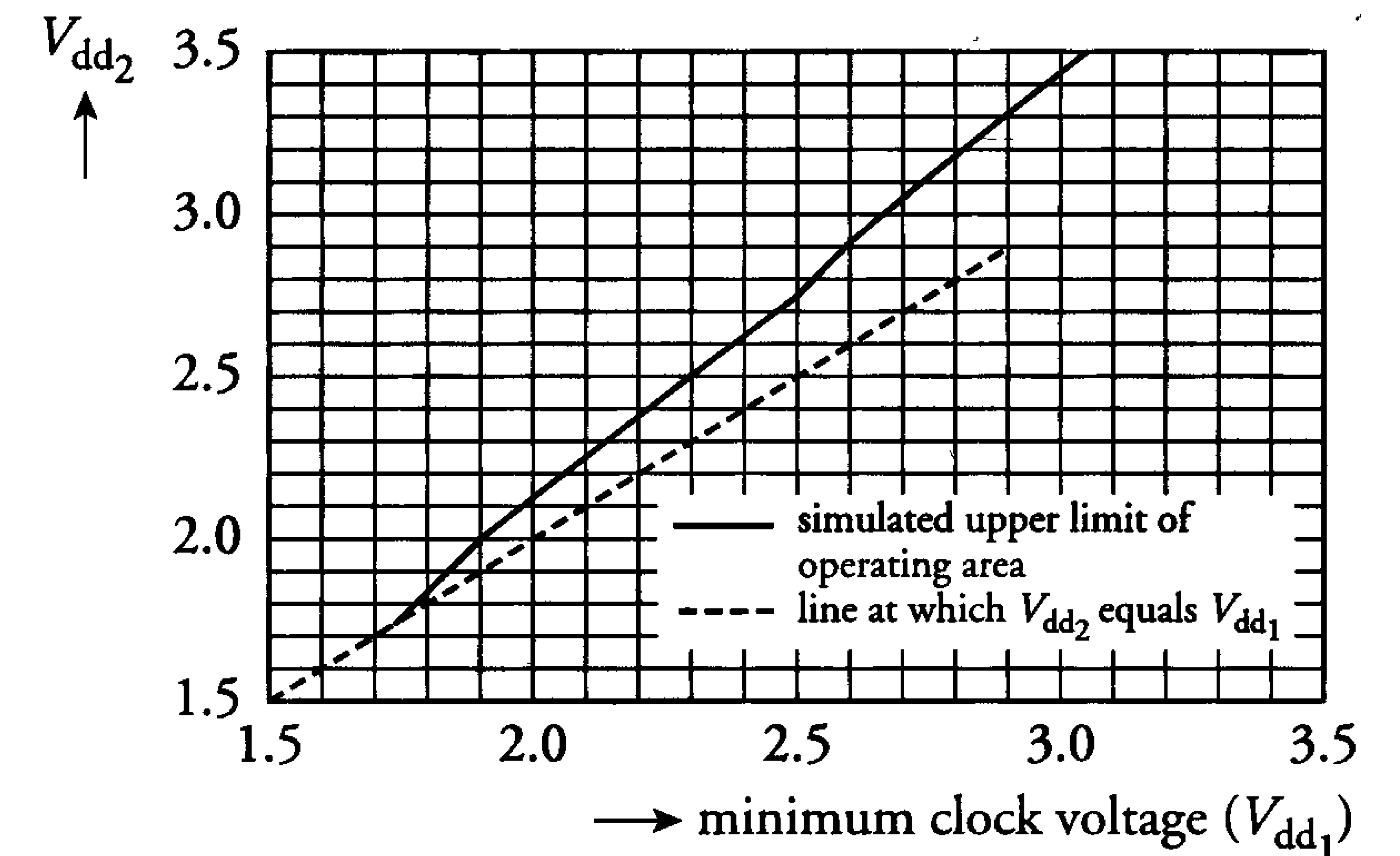


Figure 10.26: *Circuit simulation of the operating area of the latch of figure 10.25*

Below the solid line, the circuit operates correctly; above the line, it does not. If we compare this with the Shmoo plot of figure 10.24, we see the similarity between them, and one would believe that the flip-flop is the real cause of the problem.

However, before changing the flip-flop design, the outputs of the flip-flops were probed to check the diagnosis. These flip-flops happened to operate much better than simulated and thus the real cause of the problem had to be found elsewhere.

This is discussed in the following subsection on picoprobing.

10.7.4 Diagnosis via picoprobing

Picoprobing is a method that allows us to probe any node that is available at the top level metal when there is still no passivation layer (scratch

protection) on the wafer, or when this layer has been removed locally (either by etching, by laser cutting or by Focused Ion Beam: FIB).

A picoprobe 10.28 consists of a needle as thick as a hair and with a very thin tip, less than several tens of a micron. This needle is connected to the input of a FET to reduce the capacitance. Values of 10 to 1000 fF for such a FET probe are available and are so low that they can be used almost everywhere within a digital IC without affecting the probed signal itself. It is also quite possible to probe signal lines in the top level metal, which have minimum widths and are separated at minimum spacings.

This technique is a commonly used and reliable method for analysising incorrectly operating VLSI chips. However, with the advent of multi-level metal technologies, it is becoming increasingly difficult to probe a signal that is only available in the lowest metal layer(s). During the design phase, additional small probe pads in the upper metal layer could be placed at the critical nodes. Another way to cope with this problem is to adapt the design style to design for debug, see section 10.7.10.

In the previous example, picoprobes were also used to try to further locate the failure. After a while, the real cause of the failure was found. Figure 10.27 shows the corresponding schematics:

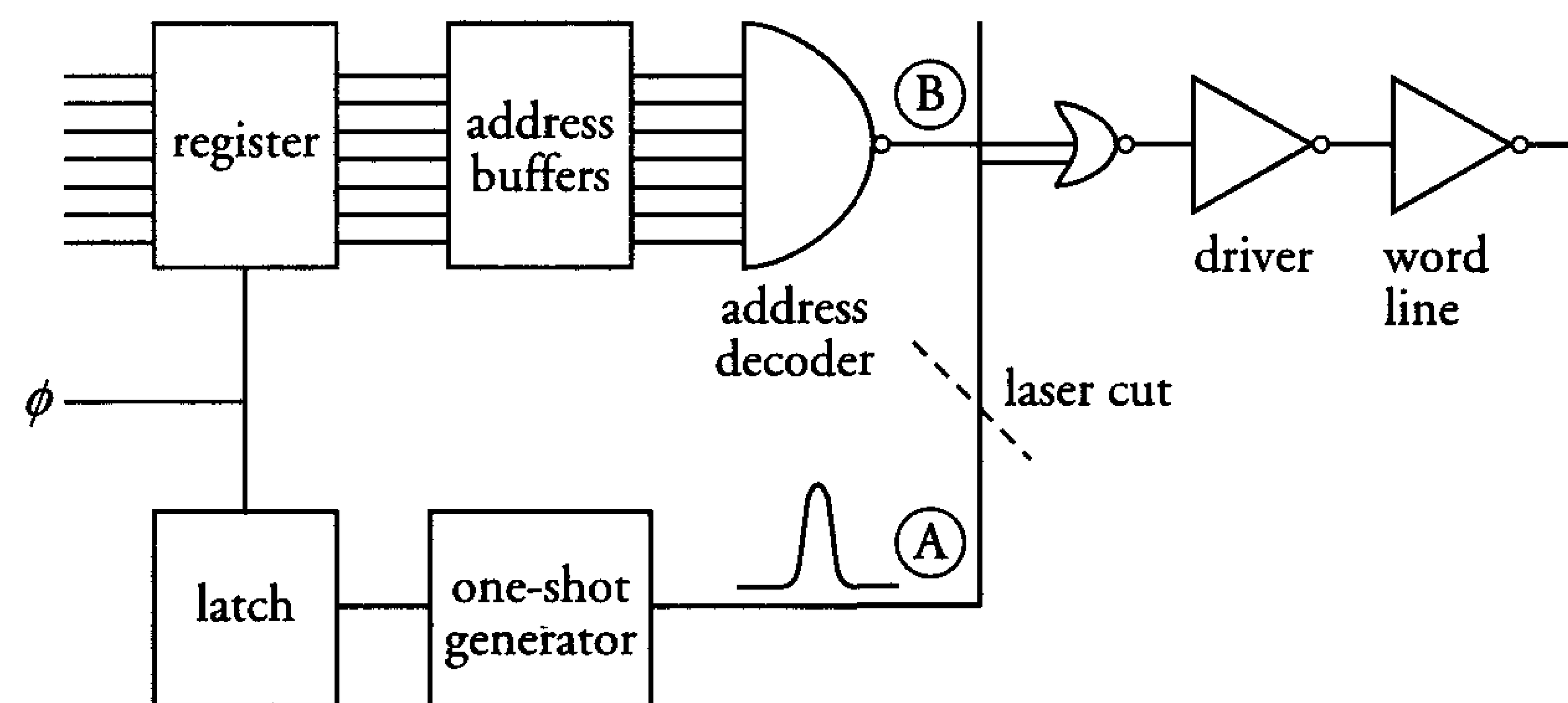


Figure 10.27: *Schematic of the real cause of failure*

Figure 10.28: *Picoprobes can be used to measure a chip's internal signals (Photo: PHILIPS)*

The one-shot pulse on node A came 100 ps too early: the old address at node B was clocked in and this resulted in reading the wrong word from the memory. By cutting track A with a laser, probing the one-shot signal right before the cut and forcing it back via a pulse generator with variable delay right after the cut, the correct pulse could be found and a new Shmoo plot was measured. Figure 10.29 shows the result.

Thus, probing identified the exact location of the failure and also a way to solve the problem. A redesign (a one-mask change only) was made and the next batch showed correctly operating devices.

Besides picoprobing, there are several other techniques to accommodate failure analysis. The techniques discussed in the following paragraphs can be very well combined with a VLSI tester. This combination can, for example, generate I_{ddq}-vectors and easily detect hot spots (small areas with more dissipation, e.g. from leakage currents).

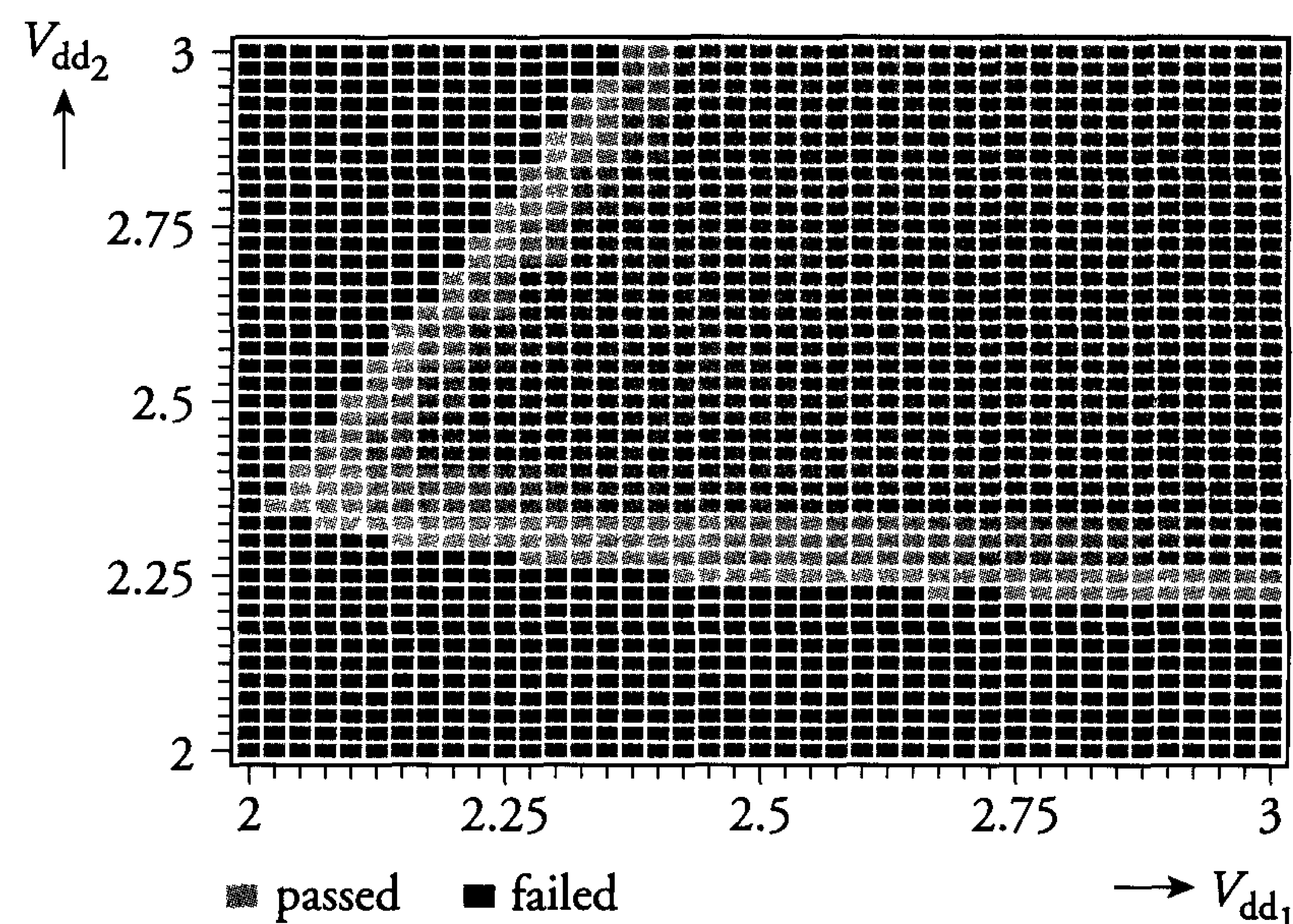


Figure 10.29: *Shmoo plot after correcting the one-shot pulse by forcing it via picoprobes*

10.7.5 Diagnosis with liquid crystal techniques

Many of the failures on an IC lead to increased local currents, causing additional power dissipation and local hot spots, as already discussed. With liquid crystal techniques, it is possible to locate and visualise these hot spots. Even clock and output buffers, which dissipate more than logic circuits, can be visualised as well. Nematic (a certain type of ordering) liquid crystals with a first-order phase transition at a clearing point T_c is most commonly used for this purpose.

Below this temperature, the liquid crystal is in the ordered state. Now, polarised light is reflected through an aligned analyser in the microscope: the low temperature state is visible as the bright area. Above the clearing point, it is isotropic and hardly any light can pass the analyser. Thus, dark spots or areas in the liquid crystal designate hot spots. A commonly-used liquid crystal is from Hoffmann LaRoche and is called ROCE1540. At room temperature, it is solid and its clearing point is at 56.5 °C. It is deposited in a saturated solution of acetone, which evaporates in one minute.

Disadvantages of this technique are the poor temperature and spatial resolution, and the absence of a temperature-mapping capability. A new option is the possibility of producing thermal maps (isothermal pictures) by generating a sequence of images. This also gives a good insight into the overall chip power distribution.

Figure 10.30(a) shows a photograph of a hot spot visualised by liquid crystal. Figure 10.30(b) shows the failure that caused the hot spot.

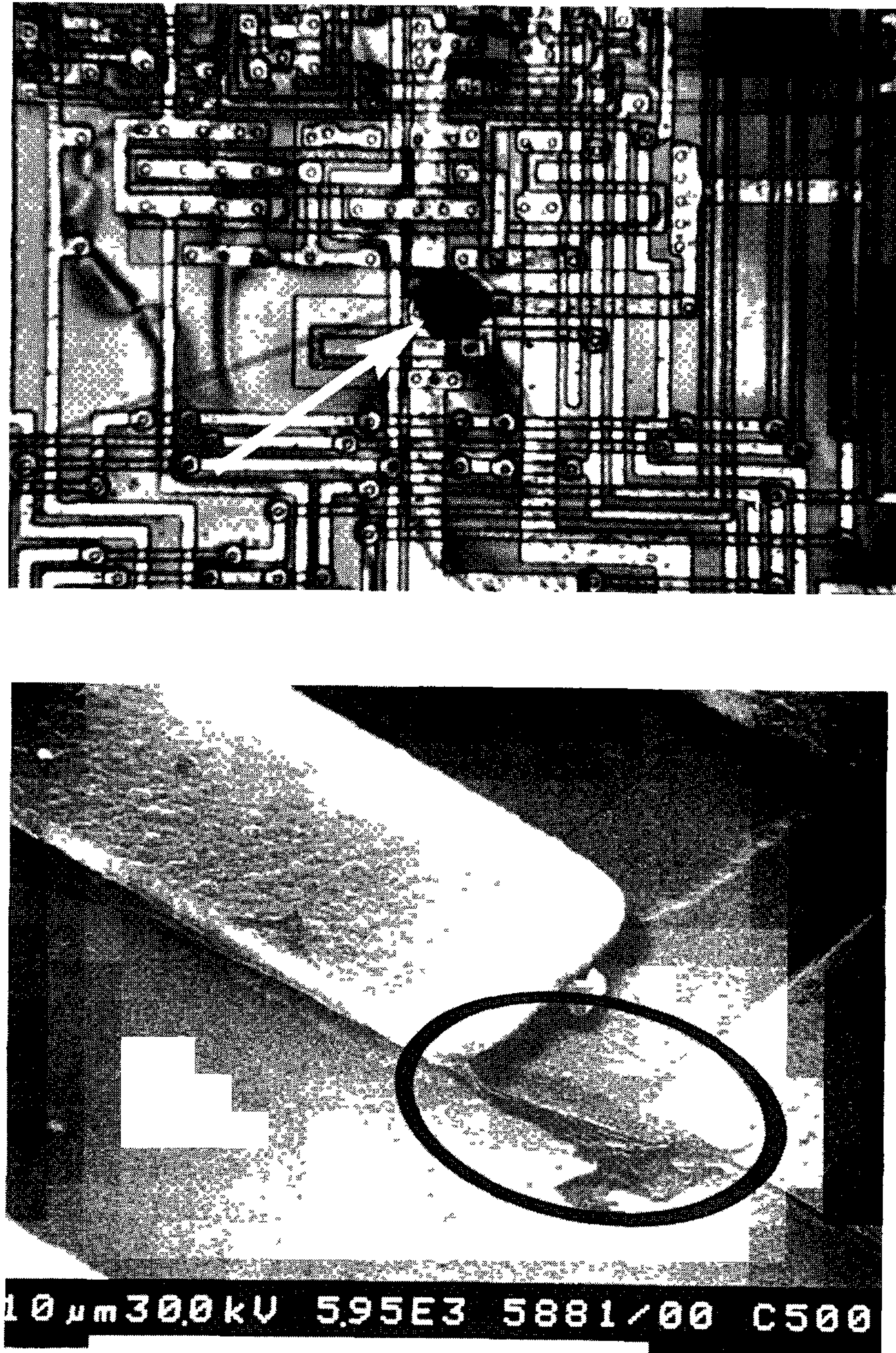


Figure 10.30: *Example of hot spot visualised by the liquid crystal technique (Photo: PHILIPS)*

10.7.6 Diagnosis by photon emission microscopy (PEM)

When charge carriers decay to a lower state of energy, the energy surplus is converted to photon emission (PE). This occurs, for instance, when electrons are accelerated in the transistor channel to excessive velocities, when carriers cross a potential barrier, or during breakdown, resulting in an avalanche of carriers [32]. It is therefore a good tool in identifying hot electrons. The light that is given off by operating ICs is collected and used for imaging. Figure 10.31 shows several hot spots on a CMOS chip layout:

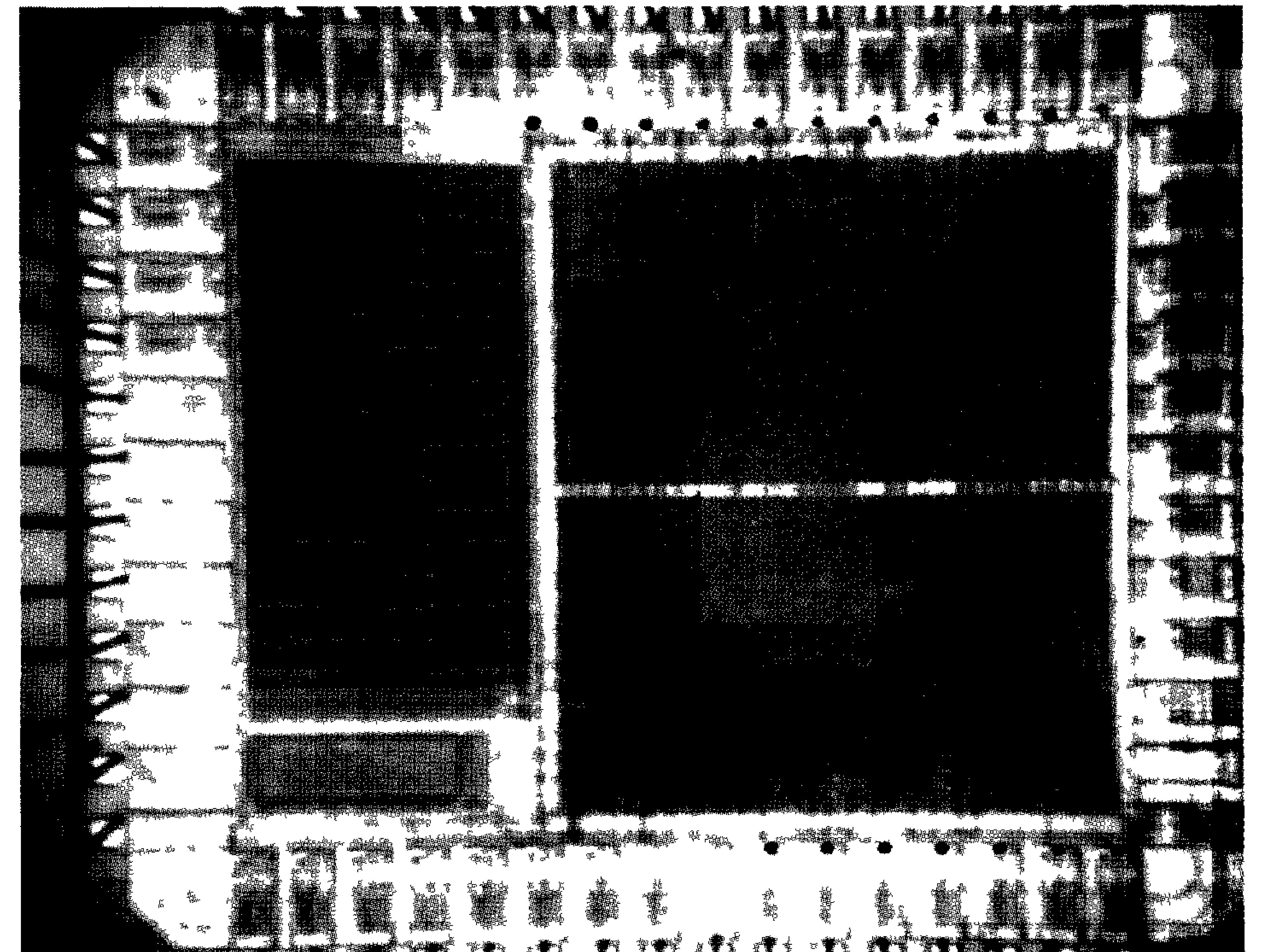


Figure 10.31: *CMOS chip showing several hot spots via the PEM technique (Photo: PHILIPS)*

Because current image intensifiers have gains up to 1,000,000, they can detect even very low-level PE. Currents below a micro-ampere can be visualised at spatial resolutions less than 1 μm. In this way, latch-up,

gate oxide defects, saturated MOS transistors, degraded (avalanching) junctions and unwanted forward-biased junctions can be detected.

The emitted photons can have energies in both the visible and near infra-red (IR) wavelengths. Therefore, IR PE can also be used for die back side analysis. With highly sensitive cameras (cooled back-illuminated CCD cameras), this greatly reduces image capture time and prevents optical obstruction by multi-level metal layers and flip-chip packaging.

10.7.7 Diagnosis by electron beam techniques

When an electron beam is directed onto a certain surface, that surface emits secondary electrons. The number of secondary electrons is then dependent on the surface potential. This allows the visualisation of the internal voltages on an IC. Ground lines basically appear bright in the image, while positive voltage lines look dark (high absorption/low emission).

By using magnetic lenses, the primary E-beam of a SEM (Scanning Electron-beam Microscope) can be focused and scanned across an IC. To reduce the charging destruction of parts of the chip, the E-beam energy is limited between 0.6 and 3 KeV. The secondary electrons are detected and the amplitude of the resulting signal modulates the intensity of a second E-beam in a cathode ray tube. There are several modes of operation.

In the *static voltage contrast mode*, part of the IC can be displayed with supply lines (at V_{dd}) absorbing and ground lines (at V_{ss}) reflecting most of the secondary electrons. This is suitable for the measurement of low frequency (or static) signals.

In the *stroboscopic voltage contrast mode*, AC signals can be measured. Here, the scan frequency is synchronous with the signal frequency on the chip but an order of magnitude lower, such that the E-beam probes the same signal phase every time it reaches a given spot. This results in a static picture.

In the *waveform measurement mode*, the beam is fixed on a certain node but now with a frequency that is much higher than the signal frequency. In this way, the signal on the node can be displayed (on an oscilloscope).

10.7.8 Alternative failure analysis techniques

Recent developments in failure analysis include *CIVA* (Charge-Induced Voltage Alteration) and *LIVA* (Light-Induced Voltage Alteration). Both are relatively simple techniques for detecting and locating open circuits and junction, contact and via defects [33].

Both techniques use the same electrical approach, but different probes. The CIVA technique uses an electron beam, which extends its usage through a multiple of dielectric and interconnection layers, just by increasing the electron energy. If LIVA, which uses a laser beam (photons), is equipped with a visible laser, it is used for front side analysis.

An IR laser LIVA analysis is performed from the rear. This technique can be used in the failure analysis at the transistor level, because they are often completely covered by one or more metal layers that shield the front side.

When the E-beam (in CIVA) or the optical beam (in LIVA) is scanned across the chip, the secondary carriers are generated at the pn junctions. These carriers create a current during recombination, as the beam passes, causing voltage changes across the chip. The image compares the voltage change to the position of the beam. The variations are used for localisation.

Advanced E-beam test systems provide software tools that allow a short time for locating and analysing faults. However, they face the same problem as picoprobes do from the multi-level metallisation. The shielding effect of these layers renders E-beam analysis ineffective on fully processed wafers.

Unless dedicated etching techniques are used, only the top-level metal can be imaged. Lower level metals can be imaged only when they are not covered by others. The lower the layer, the less image resolution will be achieved.

In the design phase of these chips, extra pads in the uppermost metal layer could be added to the most critical signal nodes. In the engineering phase, these additional pads can be created by a FIB system, which can connect such a pad to almost any point of interest. Flip-chip packaging makes E-beam failure analysis useless without disassembly.

10.7.9 Repair

Once the diagnosis has been made and verified, the chip can sometimes be repaired directly by making or breaking techniques. This might lead to fewer redesigns.

By using a laser, it is possible to fuse an interconnection (for instance to isolate circuits). A laser beam, whose spot can be positioned with less than 0.1 μm accuracy, is projected onto a certain interconnection material of which the temperature may increase to above its evaporation point, depending on the intensity of the laser beam and the absorption capability of the material.

By means of laser-induced liquid-phase metal deposition, interconnections on top of the chip can also be made during the IC evaluation phase. First, holes are made in the scratch protection, through which the connections to the underlying metals will be made. Next, a palladium (or other metal) track is grown with the aid of a laser. The minimum line width to be connected is more dependent on the sizes of holes in the scratch protection than on the minimum width of the grown tracks. A disadvantage of laser systems is that their resolution is limited.

Focused Ion Beam (FIB) systems show better resolution and, for cutting conductors, spatial resolutions of less than 0.1 μm have already been demonstrated. Because holes can be made with high accuracy, even connections between different metal layers can be made, providing the capability to rewire ICs directly on the chip. Simple FIBs are coupled to E-Beam systems, to create probe points. The deposition of the conductive material on top of the scratch protection is easy but time consuming. A modern FIB system consists of complex and expensive equipment, which is capable of removing and depositing material (metal and dielectrics) and making smooth cross-sections for SEM or TEM analysis. It is sometimes combined with a SEM column for high-resolution imaging with SIMS (Secondary Ion Mass Spectroscopy) equipment. To allow faster material removal at lower beam intensities, advanced FIB systems use *gas assisted etching*. With this technique, holes can be etched down to metal one in a five metal layer wafer. In this respect, holes with aspect ratios of 18:1 can be created [34]. Since FIB technology allows small holes to be accurately cut through the wafer, it may be a valuable tool in inspecting flip-chip packaging.

10.7.10 Design for Failure Analysis and Design for Debug

To meet the challenges posed by the IC roadmap (several hundreds of millions of transistors, with feature sizes around 0.1 μm and frequencies up to 1 GHz), different design approaches are needed to maintain a limited failure analysis and debug time.

Design for Failure Analysis includes design strategies to facilitate I_{ddq} testing, design-in of additional test points for probing (E-Beam or physical probing), as well as the addition of markers to support on-chip navigation during the use of FIB or optical microscopy equipment.

Design for Debug (fault observability) relies on software fault isolation. It will enable all flip-flops to be monitored and controlled at full functional speed. It requires additional on-chip hardware, which supports the debug software running separately from the digital tester.

10.8 Conclusions

The general requirement of a high fault coverage during the test of an IC is being challenged by an ever-increasing design complexity. Advanced test methods have been developed to maintain high fault coverages, both during IC testing and board testing. Additional hardware is included in the design to support these methods. This also reduces test time, which can be a relatively large contribution to the ultimate price of an IC.

Although packaging is not really a typical CMOS issue, the overview here shows the importance of choosing the right package. The temperature of the packaged die and the self-inductance of the package pins are important parameters which may dominate the performance and the operation integrity of a design.

Finally, the increased device complexity, combined with more levels of metal, reduces the fault observability. Advanced techniques have been developed for better observation, even from the back of the die. These techniques, used during first silicon debug, must be supported by a design-for-debug approach to allow both a rapid identification of the failure and the failure mechanism.

10.9 References

Testing and Debugging

[1] R.G. Bennetts,
'Design of Testable Logic Circuits',
Addison-Wesley, 1984

[2] G. Russel, I.L. Sayers,
'Advanced Simulation and Test Methodologies for VLSI design',
Van Nostrand Reinhold Int., 1989

[3] R. Rasjsuman,
'I_{ddq} Testing for CMOS VLSI',
Artech House Publishers, Boston 1995

[4] IEEE computer Society,
'IEEE Standard Test Access Port and Boundary-Scan Architecture - IEEE Std. 1149.1-1990',
IEEE, New York, 1990

[5] Harry Bleeker et al.,
'Boundary-Scan Test - A Practical Approach',
Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993

[6] Colin Maunder and Rodham E. Tullos,
'The Test Access Port and Boundary Scan Architecture',
IEEE Computer Society Press Tutorial, IEEE Computer Society, Los Alamitos, 1990, ISBN 0-8186-9070-4

[7] Frans Beenker et al.,
'Testability Concepts for Digital ICs - The Macro Test Approach',
Volume 3 of Frontiers in Electronics Testing, Kluwer Academic Publishers, Boston, 1995

[8] Erik Jan Marinissen and Maurice Lousberg,
'Macro Test: A Liberal Test Approach for Embedded Reusable Cores',
Paper 1.2 at 1st IEEE International Workshop on Testing Embedded Core-Based Systems, Washington D.C., November 5-6, 1997

[9] Erwin Trischler,
'Incomplete scan path with automatic test generation methodology',
Proc. IEEE International Test Conference, November 1980 (pp 153-162)

[10] Erik Jan Marinissen et al.,
'Silicon Debug of Systems-on-Chips',
Proc. of Design, Automation and Test in Europe Conference, Paris, France, February 23-26, 1998 (pp 632-633)

Yield

[11] S.M. Sze,
'VLSI technology',
McGraw Hill, New York, 1983

[12] S.M. Hu,
'On Yield projection for VLSI and Beyond Analysis of Yield Formulas',
IEEE Electron Devices Newsletter 69, 4-7, 1984

[13] K. Saito and E. Arai,
'Experimental Analysis and New Modeling of MOS LSI Yield associated with the number of Elements',
IEEE Journal Solid-State Circuits, SC-17, 28-33, 1982

[14] C.H. Stapper,
'Fact and Fiction in yield modeling',
Microelectronics Journal, Vol. 20, No. 1-2, 1989, pp 129-151

[15] G.A. Allan et al.,
'An Yield Improvement Technique for IC Layout Using Local Design Rules',
IEEE Trans. on Computer Aided Design, Vol. 11, No. 11, Nov. 1992

[16] W. Maly,
'Yield Simulation - A comparative Study',

[17] C. Kooperberg,
'Circuit Layout and Yield',
Journal of Solid-State Circuits, Vol. 23, No. 4, Aug. 1988

Packaging

[18] R. Tummala, E.J. Rymaszewski,
'Microelectronics Packaging Handbook',
Van Nostrand Reinhold, New York, 1989

[19] D. Seraphin, R. Laskym, C.Y. Li,
'Principles of Electronic Packaging',
McGraw-Hill, May 1989

[20] 'Surface Mounting and Interconnecting Chip Carrier Guideline',
IPC, Lincolnwood, Illinois

[21] Dataquest, 'Integrated Circuit Packaging-1997',
Product code: SEMI-WW-FR-9702

[22] Ho-Ming Tong,
'Micro electronics packaging: present and future ',
Materials Chemistry and Physics 40, 1995 pp 147-161

[23] R. Jensen, M. Mitchell and S. Palmquist,
'MCM, Designing for Reliability in Harsh Environments',
Advanced Packaging, January 1998

[24] G. Ginsberg and D. Schnorr,
'Multichip Modules & Related Technologies ',
McGraw-Hill, Inc. 1994

[25] 'Multichip-module market gets ready for take-off',
Electronic Components, December 1995, pp 54-64

[26] M. Töpper, J. Simon, H. Reichl,
'Redistribution Technology for Chip Scale Package using photosensitive BCB',
Future Fab International, 1997

[27] 'Integrated Circuit Packages',
Data Handbook IC26 Philips Semiconductors, 1998

[28] J. Lipman,
'New IC Packages really pack in the leads',
EDN September, 1997, pp 93-104

[29] 'Advanced packaging' magazine,
PennWell Publishing Corp.

[30] 'Step by Step', Advanced Packaging Magazine Online ,
May 11, 2000 *http://www.apmag.com/step-by-step/*

Quality and Reliability

[31] see [18]

Failure analysis

[32] K. de Kort,
'Techniques for Characterization and Failure Analysis of Integrated Circuits',
Analysis of Microelectronic Materials and Devices, John Wiley and Sons Ltd, 1991

[33] 'Finding fault with deep-submicron ICs',
IEEE Spectrum, October 1997, pp 39-50

[34] J. Orloff,
'Focussed Ion Beam Technology; Past and Future',
Future Fab International, issue 2, 1997, Vol.1, pp 239-244

10.10 Exercises

1. Why is Design for Testability an increasingly important design requirement?

2. A given CMOS manufacturing process has the following parameters at a certain point in time: $M = 0.75\,\text{cm}^{-2}$, $D_0 = 1\,\text{cm}^{-2}$.

   a) Express your opinion about this process.

   b) Calculate the expected yield for a chip with a die area of $80\,\text{mm}^2$.

   c) Three months later, a $80\,\text{mm}^2$ chip has a yield of 60.3 % and an $120\,\text{mm}^2$ chip can be produced with a yield of 49.4 %. Calculate M and D_0, assuming that the class of the clean room has not changed.

3. An IC dissipates 300 mW while its internal temperature is 32 °C. If the thermal resistance of its package is 30 °C/W, then what is the IC's ambient temperature?

4. What are the major differences between through-hole and SMD area array packages? Specify their respective advantages and disadvantages.

5. Explain the differences between test, debug and failure analysis of first silicon.

6. Why is Design for Debug a must for current and future ICs?

7. What is the drive for reducing the self-inductance (L) of the package pins?

8. What kind of tests are required to determine quality and reliability of packaged dies?

9. Explain what is meant by observability and discuss its trend with respect to future process generations.

10. In current and future processes, the transistors and lower metal layers are shielded by the upper ones. What could be done during the design phase to support failure analysis in this respect?

11. Explain how FIB supports failure analysis.

Chapter 11

Effects of scaling on MOS IC design and consequences for the roadmap

11.1 Introduction

If we continue to increase the complexity of ICs at the same pace as we have done since 1960, we will have reached a level of half a billion transistors per chip within a decade from now. Moreover, the clock period is 'expected' to be well below one nanosecond. Even if we do not believe these excessive numbers, design styles and methods have to be changed to fully exploit the potentials of IC complexity and speed, as predicted by the Semiconductor Industrial Association (SIA) roadmap (or International Technology Roadmap for Semiconductors (ITRS))[8].

This chapter discusses the consequences of the scaling process for deep-submicron IC design, with the focus on future trends of power, speed, reliability and signal integrity. The increasing dominance of physical effects that create interference and noise in VLSI designs requires more and more analogue measures to limit their influence. In the race towards a 1 Giga transistor (1 Gtor) heterogeneous System On a Chip (SOC), see figure 11.1, design styles have to be changed to make the design manageable (system design aspects) and to make a functional design (physical design aspects). Note the difference with figure 7.4, which only shows the system design aspects.

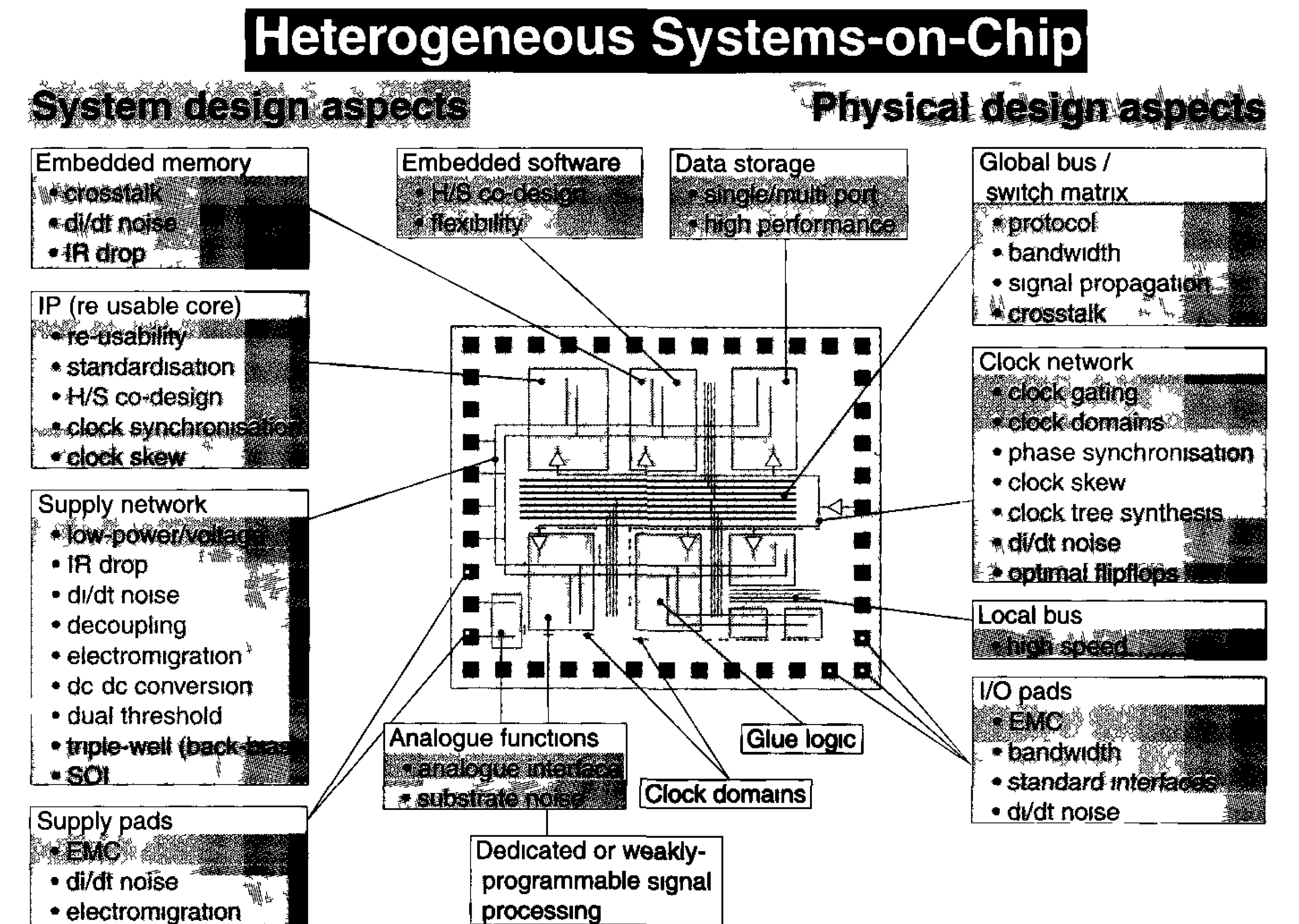


Figure 11.1: *Important aspects of a (heterogeneous) System On a Chip*

The complexity of such an SOC cannot be managed with traditional design concepts and requires:

- a platform with integrated hardware/software architecture and application development tools
- system level synthesis to improve design efficiency
- design reuse
- increased design resources per chip.

The first three items deal with system level design aspects, see chapter 7. The increased design resources, however, are not only required to manage the SOC design complexity, they are also needed to:

- develop testing, debugging and diagnosing concepts to enhance testability and observability when using IP cores
- manage clock synchronisation in different clock domains
- limit on-/off-chip noise and interference, and support EMC.

These last two items are particularly heavily influenced by the back-end part of the process (the metal interconnections). Previously, only analogue circuits were susceptible to these physical effects. In future process generations, these effects will dominate the SOC's performance and signal integrity, while some of these effects are already threatening the performance of complex VLSI chips. Future VLSI design therefore requires a more analogue approach. Design is no longer about switches and ones and zeros only, but also about resistors, capacitors, inductors, noise, interference and radiation.

Basically, a VLSI chip is just a bunch of transistors that perform a certain function by the way that they are interconnected. The next sections focus on the influence of scaling on the basic elements, the transistor and the interconnections, and the consequences for the overall performance, reliability and signal integrity of deep-submicron IC designs.

11.2 Transistor scaling effects

When scaling transistor sizes and bias voltages by a factor of s ($s \approx 0.7$), the transistor current scales by the same factor. To maintain performance, the threshold voltage is also required to scale with s. The threshold-dependent leakage current is an important factor that limits the pace of scaling. This sub-threshold leakage current can be estimated by the following, which means that this current increases by about a factor of 12 for every 100 mV decrease in V_T:

$$I_{\text{sub-threshold}}(\text{scaled}) = 12^{10(1-s)V_T} \cdot I_{\text{sub-threshold}}$$

For large SOCs, this background leakage current will be higher than a gate oxide short current, for example, which will dramatically limit the potentials of I_{ddq} testing. The eventual reduction of the threshold voltage requires alternative techniques, such as the Dual-V_T concept [1] and the Triple-well concept [2], to limit the sub-threshold power consumption during standby and test modes. Both concepts are discussed in chapter 8.

Another result of transistor scaling is the increased channel dope, caused by V_T correction and anti-punch-through implants. As a consequence, the thinner depletion layers cause higher parasitic junction capacitances.

In combination with gate oxide thickness (t_{ox}) scaling, the higher channel dope mirrors depletion charge into the gate (gate depletion),

which reduces the effective current control by the gate. Below a t_{ox} of about 2 nm [9], quantum-mechanical tunnelling of charge through the gate oxide may occur, resulting in additional standby currents and possibly a reliability problem. Finally, scaling of the channel length will increase mismatching of 'equal' transistors, as a result of increased spread of the number of dopants in the transistor channel. A minimum transistor in a 0.25 μm process contains about 1100 dopant atoms. In a 0.1 μm process, this number is only about 200. In both cases, the spread in this number causes a spread in V_T of: $3\sigma_{V_T} \approx 30$ mV, which relatively increases each new process generation. This V_T spread is additional to the process spread in V_T of about 60 to 90 mV.

11.3 Interconnection scaling effects

Scaling of widths and spacings has caused the metal interconnections to start dominating the IC's performance, reliability and signal integrity. The output load of a logic gate is equal to the total of the fan-in capacitances of its connecting gates and the total wire load of the interconnections. Table 11.1 shows the increase in the average ratio between wire load and fan-in, for large standard cell blocks (> 10 mm^2), caused by scaling. These numbers represent average values; for each individual chip, this ratio may be different from the table.

Table 11.1: Increasing interconnect dominance

Technology	Ratio: wire load/fan-in
0.35 μm	50/50
0.25 μm	58/42
0.18 μm	66/34
0.13 μm	75/25

Higher interconnection resistance leads to higher voltage drops and higher mutual capacitances lead to increased cross-talk, while the combination leads to larger signal propagation delays. At 1 GHz, the required rise and fall times should be less than 50 ps to perform some computational tasks within the available 1 ns time frame. Even on-chip wires then cause

interference with other modules. For such signal edges, line lengths of 3 mm and above become critical and require transmission line modelling. Figure 11.2 shows the *propagation delay* of an embedded metal track in different technologies.

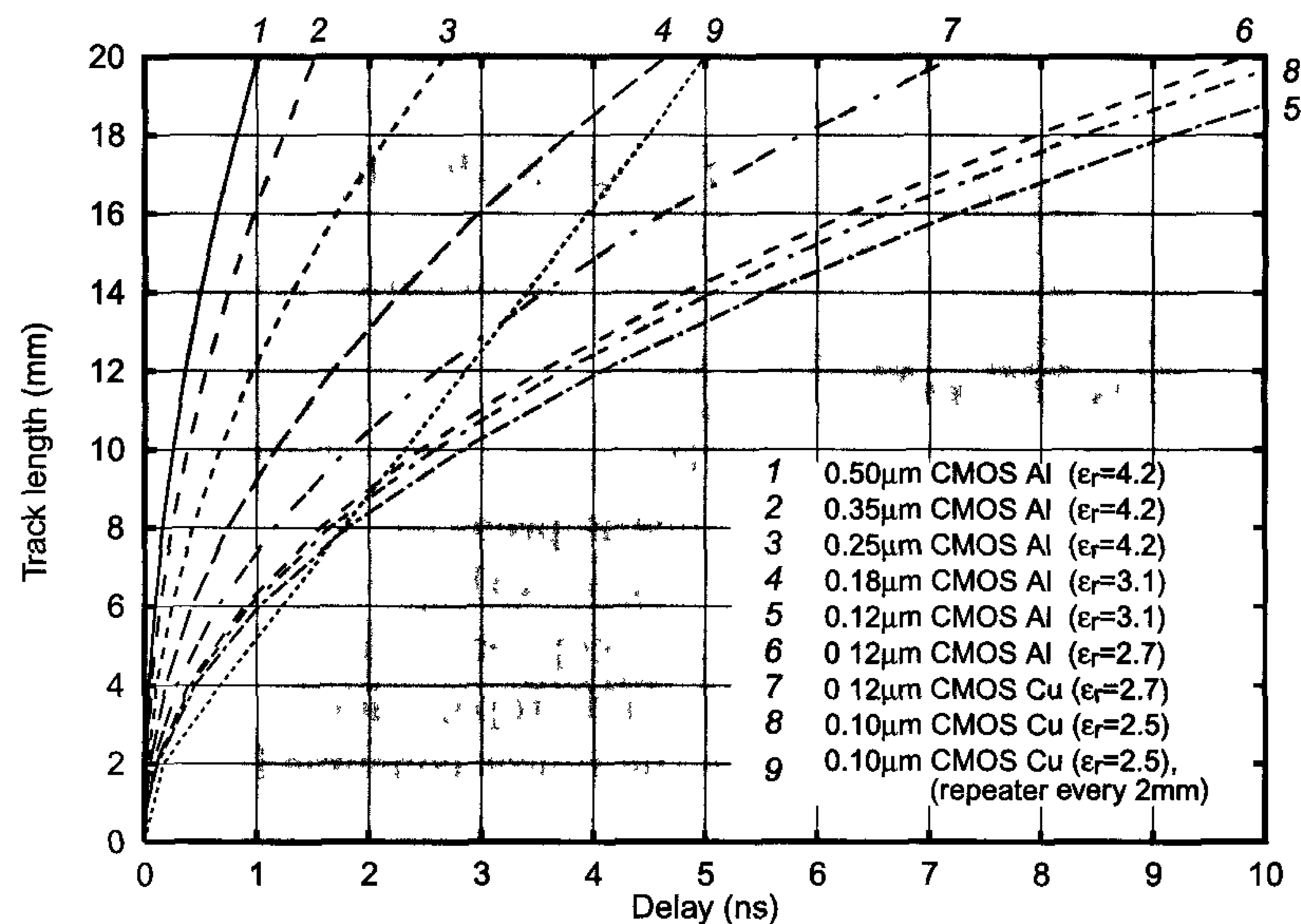


Figure 11.2: *Propagation delay of an embedded track in different technologies*

There are several approaches to reduce the negative effects of scaled interconnections. One is to reduce the capacitance, which is expressed as $C = \epsilon_0 \epsilon_r A/t_{ox}$. Current values for the dielectric coefficient ϵ_r are between 3 and 4. In the SIA roadmap for the year 2003, values around $\epsilon_r \approx 2$ are expected, leading to a capacitance reduction of a factor of 2. The second improvement we can make is to reduce the resistance. The sheet resistance of conventional aluminium alloys is around $3\,\mu\Omega$cm, while that of copper is about $1.8\,\mu\Omega$cm. However, the potentials of the reduced copper resistance cannot fully be exploited. Because copper diffuses through oxides, it cannot be deposited and etched like aluminium. By applying a damascene back-end flow, copper can be completely encapsulated within a barrier material, as shown in figure 11.3. The effective sheet resistance of copper wiring depends on the barrier material and will reach values close to $2.2\,\mu\Omega$cm. The advantage of this lower

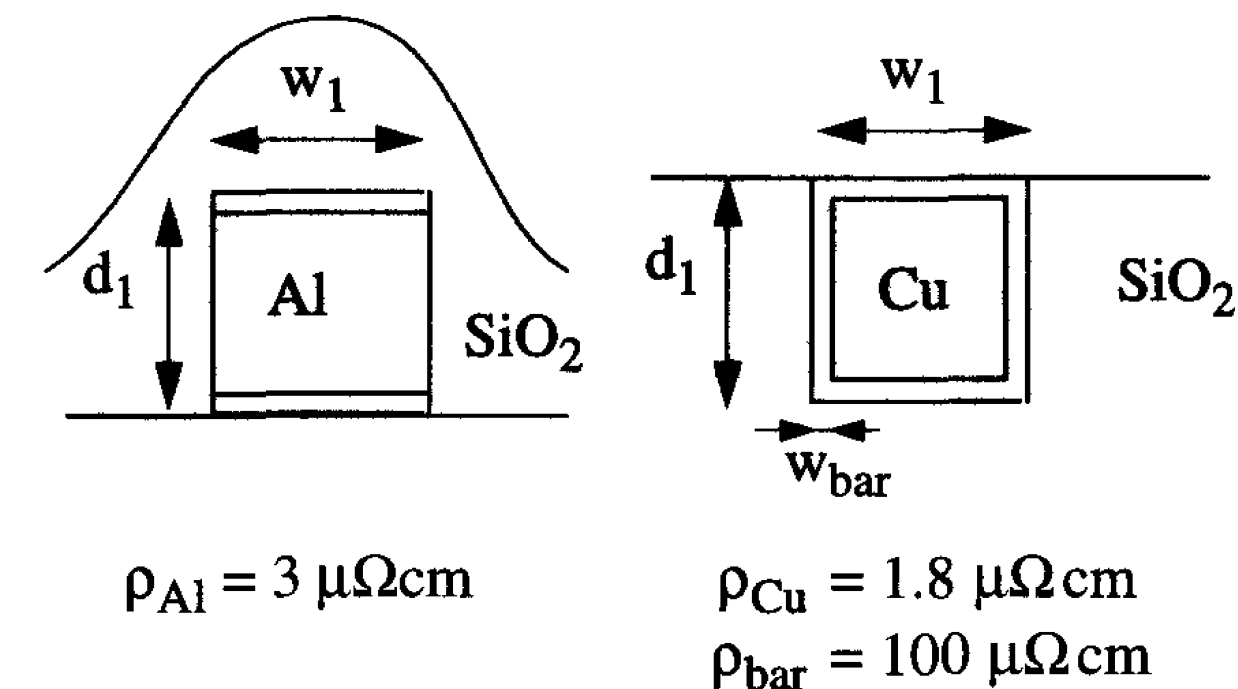
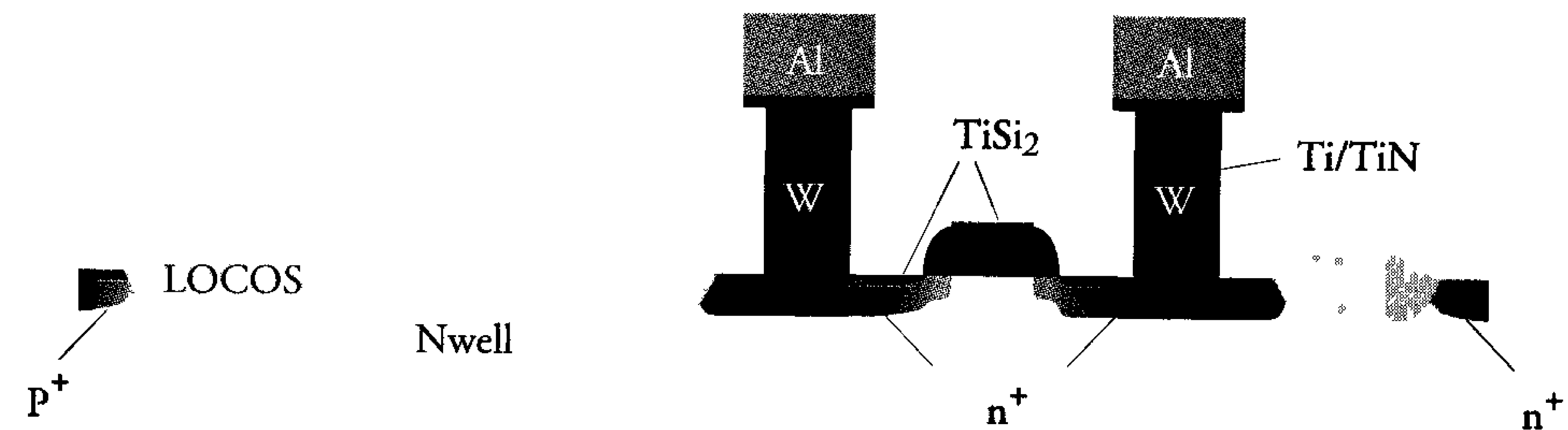


Figure 11.3: *Basic differences between the formation of aluminium and copper interconnections*

resistance will be used to reduce the track thickness, because a reduction of the mutual track capacitance is preferred to a reduction of the resistance. The total track capacitance is then reduced by a factor of about 1.25. Thus, the combination of copper with low-ϵ dielectrics may decrease the total track capacitance and propagation delay by a factor of about 2.5 at the maximum. Since the power consumed by the charging and discharging of a metal interconnection is proportional to the capacitance, this power will decrease by the same factor as well. The use of copper and low-ϵ dielectrics will help for about two generations. Figure 11.2 also shows the individual influence of copper and low-ϵ dielectrics. The signal propagation delay over a metal wire is proportional to the square of its length. The use of repeaters, however, reduces the propagation delay to a linear dependency on length. Particularly for longer wires, this may reduce the propagation delay by more than a factor of two (curves 8 and 9) The increasing clock skew and propagation delay for global signal wires are in direct contrast to the reducing clock period. Therefore there will be an increased drive to limit the size of the standard cells blocks (between one to several 100 K gates), which will also limit local interconnect lengths and clock skew. Designs will therefore become globally asynchronous and locally (within blocks) synchronous (GALS). To further relief the propagation delay problems, pipelines could be built into the global interconnects, but the bus latency will then become an important design parameter.

Figure 11.4 shows cross-sections of a 0.5 mm and a 0.1 mm transistor. Both transistors are drawn to different scales, so that the channel lengths in the drawing are equal. The figure clearly demonstrates the increased dominance of the interconnect in future deep-submicron processes.



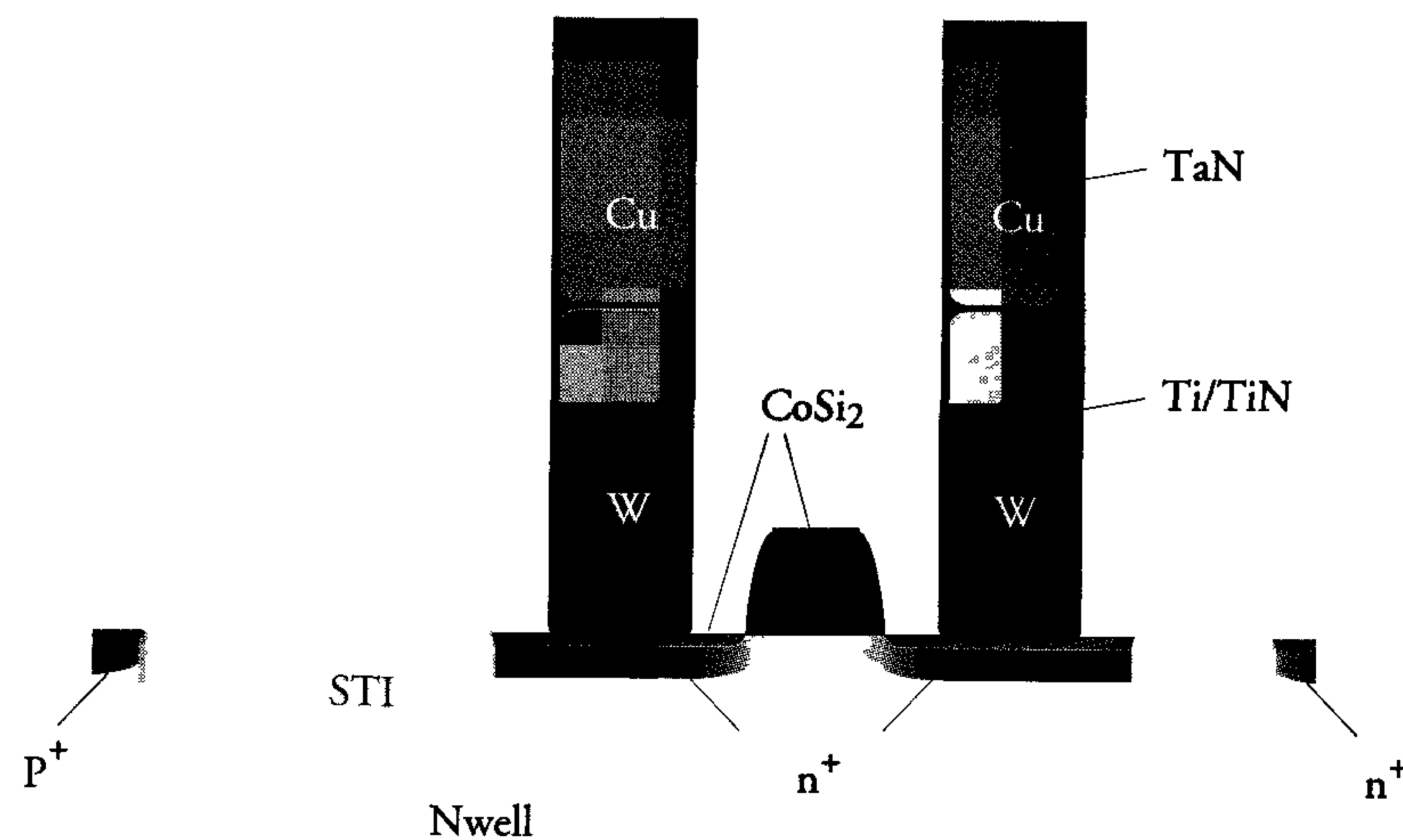


Figure 11.4: *Cross-sections of a 0.5 μm and a 0.1 μm transistor with normalised channel lengths*

11.4 Scaling consequences for overall IC design

If we integrate complete boards on one single chip, not only the functionality of the board is included, but also common board problems, such as supply noise, interference and EMI. In this section, we discuss the scaling consequences for overall chip performance, design reliability and signal integrity.

As the transistor bias voltages scale with the same factor as the transistor dimensions, the scaling from 0.35 μm processes onwards is called constant-field scaling. Table 11.2 shows the effects of constant-field scaling on the overall chip parameters. These parameters depend on the scale factor (s) between successive process generations, being about equal to $s \approx 0.7$.

11.4.1 Scaling consequences for overall chip performance

Until recently, transistor bias voltages were not scaled with each process generation. This was called constant-voltage scaling. During this conventional scaling era, speed and power efficiency scaled quite differently to the current constant-field scaling process. Figure 11.5 shows the improvements in speed and power efficiency, based on the evolution during the past fifteen years and the expected evolution in the next decade.

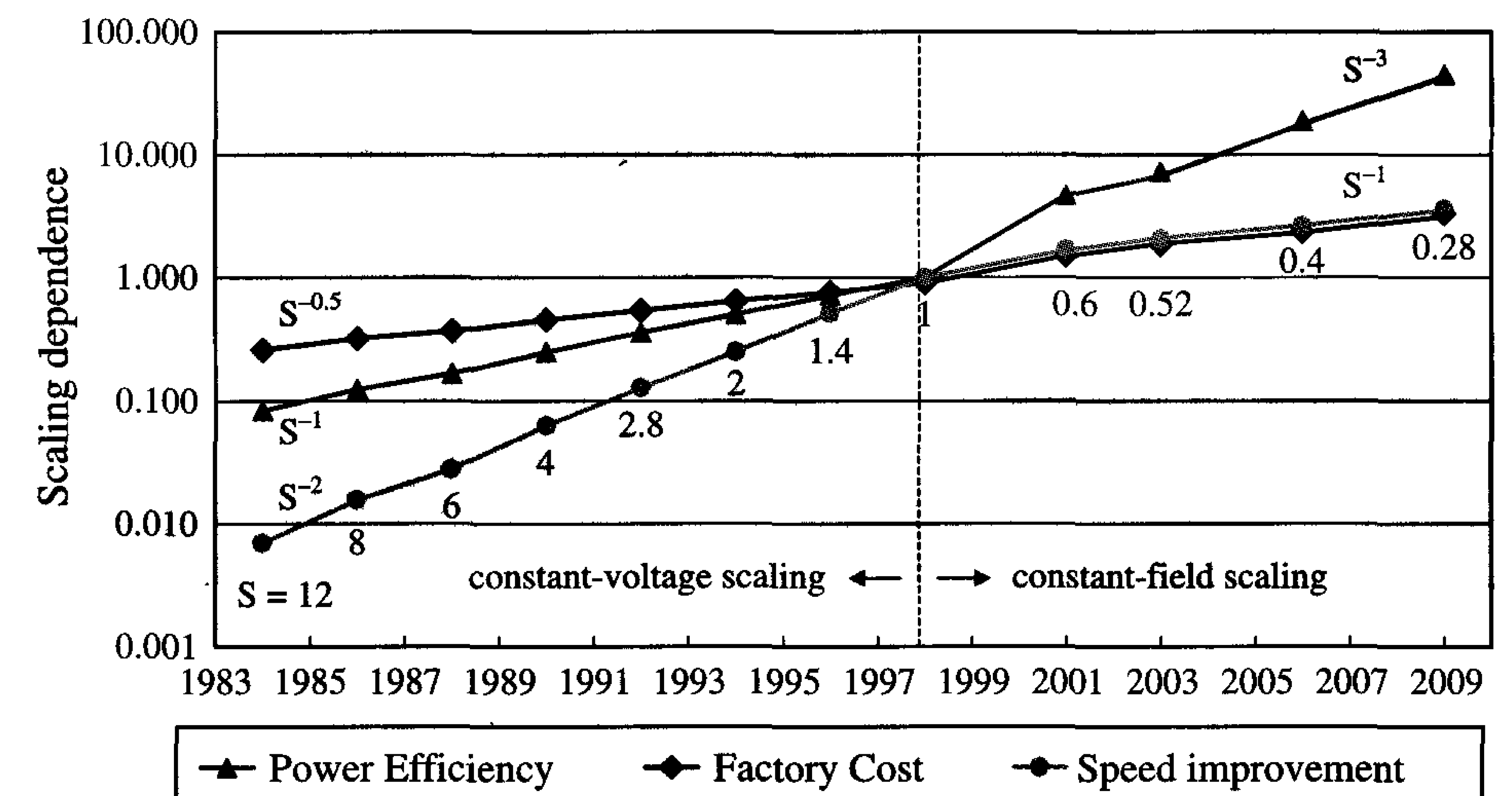


Figure 11.5: *Change in scaling dependencies of power efficiency, speed and factory costs*

Table 11.2: Effects of scaling on transistor and chip characteristics

Property	Effect	Scale factor dependence
electrical field strength		1
device demensions (W L, t_{ox} junction depth)		s
transistor area (logic gate area) (W L)		s^2
capacitances per unit area (C_u)		$1/s$
capacitances ($C = W \cdot L$ C_u)		[illegible]
bias voltage (V) and V_T		s
body effect (K factor)		$\sqrt{s}$
bias currents ($I = \mu C_{ox} \frac{W}{L} (V_{gs}\ V_T)^2$)		s
packing density (logic gates per unit area)	complexity	$1/s^2$
number of pins	bandwidth	$1/s$
gate delay (T_{min} = [illegible])	speed	[illegible]
power dissipation/gate (C V^2 f_{max})	power	s^2
power delay product (I V T_{min})	efficiency	s^3
ESD susceptibility ($I/$ t_{ox})	ESD	$1/s$
current density ($I/area$)	electromigration	$1/s$
power density (I $V/area$)	heating	1
latch-up sensitivity ($V > 1\ 2V$)	latch up	s
hot carrier degradation ($1\ 5V < V < 3\ 0V$)	hot carrier	s
sub threshold current	leakage	$12^{10(1\ s)V_T}$
noise margin	signal integrity	s
signal interference	mutual cross-talk	1/s
current density slew rate ($\frac{1}{area}$ $\frac{dI}{dt}$)	EMC	$1/s^2$
L $\frac{dI}{dt}$ noise density ($\frac{dI}{dt}/\#pins$)	supply bounce	$1/s$
ΔV per unit wire length (I R/V)	voltage drop	$1/s$
α-particle sensitivity	radiation	$1/s^2$

Until recently, the costs of building and equipping a factory increased by a factor of 1.5 every three years [3]. However, because of the rapidly increasing demands on the accuracy of the equipment and the class of the cleanroom, the factory costs are expected to double every factory generation (about a three-year period, twice the average period between

successive process generations). Because the inverse of the average scale factor is about equal to 1.5, while its square is about equal to 2, the factory costs can also be depicted in figure 11.5. As can be concluded from this figure, we might state that, until recently, chip size reduction and speed improvement were the drivers behind scaling. However, now, with constant-field scaling, power efficiency (i.e. computing power per Watt) has become the main driver behind the scaling process.

Finally, constant-voltage scaling is threatening mixed analogue/digital designs, which are becoming increasingly challenging and become questionable in performance and cost efficiency below 1.5 V.

11.4.2 Scaling consequences for overall design reliability

Reliability deals with several different aspects. In this paragraph, only the design-related reliability aspect is briefly discussed. For associated scaling dependencies, please refer to table 11.2.

Latch-up

In current 0.25 μm CMOS processes, the thyristor trigger voltage is around 1.8 V. In scaled processes, the bias voltages are scaled as well. Generally, at voltages below 1.5 V, hardly any internally generated latch-up is possible. However, as the latch-up requirements at the I/Os are at least maintained, I/O latch-up sensitivity will increase by a factor of $1/s$, as a result of closer n^+ to p^+ spacing. The I/Os will therefore require a relatively larger area to limit the possibility of latch-up.

Electromigration

Because electromigration is proportional to the current density, electromigration will increase by a factor of $1/s$. To maintain a proper overall power distribution, wider metal tracks will be required. The use of copper may also alleviate the electromigration problem. Experiments with copper at IBM [4] showed first improvements in electromigration properties (current capability) of about a factor of 10 compared with aluminium, see figure 11.6.

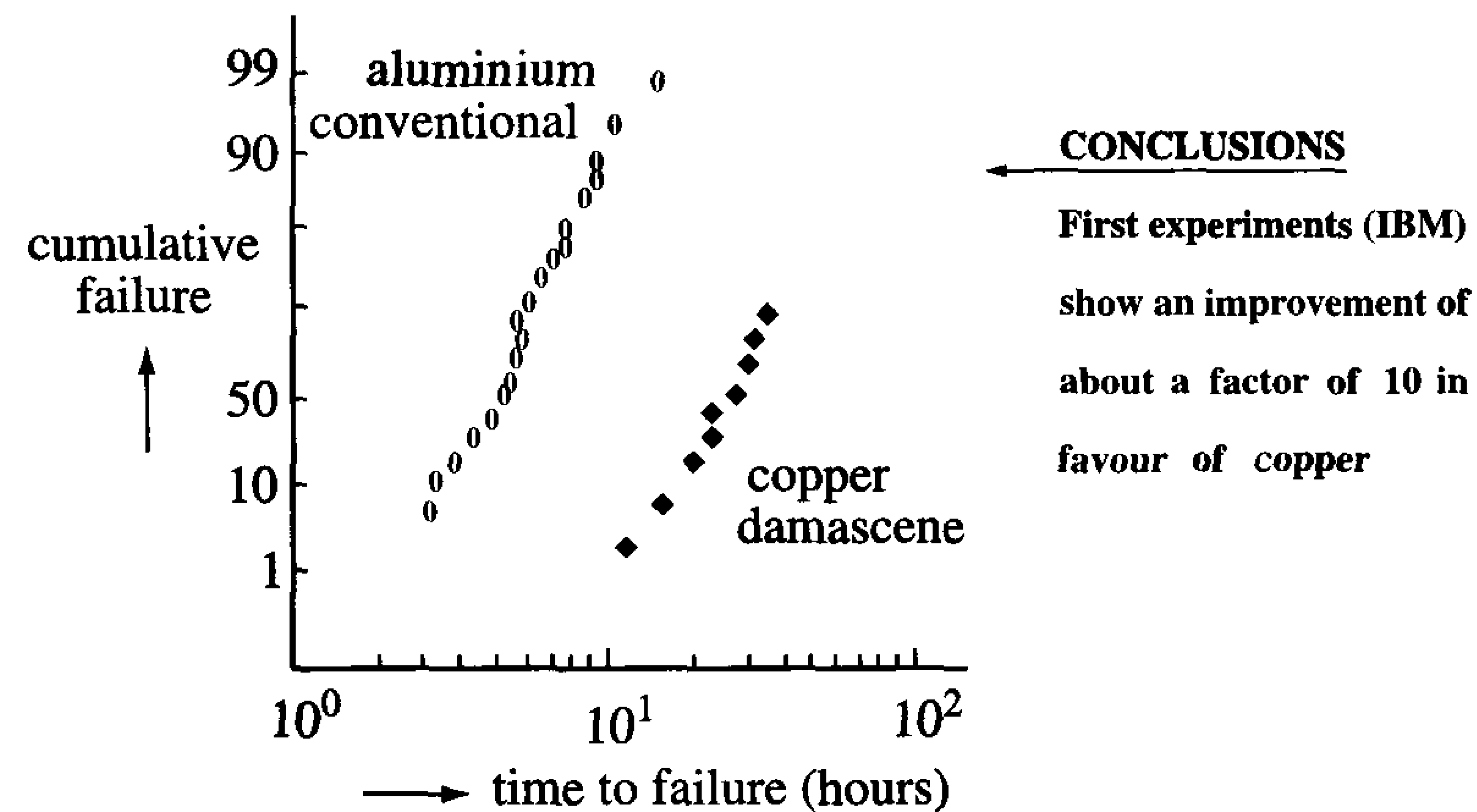


Figure 11.6: *Difference in electromigration properties of copper and aluminium*

Hot-carrier effect

The hot-carrier effect generally occurs when electrons and holes achieve energies above 3.2 eV and 3.8 eV, respectively. A simple conclusion could be that in a 2.5 V process, as an example, hardly any carrier can achieve energies higher than 2.5 eV. However, in the pinch-off region near the drain, some electrons can still achieve energies exceeding 3.2 eV after multiple collisions. As a general conclusion, we might state that devices become less susceptible to hot-carrier degradation and, below 1.5 V, it is no longer considered as a dominant effect.

Electrostatic discharge (ESD)

As a result of the scaling of the gate-oxide thickness, the input transistor sensitivity for ESD will increase by a factor of $1/s$. Because the protection must sink current to limit or clamp voltages to a certain level, power will be consumed inside the protection circuits. To limit power density, protection circuits cannot be scaled similarly to functional circuits. They will either occupy relatively larger areas or they will require novel structures.
In summary, the phenomenae that determine the reliability change every process generation according to figure 11.7.

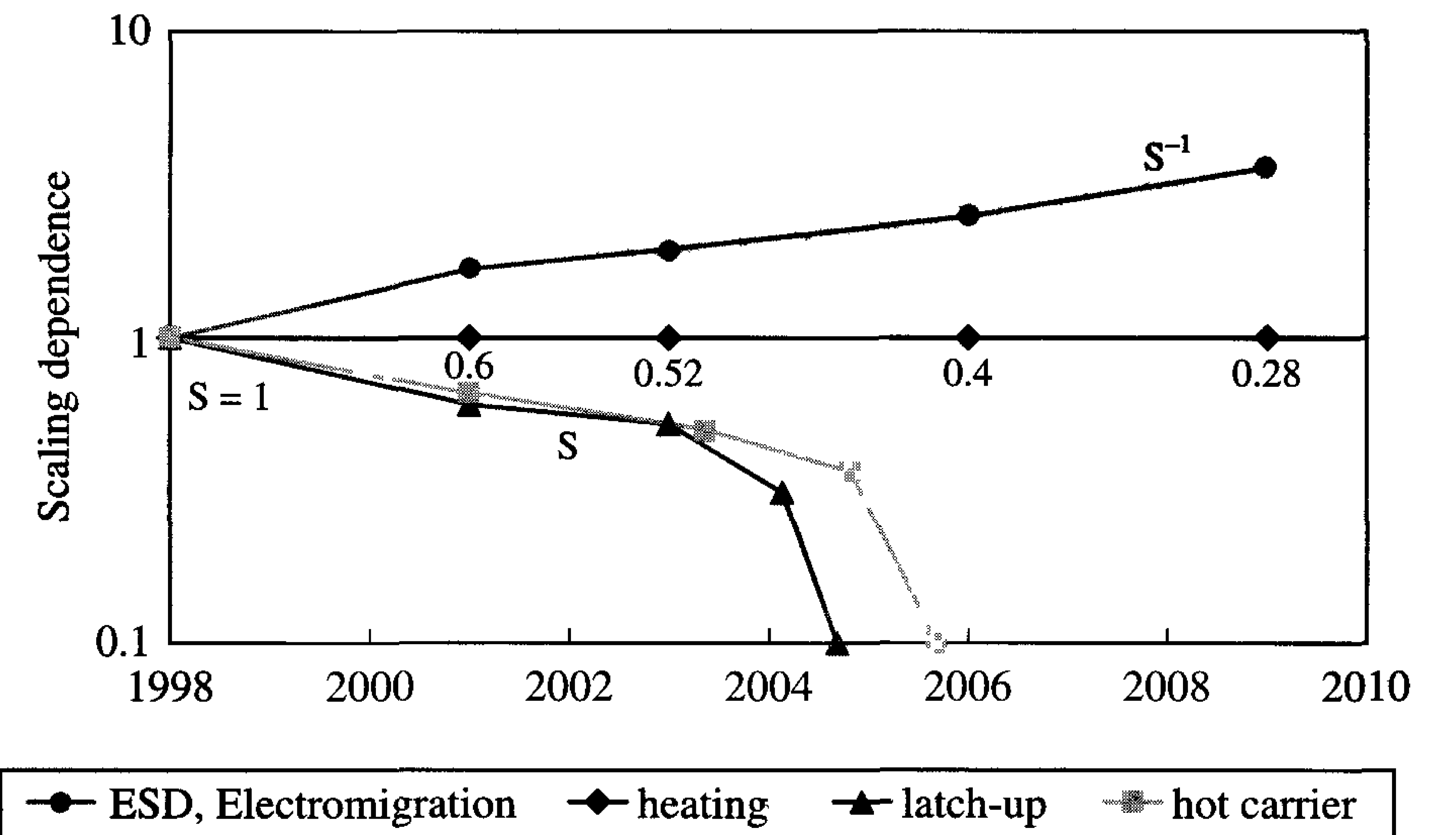


Figure 11.7: *Scaling dependencies of reliability parameters*

11.4.3 Scaling consequences for overall signal integrity

As a result of voltage scaling, noise margins reduce by a factor of s. At the same time, scaling the interconnection layers leads to increased resistances and mutual capacitances, which also starts to threaten the on-chip signal integrity.

Cross-talk

Narrower metal tracks at closer distances cause the signal interference (cross-talk) to increase by a factor of 1/s. Figure 11.8 shows a victim wire, embedded in between two source wires. The cross-talk is expressed as:

$$\Delta V_{\text{victim}} = \frac{2C_{\text{lateral}}}{C_{\text{vertical}} + 2C_{\text{lateral}}} \times \Delta V_{\text{source}}$$

In a $0.25\,\mu$m CMOS process, the cross-talk level between minimum spaced signal tracks could be as high as a factor of 0.8.

Especially victim wires connected to floating nodes can be put and left in a different logic high-ohmic state after a transient on the source wire. Tri-state buses and precharged bit lines (in memories) are very critical in this respect. Even if the lines are driven by relatively low-ohmic drivers, the cross-talk can reach unacceptable levels in large standard cell blocks.

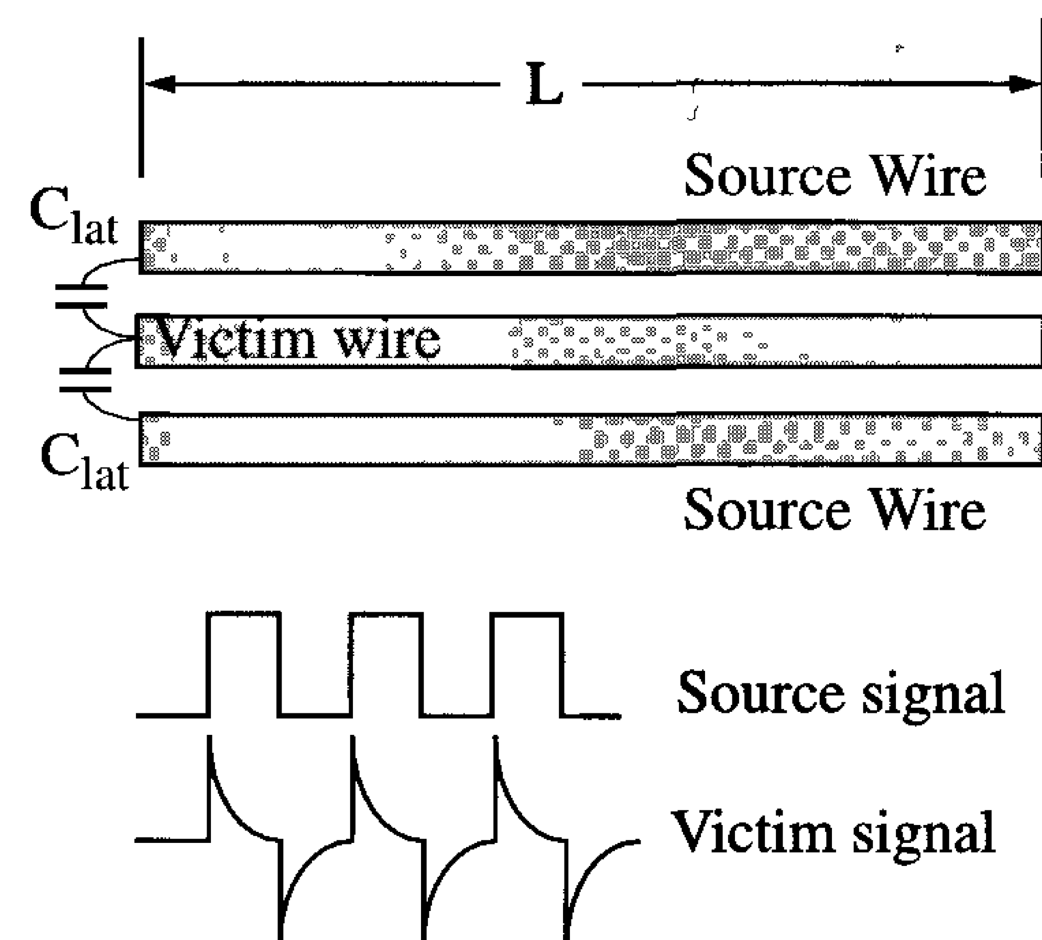


Figure 11.8: *Cross-talk from source wires to victim wires*

Particularly long signal wires, such as buses, scan control, reset and other global control signals are very susceptible to cross-talk. Tool vendors (e.g. Cadence) are developing cross-talk features, which support verification of the cross-talk in a logic block and allow rerouting of the critically interconnected nodes.

Voltage drop

Increased currents, combined with narrower metal tracks, lead to larger voltage drops ($\Delta V = IR$; which is proportional to $1/s$) across the supply lines. Therefore, a very structured and low-ohmic supply network is required to limit the influence of this voltage drop on the performance.

Switching ($\mathrm{d}i/\mathrm{d}t$) noise; supply bounce

Bus widths of microprocessors and memories have increased from 4 and 8 bits during the seventies to 32 and 64 bits today. Even 128-bit wide communication buses have been published [5]. Also, increased instruction words, such as used in Very Large Instruction Word (VLIW) architectures, will result in higher switching activities. This, combined with an increase in circuit density, results in a larger voltage drop across the resistance of the on-chip supply network ($\Delta V = IR$) as well as across the self-inductance of the bond wires and package leads ($\Delta V = L\mathrm{d}i/\mathrm{d}t$). This may lead to a large supply and substrate bounce. Particularly in mixed analogue/digital ICs and high-performance digital ICs, substrate bounce is one of the major threats to the integrity of operation of the chip.

There are three ways to reduce the level of this bounce. The first one is to reduce the self-inductance by adding more supply pins and/or using low-inductance packages. The second one is to lower the $\mathrm{d}i/\mathrm{d}t$. An important measure to achieve this is to optimise the clock network, since it contributes a relatively large amount to the total supply bounce. Tailored driver turn-on is another means to lower simultaneous activity. Finally, the third way to limit supply bounce is to incorporate decoupling capacitors on the chip.

Because the gate-oxide thickness is the smallest dimension in the manufacture of CMOS ICs, the gate capacitance is most commonly-used for implementing on-chip decoupling capacitors. Initially, only an nMOS transistor was directly connected between V_{dd} and V_{ss}, see figure 11.9.

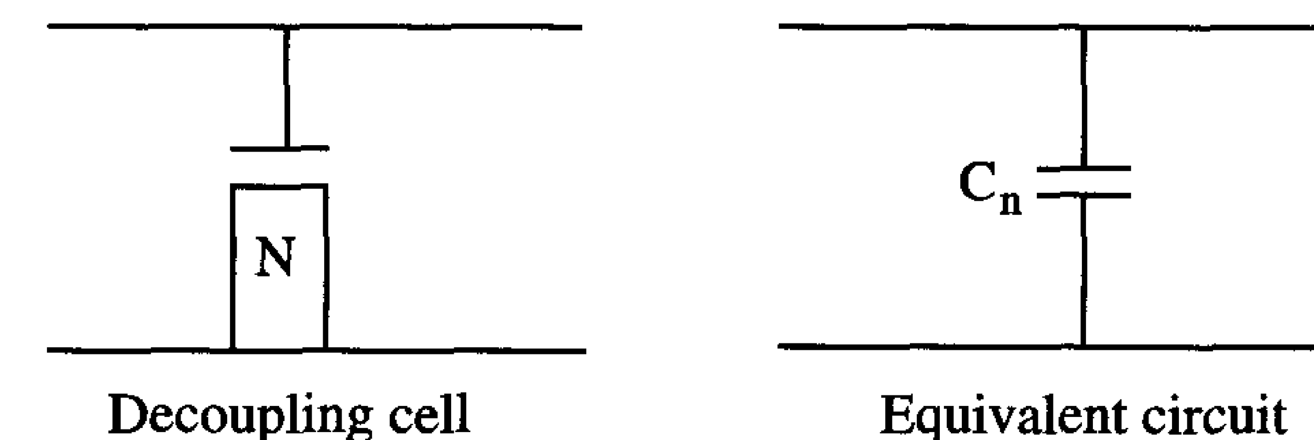


Figure 11.9: *Basic conventional MOS capacitor*

However, as a result of ESD requirements, it is not advised to connect an nMOS transistor's gate directly to either a V_{dd} or V_{ss} line. This means that the nMOS transistor has to include some resistance between its gate and the V_{dd} supply. Figure 11.10 shows a very efficient implementation.

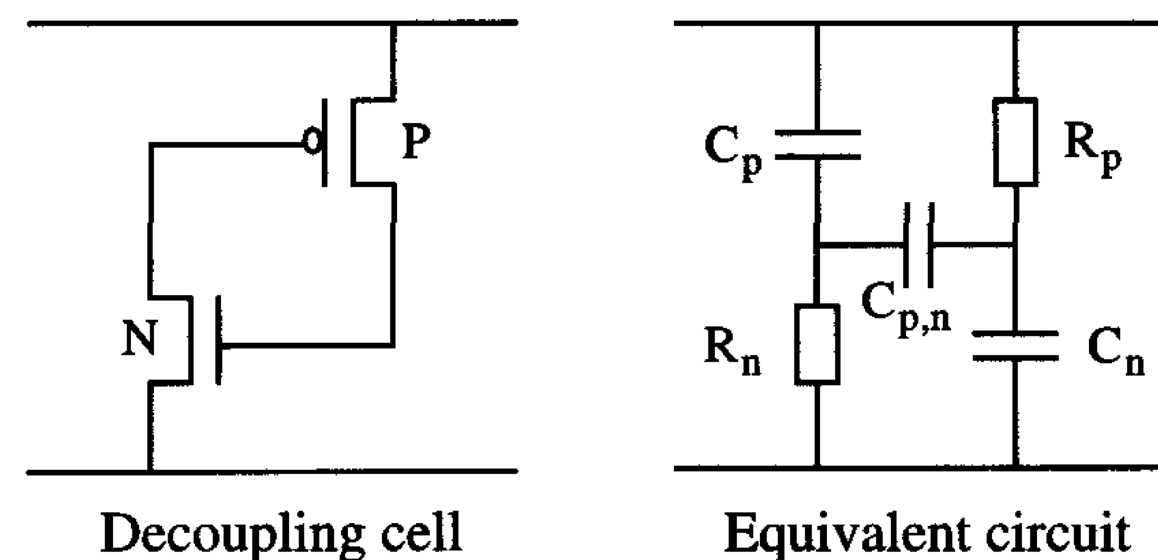


Figure 11.10: *Tie-off cell used as decoupling capacitor*

This cell generates dummy V_{dd} and V_{ss} levels, which allow direct connection to transistor gates where required. It is therefore also called a 'tie-off' cell. The advantage of this cell is that both the nMOS and pMOS transistors are always on and that both gate capacitances contribute to the total decoupling capacitance between V_{dd} and V_{ss}. Figure 11.10 also shows the equivalent circuit diagram. It can easily be seen that the pMOS transistor serves as a resistor through which the nMOS gate capacitor is charged and discharged.

Another advantage of this capacitor cell is that it consists of both types of transistors. This allows this cell to be placed inside standard cell blocks, to compensate for supply noise right where it is generated: inside the logic blocks. Even in a five- to six-metal layer process, complex standard cell blocks have utilisation factors of 85% to 90%. This means that 10% to 15% of the standard cell blocks are still 'empty' and only used for routing purposes. The decoupling cell of figure 11.10 easily fits in the library and is implemented in different versions, each occupying a different number of layout grids. These cells can automatically be placed in the empty locations as part of the design flow, see also chapter 9.

Because these cells only need first metal for interconnection, they are transparent for routing and do not obstruct routing of the logic block. At a utilisation factor of 85%, a 15% area fill with these capacitor cells will increase the decoupling capacitance of the total logic block by almost a factor of two. As a result, supply bounce is reduced by the same factor, without any area penalty. For logic blocks with a high switching activity, additional area must be created for decoupling capacitances, leading to a larger logic block area.

EMC

High current peaks not only generate on-chip supply noise; when flowing through bond wires, they also introduce electromagnetic noise on PCBs. As a result they generate an electromotive force (emf):

$$emf = -L\frac{\mathrm{d}i}{\mathrm{d}t}$$

As bond wires, package leads and board wiring act as antennae, they can radiate large EMPs (electro-magnetic pulses). Because both the density and the $\mathrm{d}i/\mathrm{d}t$ scale by a factor of $1/s$, the total radiation scales by $1/s^2$. In electro-magnetically compatible circuits, the maximum allowed value of an EMP is limited according to internationally-defined standards. On-chip measures to reduce EMI (electro-magnetic interference) include the just-discussed measures to limit supply bounce.

α-particle sensitivity

If an output node of a logic gate in a digital logic block is hit by an α-particle, the logic value at that node may temporarily be destroyed. This is because the input signals to that logic gate are maintained, thereby regenerating the original output state again after a very short time. This kind of α-particle disturbance will hardly result in an operating failure. However, if the diffusion area of an output node of a latch (or flip-flop) is hit by an α-particle, this might cause a permanent change of state of that latch. The stored charge on an output node of a minimum sized cross-coupled latch in a 0.25 μm CMOS technology is roughly 2 to 10 fC.

This amount of critical charge to disturb proper circuit operation is much less than the maximum charge (130 fC), which an α-particle can deposit on a single node of a present-day device [6]. Therefore, it is not a question of whether the latch will change its state when it is hit by an α-particle, but what is the chance of being hit? This chance is relatively low, because the critical latch area (output diffusion area) is much smaller than the total latch area. Also, the number of latches in a standard cell block is very low, compared to the many memory cells in a large memory chip. However, with scaling, this chance increases by a factor $1/s^2$, as the number of flip-flops per mm^2 increase s by that factor. In summary, the effects that influence the signal integrity tend to increase every process generation. Figure 11.11 shows their dependency on the scaling.

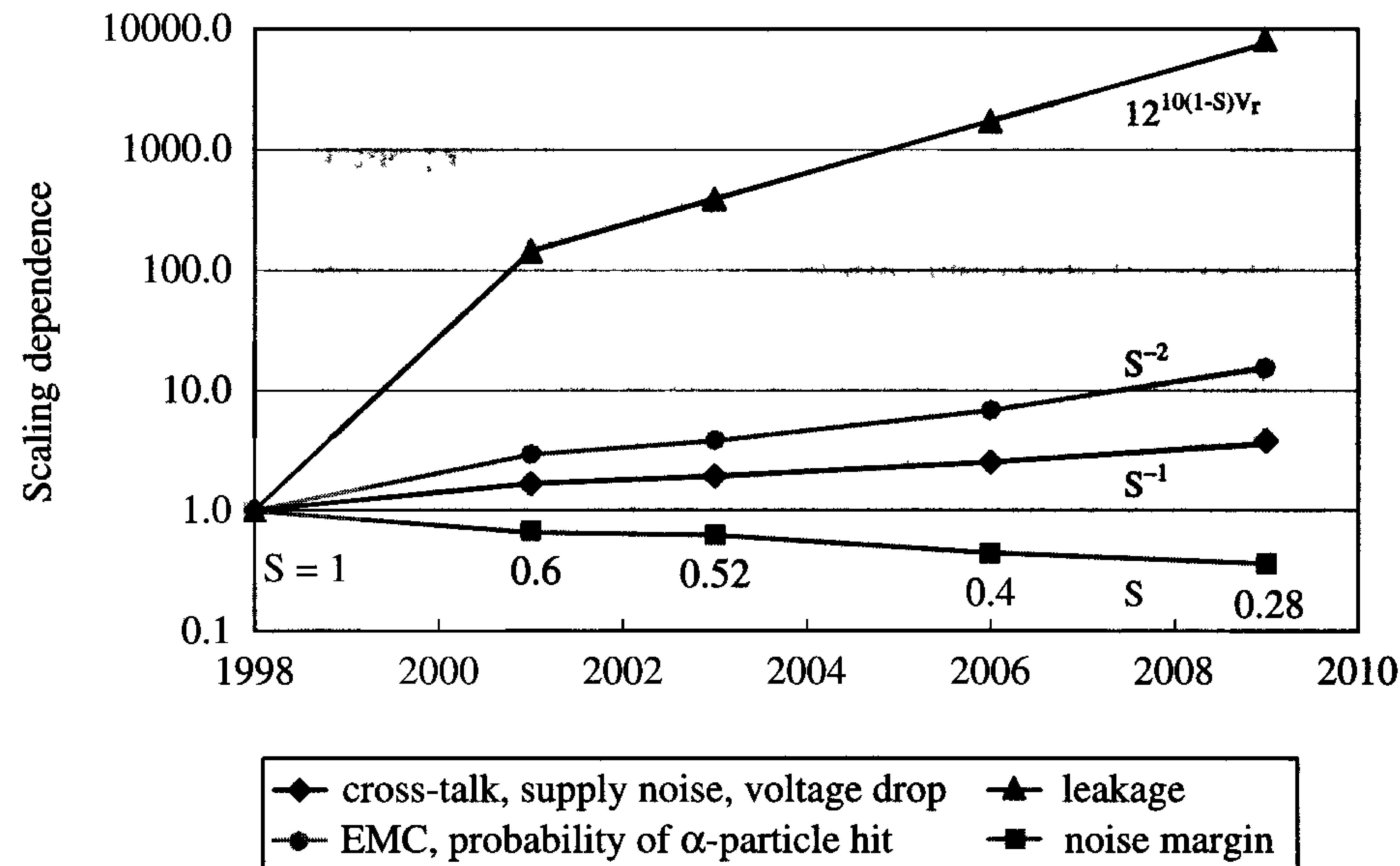


Figure 11.11: *Scaling dependencies of integrity issues*

11.5 Potential limitations of the pace of scaling

The observation of Gordon Moore's law (a quadrupling of IC complexity every three years) has proven its validity from the invention of the chip until now. It is sometimes called a self-fulfilling prophecy and is viewed as a measure for future trends and sets the pace of innovation. Almost according to this law, the Semiconductor Industrial Association has set up its roadmap for the next couple of years. Table 11.3 shows the most important parameters of this roadmap [8].

The previously-discussed scaling trends show that there are potentially three important key factors that may limit the pace of scaling. The complexity of MOS ICs increases exponentially with time, as can be seen from the table. However, the complexity of the design and test tasks is accelerated and forms a potential barrier to obtaining full exploitation of the available manufacture potentials. The overall success of the semiconductor industry will be increasingly dominated by how the complex design, engineering and test challenges will be addressed [8]:

- increased design databases
- increased complexity of design flows
- increased hierarchy of synthesis
- increased integration of reusable cores
- increased complexity of timing modelling (clocks, interconnections)
- increased design effort (design teams)
- increased noise and reliability problems
- increased number of redesigns
- increased complexity of failure analysis
- increased costs of testing (test time, hardware and software).

The ability to completely verify, test, debug and diagnose future complex designs will reduce dramatically. It is therefore likely that current design styles with fixed and dedicated logic will be replaced by design styles that allow flexibility and configurability. This flexibility can be enhanced by software solutions (programmability) as well as hardware solutions (*configurable computation* such as embedded *FPGA* and/or sea-of-gates architectures). Remaining bugs can then be bypassed by changing the program or by remapping the function, respectively.

Another potential key factor in lowering the pace of process innovation, which is already discussed, is formed by the economics involved. From 1966 to 1995, the costs of a wafer factory increased by a factor of 100, from about $15 million to $1.5 billion respectively [3]. The first $3 billion factory was built in 1998. As long as the performance of ICs improves faster than the costs increase, these heavy investments are worthwhile. While the chip performance between 1984 and 1990 tripled, the investments in wafer factories only doubled. For the near future, the investments per generation will double again. However, as a result of the supply voltage reduction every generation, the performance increase will be reduced to only a factor of 1.5.

If this trend continues, the costs of a wafer factory will reach about $10 billion by the year 2005. These investments can only be raised when competitors start alliances to share new wafer factories. As a result of the previous trends in IC performance and wafer factory costs, the price

per transistor can bottom out somewhere around 2005. It might thus occur that there will be no reason to scale transistors any further after some period of time [3].

The third key factor that may limit the pace of scaling is represented by the increased manifestation of physical and electrical effects in deep-submicron technologies. Larger current slew rates (di/dt) and mutual signal track capacitances will bring the circuit noise to unacceptable levels. In addition to this, the noise margins of future processes will further decrease due to the reduction of the supply and threshold voltages (figure 11.12). Every new technology requires additional design and/or technology measures to reduce the noise and increase the gap between the noise and the noise margin. However after scaling to the next technology, the problem is the same again and new measures are required. Relatively large additional chip areas must be devoted to on-chip measures like decoupling capacitances and to more widely-spaced buses and other global signal interconnections etc. These deep-submicron effects, which are extensively discussed in chapter 9, reduce the chance of fully exploiting the potentials of the new process generations. The level to which these effects will limit the efficient use of chip area cannot be predicted because it also depends on the creative design alternatives that will be developed in the near future.

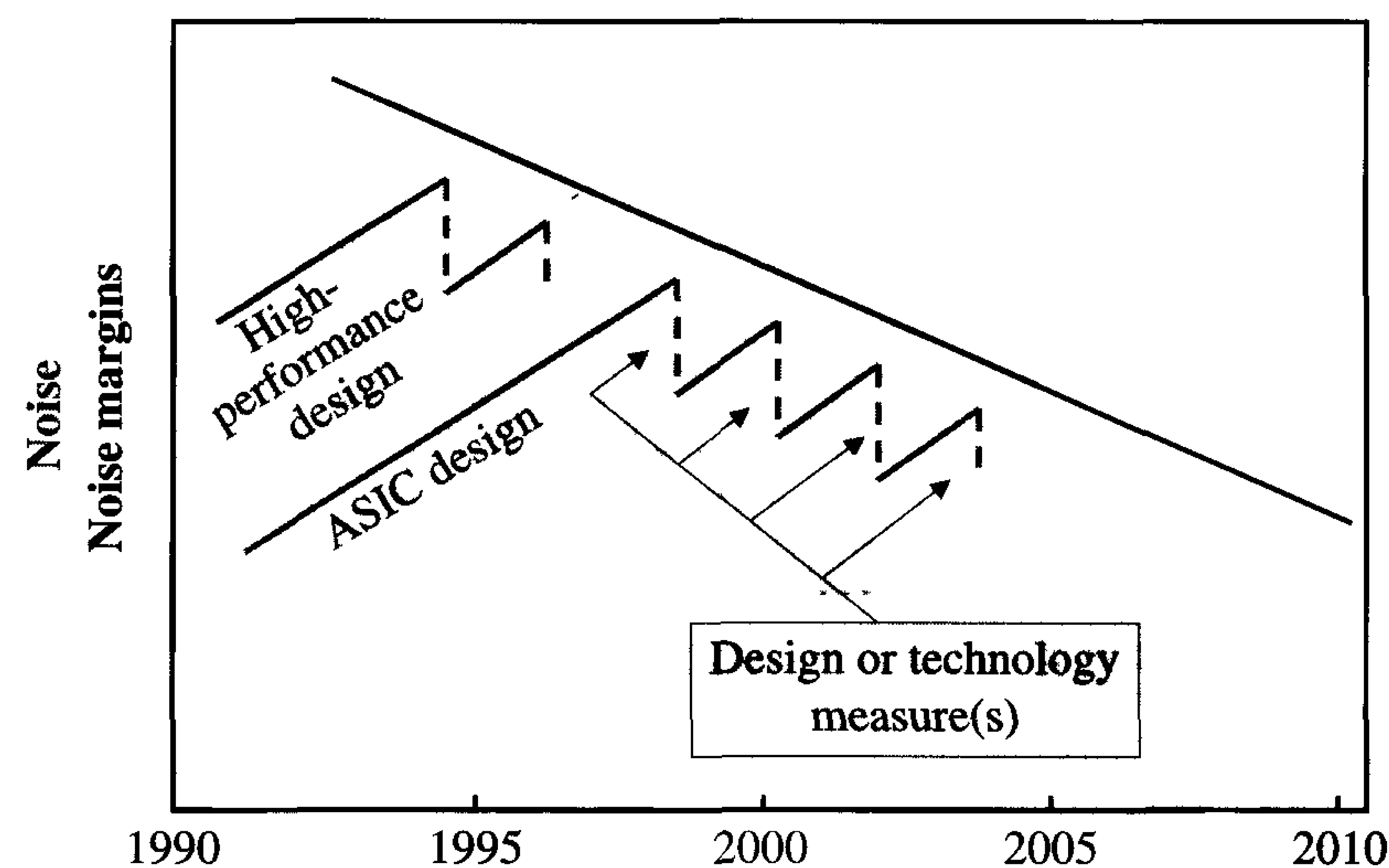


Figure 11.12: *The reducing gap between noise and noise margins.*

Table 11.3: Most important IC characteristics and their change according to SIA (ITRS) roadmap

Year of First IC Shipment Technology Generation	...1999...... 180nm	...2001...... 150nm	...2003...... 130nm	....2005..... 100nm	....2008.... 70nm	....2011..... 50nm	....2014..... 35nm
Power: Single-Chip Package (Watts)							
Low-cost	not available	not available	not available	not available	not available	not available	not available
Hand-held	1.4	1.8	2.2	2.4	2.5	2.6	2 7
Cost / Performance	48	61	75	96	104	109	115
High-Performance	88	108	129	160	170	174	183
Harsh	14	14	14	14	14	14	14
Chip Size (mm^2)							
Low-cost	53	55	61	65	72	81	90
Hand-held	53	57	61	65	72	81	90
Cost / Performance	170	170	214	235	270	308	351
High-Performance	450	450	567	622	713	817	937
Harsh	53	57	61	65	72	81	90
I/O Bus Widths (Bits)							
Low-cost	32	64	64	128	128	256	256
Hand-held	64	64	128	128	256	256	256
Cost / Performance	64	128	128	128	256	256	256
High-Performance	128	256	256	256	512	512	512
Harsh	32	32	32	64	64	64	128
Performance: On-Chip (MHz)							
Low-cost	300	415	510	633	840	1044	1250
Hand-held	300	415	510	633	840	1044	1250
Cost / Performance	600	727	890	1100	1400	1800	2200
High-Performance	1200	1454	1724	2000	2500	3000	3600
Harsh	25	60	60	60	100	100	100
Performance: Chip-to-Board for Peripheral Buses (MHz)							
Low-cost	75	100	100	100	125	125	150
Hand-held	75	100	100	100	125	125	150
Cost / Performance	133/300	150/362	150/445	150/550	175/700	200/900	225/1100
High-Performance	600	727	862	1000	1250	1500	1800
Harsh	25	60	60	60	100	100	125
Memory (D/SRAM)	133/360	150362	150/445	150/550	175/700	200/900	225/1100
Logic (High-Volume : Microprocessor) Cost-Performance at ramp-up							
MTransistors/cm^2 packed including on chip SRAM	7 M	14 M	22 M	41 M	100 M	247 M	609 M
' Affordable ' Cost/Transistor @ (Packaged-microcents)	1735	868	434	217	-	27	-
Package Pincount							
Low-cost	80-290	90-338	109-395	127-460	160-580	201-730	254-920
cost/performance	370-740	432-912	503-1123	587-1384	740-1893	932-2589	1174-3541

In this SIA roadmap, the following definitions are used for the different IC categories:

- Low-cost < $300 consumer products, microcontrollers, disk drives, displays
- Hand-held < $1000 battery-powered products; mobile products, hand-held cellular and other hand-helds
- Cost-performanc < $3000 notebooks, desktop personal computers, telecommunications
- High-performance > $3000 high-end workstations, servers, avionics, supercomputers, most demanding requirements
- Harsh under-the-hood and other hostile environments

11.6 Conclusions

Conventionally, the drive for a continuous scaling of integrated circuits has been the shrinkage of the circuits and of the systems built from them, plus the increased speed that accompanied the advent of every new generation. However, scaling not only influences the system sizes and performance positively, it also has major negative effects on the reliability and signal integrity of deep-submicron ICs.

These effects have increased to such an extent that digital ICs can no longer be regarded as circuits that propagate ones and zeros in a certain order to perform certain functionality. The design of digital circuits increasingly requires an analogue approach to maintain reliability and signal integrity at a sufficiently high level. The manifestation of the responsible physical effects, which grows with scaling of the feature sizes, will even be a further challenge, if not a threat to reliability and signal integrity of future VLSI designs. An understanding of the effects of scaling is essential for the efficient exploitation of the full potential of modern deep-submicron IC manufacture processes.

These effects place high demands on the design and test strategies used for modern ICs and systems. Additional measures in the design are needed to maintain testability, observability, reliability and signal integrity at a sufficiently high level. These measures all require additional chip area, which limit an efficient exploitation of the potentials of the new process generations.

Before the year 2010, we will face the fact that a move to the next process generation will no longer be commercially attractive for a lot of products. For cheap high-volume consumer products, however, this point of time will already be reached within only a few process generations.

11.7 References

[1] M. Izumikawa, et al., 'A 0.25 μm 0.9 V 100 MHz DSP core', IEEE-JSSC, Jan. 1997, pp 52-61.

[2] T. Kuroda et al. 'A 0.9 V, 150 MHz, 10 mW, 4 mm^2, 2-D DCT Core Processor with variable Threshold voltage Scheme', IEEE-JSSC, Nov. 1996, pp 1770-1779.

[3] R. Schaller, 'MOORE'S LAW: past, present and future', IEEE Spectrum, June 1997, pp 53-59.

[4] D. Edelstein, 'Full Copper Wiring in a sub-0.25 μm CMOS ULSI Technology', IEDM 1997, Digest of Technical Papers, pp 773.

[5] nvidea, 'RIVA 128ZX processor', *www.nvidia.com/products/frames_riva128zx.html.*

[6] Robert H. Dennard, 'Future CMOS scaling: approaching the limits?', Future FAB international, 1997

[7] B. Davari, R.H. Dennard, G.G. Shadidi, 'CMOS scaling for High-Performance and Low-Power; The Next Ten Years', Proceedings of the IEEE, vol 83, No 4

[8] Semiconductors Industrial Association, 'The International Technology Roadmap for Semiconductors', 1999 edition

[9] S.H. Lo, et al., 'Quantum-Mechanical Modeling of Electron Tunneling Current from the Inversion Layer of Ultra-Thin-Oxide in MOSFET's', IEEE Electron Device Letters, Vol. 18, No 5, 1997, pp 209-211

[10] Serge Luryi, et al. 'FUTURE TRENDS IN MICROELECTRONICS, The Road Ahead', John Wiley & Sons, 1999

11.8 Exercises

1. Explain the differences between conventional constant-voltage scaling and the currently-applied constant-field scaling process. How did they influence the main driving force behind the scaling process?

2. Why was copper not used in the early MOS processes? What is the result of using copper instead of aluminium for the interconnection patterns of an IC?

3. An IC with channel lengths of 0.25 μm is manufactured in a 0.18 μm CMOS process and used in a particular application. Suppose this IC is scaled by a factor of 0.8 and manufactured in the same process. What would happen to the following parameters when this IC is used in the same application:

   a) the transistor gain factors β_n and β_p

   b) the threshold voltages V_{T_n} and V_{T_p}

   c) the chip's power dissipation

   d) the chip's power density

   e) the noise on the chip's supply and ground lines.

4. What are the consequences of the scaling on the performance and reliability of the circuit in exercise 3 if the geometrical effects mentioned in chapter 2 are also taken into account?

5. Suppose that the additionally required decoupling capacitance on a chip results in an area penalty of 50 percent. How could the capacitance density (i.e. capacitance value per unit area) be increased by technology means?

Index

1 T-cell, 242
2-phase, 184
3 T-cell, 272

α-particle radiation, 243, 251, 518
abstraction level, 280
accelerating voltage, 98
acceptor, 8
access time, 234, 239, 246
accumulation
  ~ capacitor, 10
  ~ layer, 37
  ~ process, 10
active area, 112, 188
ACTIVE mask, 107, 109, 188
  ~ programmed ROM cell, 254
activity factor, 339, 361, 422
adaptive clock skew control, 409
address buffer, 236
allocation, 292
ALU, 282
aluminium gate process, 4, 10, 109, 507
AND
  ~ matrix, 308
  ~ function, 146
anisotropic etching, 71, 90
annealing, 98, 123
antenna effect, 91, 105
anti-punch-through implant, 66, 119
Anti-Reflective Coating, 120
APCVD, 95
application-specific
  ~ integrated circuit, 224, 274
  ~ standard product, 274, 276
APS, 214
APT implant, 66, 119
ARC, 120
Area Usage Factor, 450
ASIC, 224, 274
  ~ turn-around time, 274, 276
aspect ratio, 133, 152, 158
ASSP, 274, 276
asynchronous
  ~ circuits, 368
  ~ design, 368
ATE, 443
AUF, 450

β_n, 156
β_p, 157
back-bias, 388
  ~ controlled V_T, 341
  ~ effect, 26, 122, 150, 151, 480
back-gate effect (see back-bias)
balanced clock tree, 404
Ball-Grid Array, 413, 466
bandgap, 6
basic
  ~ CMOS process, 112
  ~ MOS technologies, 107
  ~ nMOS process, 107
batteries, 338
battery RAM, 263
BCCD, 210, 213
  ~ surface-state immunity, 213
BCD counter, 367
BGA, 413, 466
BICMOS, 220
  ~ characteristics, 222
  ~ circuit performance, 223
  ~ digital circuits, 220

~ NAND gate, 223, 224
~ performance, 225, 226
~ technology, 209, 220
bipolar
~ gain factor, 222
~ noise, 222
bird's beak, 92
BIST, 446
bit
~ density, 443
~ line select, 236
~ -parallel operation, 284
~ -slice layout, 305
board substrate, 460
body
~ effect, 26, 122, 150, 151, 480
~ factor, 26
bond
~ pad, 415, 442
~ wire, 413
bootstrap, 146
~ -capacitance, 136
~ load, 136
bottom-up implementation, 280
Boundary Scan Test, 311, 446
BPSG, 95
BRAM, 263
breakdown
~ mechanism, 479
~ voltage, 218
bridging defects, 444
BST, 446
buffer circuits, 158
Built-in Self Test, 446
buried
~ contact, 189
~ -channel CCD, 210, 213
burst
~ mode, 239
~ rate, 249
bus latency, 508
bus width, 412
buses, 433

C-4 bonding, 460
CAD tools, 274, 291, 293
capacitances in MOS device, 36, 40, 122
capacitive
~ coupling, 428
~ load, 140
carrier mobility degradation, 55
$\overline{CAS}$, 245
Cascode Voltage Swing Logic, 182
CCD, 210
~ applications, 214
~ cell, 211
~ operating frequency, 213
~ shift register, 210
CCO, 408
cell-based IC design, 276
channel
~ dope, 5
~ length modulation, 57
~ resistance, 30
~ stopper, 33, 107, 119
~ -free gate array, 313
~ -led gate array, 312
~ -less gate array, 313
characterisation, 448
charge
~ bucket, 210
~ characteristic, 156
~ -coupled device, 210
~ distribution, 11
~ mobility, 31, 54, 57
~ neutrality, 23, 38
~ sharing, 172, 180, 444
~ transfer, 212
~ -Induced Voltage Alteration analysis, 494
Chemical
~ Mechanical Polishing, 101
~ Vapour Deposition, 94
chip, vi
~ select, 236
Chip Scale Package, 468
circuit
~ density, 113
~ simulation program, 46

~ testability, 443
~ -analysis program, 145
CIVA analysis, 494
class of clean room, 449
clean room, 449
clock
~ activity, 372
~ forwarding, 407
~ generation, 408
~ jitter, 411, 415
~ signals, 173
~ skew, 175, 179, 184, 404
~ tree synthesis, 405
~ trunk, 405
~ -phase synchronisation, 409
clocked CMOS circuits, 173
clocking strategies, 184, 409
CMOS, 149
~ buffer design, 162
~ driver, 186
~ image sensors, 209
~ inverter, 151
~ inverter dissipation, 158
~ inverter transfer characteristic, 152
~ latch, 173
~ NAND gate, 224
~ output buffer, 186
~ parasitic bipolar device, 475
~ process, 112, 187
~ transmission gate, 170
CMP, 101
column decoder, 236
commodities, 278
compiled cell, 276
Complementary Pass-Transistor Logic, 357
computing power, 378, 512
concentration gradient, 71
conduction band, 6
configurable computation, 520
constant
~ -field scaling, 510
~ -voltage scaling, 512
CONTACT mask, 108, 109, 189
~ mask programmed ROM cell, 257
contact
~ plug, 114, 121
~ (re)fill, 121
~ resistance, 120
continuous array, 314
control
~ bus, 279
~ path, 281
controlled-collapse chip connection, 460
copper, 122, 507, 512
core, 276, 311
correct by design, 386
cost
~ of interconnect, 264
~ of wafer factory, 520
CPL, 357
CPLD, 320
critical
~ delay, 406
~ path, 290, 407
~ process steps, 451
cross-talk, 181, 427, 444, 514
current
~ density, 2, 96
~ fluctuations, 412
~ hogging, 218
~ slew rate, 412
~ -controlled oscillator, 408
custom IC, 276
customer return, 442
customisation, 274, 278, 312
CVD, 94
CVSL, 182
cycle
~ stealing, 407
~ -time, 234

ΔV_T implant, 62
damascene
~ back-end flow, 507
~ patterning, 121
dark current, 212

data
~ base organisation, 436
~ bus, 279
~ input buffer, 236
~ output buffer, 236
~ path, 281
~ retention time, 231, 261, 268
DDR, 250
decision tree, 288
decoupling capacitor, 420, 422, 516
deep-submicron
~ design, 385, 428, 505
~ technology, 114
defect density, 450
delay-locked loop, 411
depletion
~ area, 59
~ layer, 11
~ load, 138
~ process, 11
~ transistor, 31, 142
deposition, 94
Depth Of Focus, 101
design
~ documentation, 436
~ efficiency, 504
~ flow, 291
~ for debug, 496
~ for manufacturability, 450
~ for reliability, 386, 512, 514
~ for signal integrity, 398, 514, 519
~ hierarchy, 437
~ organisation, 436
~ path, 279
~ productivity, 264
~ resources, 504
~ rules, 187
~ -rule-check, 84, 303
~ style, 520
~ -verification, 294
~ of a CMOS inverter, 156
destructive read-out, 242, 253
diamond
~ saw, 453
~ tipped scribe, 453
DIBL, 75
dielectric relaxation time, 37
Differential Split Level Logic, 183
diffusion, 97
~ barrier, 507
~ coefficient, 97
digital
~ circuits, 167, 275
~ potentiometer, 283, 294
DIL, 454
direct
~ contact, 189
~ slice writing, 87, 275
discharge
~ characteristic, 156
~ of a capacitance, 144
dishing, 102
distributed clock network, 404
DLL, 411
DMOS transistor, 217
DOF, 101
DOMINO-CMOS, 177
donor, 8
dope
~ fluctuations, 435, 506
~ profile, 68, 98
Double Pass-Transistor Logic, 358
double
~ data rate, 250
~ -diffused MOS transistor, 217
~ -flavoured, 94
~ -flavoured polysilicon, 112, 152
DPL, 358
Drain-Induced Barrier Lowering effect, 75
drain, 4
~ extension, 108, 119
~ series resistance, 72
DRAM, 74, 232, 242
~ access time, 246
~ architectures, 245
~ cell, 242, 244
~ high performance, 247

drc, 303
drive current, 57
driver transistor, 132, 144
DRO, 242, 253
DSL, 183
DSW, 275
dual
~ -damascene, 121
~ -dope polysilicon, 152
~ -edge triggered flip-flops, 373
~ -in-line-package, 459
~ polysilicon, 112
~ -V_T concept, 505
dynamic
~ CMOS, 176
~ CMOS latch, 178
~ CMOS shift register, 178
~ flip-flop, 178, 179
~ memory, 232
~ power consumption, 339
~ power dissipation, 159
~ RAM, 232, 242
~ shift register cell, 178

Early voltage, 60
early failure rate, 476
E-beam, 87
~ microscopy, 493
~ pattern generator, 84
~ test techniques, 448
EBPG, 84
EDO, 247
E/D technology, 138
EEPLD, 277
EEPROM, 259, 260
effective channel length, 58, 108
electric field, 11, 70
electrical endurance test, 475
electro-optic sampling, 448
electromagnetic
~ compatibility, 419
~ pulse, 419
electromigration, 96, 396, 512
electromotive force, 412, 419, 518
electron
~ mobility, 2, 54, 55, 152
~ valve, 1
electrostatic
~ charge, 449
~ discharge, 390, 475, 513
embedded
~ arrays, 311
~ FPGA, 520
~ logic, 265
~ memory, 231, 265
~ software, 293
EMC, 419, 518
emf, 412, 518
emitter ballasting, 218
EMP, 419
emulation, 291, 294
endurance characteristic, 261
energy
~ band, 6
~ band diagram, 14
~ band theory, 5
~ barrier, 66
~ gap, 6
engineering, 448
enhancement transistor, 31, 141
epitaxial layer, 94, 389
epitaxy, 94
EPLD, 277
EPROM, 259
ESD, 123, 390, 423, 475, 513
~ test models, 391
e-sort, 442
etching, 89
evaporation, 96
EXOR, 171, 372, 376
Extended Data Out, 247

failure analysis, 482
fan-in, 506
Fast Page Mode, 247
fat zero, 212
fault coverage, 442
feature size, 85
Fermi level, 8
ferroelectric RAM, 252

FGDW, 452
FIB, 495
field
~ -effect principle, 1
~ oxide isolation, 314
~ -programmable device, 277, 315, 520
FIFO, 233
fill factor, 214
fingertip method, 391
firm core, 277
first-time-right silicon, 385
first-silicon debug, 482
flash memory, 262
flat-band condition, 15
flip-chip bonding, 413, 460
flip-flop, 173, 174, 211, 314
~ scannable, 400
floating gate, 259
floorplan, 290
Focused Ion Beam, 495
Fowler-Nordheim tunnelling, 260
FPGA, 317, 520
FPM, 247
FRAM, 252
full adder, 147, 286
full-CMOS SRAM cell, 239
full-custom IC, 276
full-featured EEPROM, 260

gain factor, 25, 54, 55, 148, 157
GALS, 508
gate, 4
~ array, 312
~ capacitance, 422
~ density, 443
~ depletion, 94, 505
~ -drain overlap capacitance, 109
~ forest, 313
~ -isolation technique, 314
~ oxidation, 108
~ oxide, 93
~ oxide tunneling, 506
~ -source overlap capacitance, 109
gated clock, 374, 405
GDSII, 329
geometric layout description, 329
GLDL, 329
glitches, 371
globally asynchronous and locally synchronous, 508
glue logic, 281, 317
graded-drain transistor, 70
Gray code counter, 367
ground bounce, 412
guard ring, 33, 388

handcrafted layout, 303
hard core, 277
hardware
~ accelerator, 293
~ /software co-design, 291
HDD, 120
HDGA, 313
HDP, 90
hetero-epitaxy, 94
heterogeneous system, 281, 408, 503
hierarchical design approach, 326
High-Density Plasma, 90
high-density gate array, 176, 313
high-impedance node, 427
Highly-Doped Drain, 120
high performance DRAM, 247
hillocks, 396
hole mobility, 152
hole, 7
homo-epitaxy, 94
horizontal electric field, 56
hot-carrier
~ degradation, 67, 69, 398
~ effect, 67, 119, 513
hot-electron
~ degradation, 70
~ effect, 259, 70
hot spot, 218
human body model, 391
humidity
~ sensitivity, 475
~ test, 476

IC, vi
~ characterisation, 448
~ engineering, 448
~ infant mortality, 476
~ intrinsic failure rate, 476
~ lifetime, 93, 96
~ package, 476
~ quality, 475
~ reliability, 459, 475
~ wearout, 476
I_{ddq} testing, 444, 482
image sensor, 214
imaginary drain, 58
impact ionisation, 67, 69
implantation, 97, 98
In-System Programmability, 317
inductance, 413
inert
~ gas, 476
~ liquid, 476
infant mortality, 476
inner lead bonding, 460

input
~ protection, 123
~ TTL, 185
integrated circuit, vi
intellectual property, 276, 293, 408
interconnect
~ cost of, 264
~ parasitics, 426
~ sheet resistance, 507
interstitial dope atoms, 98
intrinsic failure rate, 476
intrinsic silicon, 8
inverse narrow-width effect, 64
inversion layer, 17
~ transistor, 4
inverter, 130, 131, 151, 152
~ chain, 163
~ DC behaviour, 132
ion, 7
~ acceleration, 98
~ implantation, 97
~ implanter, 97
ionisation energy, 8
IP, 276, 293, 408
isotropic etching, 89
ISP, 317
iterative multiplier, 283
ITRS roadmap, 434, 503, 522

junction spiking, 100

K-factor, 26, 28, 152
KGD, 278, 474
known-good
~ core, 278
~ die, 278, 474

Laser-Beam Pattern Generator, 84
laser beam
~ technique, 495
~ -fusing, 252
latch, 173
latch-up, 122, 222, 386, 475, 481, 512
lateral diffusion, 108
layout, 82, 279
~ description, 410
~ implementation, 303, 328
~ level, 280, 290
~ process, 187
LDD, 71, 72
lead frame, 460
leakage
~ current, 74, 178, 241, 341, 505, 511
~ power consumption, 339, 340, 511
LEAP, 358
lifetime, 93, 96
Light-Induced Voltage Alteration analysis, 494
lightly doped drain, 71
line
~ capacitance, 427

~ delay, 433
~ resistance, 426
linear region, 18, 19
liquid crystal technique, 490
lithography, 82
LIVA analysis, 494
load
~ elements, 131, 132
~ characteristics, 132
~ transistor, 133, 134
Local Oxidation of Silicon, 92
Locally Sealed LOCOS, 93
LOCOS, 92, 107, 112
logic
~ block, 276, 311
~ -gate level, 286
~ implementation, 183
low-end IC market, 273
low-power, 337-384
~ library, 356
low-voltage design, 351
LPCVD, 95

macro cell, 276, 311
majority charge carrier, 10
making or breaking techniques, 495
Manhattan skyline effect, 327
mapping, 292
mask, 82, 188
~ -programmable ROM, 257, 309
master cell, 313
matching of transistors, 435, 506
MCM, 470
mechanical probing techniques, 448
meet-in-the-middle strategy, 327
mega cell, 276
memory
~ access time, 234, 239, 246
~ address, 233
~ array, 231
~ bank, 236, 249
~ cell, 231-263, 272
~ matrix, 231
~ word, 233
merged memory logic, 265
metal gate process, 4, 10, 109
METAL mask, 108, 109, 189, 194
microcode instruction, 309
microcontrol unit, 281
microprocessor core, 293
military specifications, 4
minimum feature size, 85
minority carrier, 17
MISR, 478
mixed analog/digital circuit, 417-419
MLC, 262
MML, 265
mobility, 2, 31, 54, 57
~ degradation, 55
~ reduction factor, 57
modelling small-channel effects, 65
module generator, 326
molybdenum, 4
~ gate, 109
MOS, 1
~ capacitance, 10, 36, 39
~ diode, 49, 393
~ formulae, 22
~ transistor, 5, 30, 60, 73, 74, 218
multi-chip module, 470
multi-project wafers, 86
MultiLevel Cell, 262
multilevel flash memory, 262
Multiple Input Signature Register, 478
multiple threshold CMOS, 342
Murphy's law, 122

n-channel MOS transistor, 30
n-tub CMOS process, 112, 130
n-type silicon, 3
n-well
~ CMOS process, 112, 130
~ contact, 194, 195, 198
NAND-gate flip-flop, 176
narrow-channel effect, 63, 64
netlist, 274, 286

nMOS
~ circuit transient behaviour, 140
~ mostly, 130, 176
~ process, 107
~ transistor, 4, 150, 156
nMOSt, 4, 150, 156
noise, 444, 515, 521
~ immunity, 184
~ margin, 131, 166, 444, 521
non-overlapping clocks, 175, 179, 402, 403
non-recurring engineering costs, 301
non-saturated enhancement transistor, 134, 142
non-volatile memory, 231
non-volatile RAM, 263
normally-on transistor, 31
normally-off transistor, 31
NPLUS mask, 188
NRE costs, 301
number representation, 363
numerical aperture, 85
NVRAM, 263
NWELL mask, 188

one-time-programmable EPROM, 259
optical beam analysis, 494
OR
~ -function, 146
~ matrix, 308
OTP EPROM, 259
outer lead
~ bonding, 460
~ frame, 460
output
~ buffer, 186
~ enable, 236
~ impedance, 25
~ protection, 123
~ resistance, 30, 60, 218
oxidation, 87, 91

p-channel MOS transistor, 30
p-type substrate, 4
package
~ cavity, 458
~ corrosion, 476
packaging, 453
page mode, 247, 248
PAL, 309
parallel multiplier, 284
parallelism, 348
parasitic
~ capacitances, 122
~ MOS transistor, 32
~ thyristor, 222, 386
partial product, 284
pass transistor, 170, 210, 211
pass-transistor logic, 171, 356
passivation layer, 109
PCM, 452
PECVD, 95
PEM, 492
penetration depth, 98
performance of CMOS circuit, 183, 510, 511
periodic system of elements, 8
Perovskite crystals, 253
PGA, 413
PGDW, 451
phase-locked loop, 408
photolithography, 82
photon emission microscopy, 492
photoresist layer, 87
physical design aspects, 503
picoprobing, 486
pillar, 121
pinch-off
~ point, 21
~ region, 59
pinhole, 479
pinning assignment, 418, 419
pipelining, 348
PLA, 308, 309
placement and routing, 310, 327
planar
~ DRAM cell, 243
~ IC technology, 33
~ silicon technology, 4

planarisation, 100
plasma, 95
  ~ etching, 90
PLD, 277
PLL, 408
plug, 114, 121
pMOS transistor, 31, 32, 150, 151, 157
Poisson's law, 11
POLY mask, 108, 109, 188
polycide, 120
polycrystalline silicon, 4, 82
polygon pusher, 303
polymide film, 460
polysilicon gate, 108
Polysilicon-Encapsulated LOCOS, 93
positively-charged ion, 7
power
  ~ device, 459
  ~ dissipation, 158, 183, 337-384, 463
  ~ MOSFET, 209, 217, 220
  ~ reduction techniques, 337-384
  ~ transistor, 217
  ~ -delay product, 31, 345
  ~ -down mode, 374
PPLUS mask, 188
pre-deposition, 97
probe
  ~ card, 442, 478
  ~ station, 442
process
  ~ control module, 452
  ~ cross-section, 106, 114, 197
product term, 308
production disturbances, 448
programmable
  ~ array logic, 309
  ~ logic array, 308
  ~ logic device, 277, 315
  ~ read-only memory, 258
PROM, 258
propagation delay, 406, 415, 433, 507, 508
protection circuit, 185, 393
prototyping, 312, 315
pseudo-nMOS logic, 156, 169, 347
pseudo-static RAM, 233
punch-through, 66, 119
p-well, 118, 119

R-load SRAM cell, 240
race, 175
RAM, 231
Rambus DRAM, 247
random access, 233
  ~ memory, 231, 235
$\overline{RAS}$, 245
ratioed logic, 133
RC delay, 433, 507
RDRAM, 247
reactive ion etching, 90
read-only memory, 231, 254
recombination, 37, 56
reduced voltage swing, 354
redundancy, 252
refresh
  ~ amplifier, 242
  ~ operation, 243
register-transfer language, 283, 293
reliability, 386-398, 459, 475, 512-514
repair techniques on first silicon, 495
repeaters, 508
resistive load, 139
reticle, 84
retrograde
  ~ profile, 98
  ~ -well, 118
reuse, 277, 293, 315, 408, 504
reverse short-channel effect, 62
roadmap, 503
ROM, 231, 254
  ~ layout, 306
  ~ logic function, 306
ROR, 245
routing, 310, 327
  ~ channel, 312
row
  ~ decoder, 235

  ~ refresh, 243
RTL, 283, 293
  ~ description, 283

sacrificial gate oxide, 107
SACVD, 95
safe operating area, 217
salicide, 120
saturated enhancement transistor, 133, 141
saturation
  ~ current, 21
  ~ region, 18, 21
  ~ velocity, 57
scaling, 505
  ~ consequences, 510, 512, 514
  ~ effects, 503
  ~ limitations, 519
  ~ properties, 119
scan
  ~ chain, 446
  ~ flip-flop, 434
  ~ -test, 315
Scanning Electron-beam Microscopy, 493
SCCD, 210
scheduling, 292
scratch-protection layer, 109, 476
scribe lane, 453
SDRAM, 247
sea-of-gates, 313
sea-of-transistors, 314
secondary ion mass spectroscopy, 495
self-aligned
  ~ drain, 82, 108
  ~ salicide, 120
  ~ source, 82, 108
  ~ source/drain implantation, 113
self-inductance, 413, 463
self-test, 311, 446
self-timed circuits, 368
SEM, 493, 495
semi-custom IC, 278
semiconductor
  ~ doping, 8
  ~ material, vi
sense amplifier, 236, 243
serial memory, 231, 234
serial ROM, 258
set-up time, 415
shadow RAM, 263
Shallow-Trench Isolation, 93, 114
sheet resistance, 108, 139
shift register, 178
shift-and-add operation, 284
Shmoo plot, 483
short-channel effect, 61
short-circuit dissipation, 158, 159, 186, 339
SIA roadmap, 434, 503, 522
side wall capacitance, 428
sign-magnitude notation, 364
signal
  ~ interference, 427
  ~ integrity, 398-437, 514-519
  ~ processor, 279
  ~ propagation, 406, 415, 433, 507, 508
signature, 478
silicide, 100, 120, 198
silicon
  ~ atom, 5
  ~ crystal, 6
  ~ dioxide, 2
  ~ -on-anything, 124
  ~ -on-glass, 124
  ~ -on-insulator CMOS, 122, 389
  ~ -on-sapphire, 123, 389
SIMOX, 123, 389
SIMS, 495
simulation, 293
single-edge triggered flip-flop, 373
single-phase clocking, 184, 399
SiO_2, 2
slack borrowing, 407
sleep mode, 374
slew-rate control, 416
SLI, 317
slurry, 101
small-channel effect, 61-65

SMD package, 454
SMIF environment, 449
SOA, 124, 217
SOC, 264, 291, 398, 503
soft
~ core, 276
~ error, 251, 518
SOG, 100
SOI-CMOS, 122, 389
SOS-CMOS process, 123
source, 4, 72
~ -synchronous timing, 407
specification, 436, 481
Spin-On-Glass, 100
spurious transitions, 371
sputter
~ deposition, 96
~ etching, 90, 96
SRAM, 232, 235
~ memory cell, 239-241
SRPL, 358
stacked capacitance cell, 244, 266
stand-alone memory, 231
stand-by
~ current, 74, 340, 342
~ mode, 241, 263, 341, 374
~ power, 74, 264, 340
standard
~ cell, 310, 356
~ commodities, 278
~ IC, 301
~ logic IC, 278
~ product, 278
static
~ CMOS circuits, 155, 167
~ CMOS flip-flop, 174-176
~ column access, 246
~ drain feedback, 57, 59
~ memory, 232
~ power consumption, 339
~ RAM, 232, 235
~ RAM cells, 239
STC, 244
steady-state current, 445
step-and-repeat lithography, 84
step coverage, 96
stepper, 84-87
STI, 93, 114
stick diagram, 192
storage
~ gate, 210, 211
~ time, 212
stuck-at
~ fault, 444
~ -one fault, 444
~ -zero fault, 444
stud, 121
sub-threshold
~ behaviour, 74
~ current, 74, 178, 241, 340, 505, 511
~ leakage, 74, 178, 340, 505, 511
~ region, 74
substrate
~ contact, 195
~ dope, 5
~ current, 67, 69
~ resistance, 388
supply bounce, 412, 420, 515
surface mount
~ area array packages, 454
~ dual/quad packages, 454
surface
~ scattering, 56
~ states, 212
~ -channel CCD, 210
Swing Restored Pass-Transistor Logic, 358
switching
~ ($\mathrm{d}i/\mathrm{d}t$) noise, 412, 416, 459, 515, 521
~ activity, 339, 361, 422
symbolic layout, 329
synchronisation, 407-409, 415
synchronous
~ CMOS circuits, 173
~ DRAMs, 247
synthesis, 292, 294
system design aspects, 279, 291, 503
system on chip, 291, 293, 398
systems on silicon, 293, 317

τD-product, 31, 345
TAB, 413, 460
tailored driver switch-on network, 416
Tape Automated Bonding, 413, 460
tapering factor, 163, 164, 346
TEM, 495
temperature
~ expansion coefficient, 476
~ sensitivity, 475
~ variation cycle, 476
~ -cycle test, 476
TEOS, 95
testability, 442, 444
test vectors, 444, 477
thermal
~ energy , 7
~ generation, 37
~ oxide, 91
~ resistance, 463
thermocompression step, 460
thick oxide, 91, 107
threshold voltage, 15, 28, 150, 480
~ adjustment implantation, 5, 107
~ loss, 134, 170
~ roll-off, 62
~ temperature dependence, 55
through-hole package, 454
thyristor, 386
tie-off cell, 423
tiles, 105
time stealing, 407
timing
~ margins, 406
~ problems, 398-407
~ verification, 448
titanium, 120
~ nitride, 120
top-down design process, 279
transconductance, 3, 30
transfer
~ efficiency, 212
~ gate, 170, 210, 211, 356
transient response, 140
transistor matching 435, 506
transition region, 20
transmission gate, 170, 173
transparency, 175, 179
trench capacitance cell, 244, 266
tri-state
~ buffer, 186
~ buses, 429
triode region, 18
triple-well device, 341, 505
TTL input, 185, 414, 480
tunnelling, 93
turn-around time, 258, 274, 315
twin-well CMOS process, 118
two's complement notation, 364
two-phase clocking, 402

under-etch, 89, 448
usable gates, 278
user-specific integrated circuit, 274
USIC, 274
utilisation factor, 278, 422

valence
~ band, 6
~ electron, 6
VCO, 408
VDMOST, 219
velocity saturation, 57
Verilog, 293
version management, 437
vertical
~ DMOS transistor, 219
~ electric field, 56
VHDL, 293
video memories, 234
video RAM, 234
virtual component, 276
virtually static RAM, 233
voids, 396
volatile memory, 231
voltage

~ drop, 186, 426, 515
~ regulator, 353
~ -controlled oscillator, 408
VRAM, 234

wafer
~ diameter, 448
~ probing, 448
Wallace tree multiplier, 284, 372
waveform measurements, 448
weak-inversion behaviour, 73
wearout, 476
well resistance, 388
wet-etching, 89
wire bonding, 458
word line, 235
work function, 15
worst-case delay path, 290
write enable, 236

x-decoder, 235

y-decoder, 236
yellow room, 449
yield, 448
~ control, 452
~ degradation 448,449

zero-resistance method, 391