Philips Research Book Series Volume 5

Amlware

Hardware Technology Drivers of Ambient Intelligence

Edited by
Satyen Mukherjee, Emile Aarts,
Raf Roovers, Frans Widdershoven and
Martin Ouwerkerk

Springer

AmIware

Philips Research

VOLUME 5

Editor-in-Chief
Dr. Frank Toolenaar
Philips Research Laboratories, Eindhoven, The Netherlands

SCOPE TO THE *'PHILIPS RESEARCH BOOK SERIES'*

As one of the largest private sector research establishments in the world, Philips Research is shaping the future with technology inventions that meet peoples' needs and desires in the digital age. While the ultimate user benefits of these inventions end up on the high-street shelves, the often pioneering scientific and technological basis usually remains less visible.

This 'Philips Research Book Series' has been set up as a way for Philips researchers to contribute to the scientific community by publishing their comprehensive results and theories in book form.

Ad Huijser

AmIware

Hardware Technology Drivers of Ambient Intelligence

Edited by

Satyen Mukherjee
Philips Research North America, New York, USA

Emile Aarts
Philips Research Laboratories, Eindhoven, The Netherlands

Raf Roovers
Philips Research Laboratories, Eindhoven, The Netherlands

Frans Widdershoven
Philips Research, Leuven, Belgium

and

Martin Ouwerkerk
Philips Research Laboratories, Eindhoven, The Netherlands

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4197-7 (HB)
ISBN-13 978-1-4020-4197-6 (HB)
ISBN-10 1-4020-4198-5 (e-book)
ISBN-13 978-1-4020-4198-3 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

Printed in the Netherlands.

Contents

Contributing Authors

Ronald M. Aarts
Philips Research Eindhoven
ronald.m.aarts@philips.com

Anteneh Abbo
Philips Research Eindhoven
anteneh.a.abbo@philips.com

Joerg Appenzeller
IBM Research
joerga@us.ibm.com

Jessy Baker
University of California at Berkeley

Srinivasan Balakrishnan
Philips Research Eindhoven
srinivasan.balakrishnan@philips.com

Kees van Berkel
Philips Research Eindhoven
Eindhoven University of Technology
kees.van.berkel@philips.com

Eric Beyne
IMEC, Leuven
eric.beyne@imec.be

Neil C. Bird
Philips Research Eindhoven
neil.bird@philips.com

Erik Bodegom
Department of Microelectronics/DIMES, Delft University of Technology
bodegom@pdx.edu

Benton H. Calhoun
University of Virginia
bcalhoun@virginia.edu

Eugenio Cantatore
Philips Research Eindhoven
eugenio.cantatore@philips.com

Eric Carleton
University of California at Berkeley

Anantha Chandrakasan
Massachusetts Institute of Technology
anantha@mtl.mit.edu

Yuen-Hui Chee
Berkeley Wireless Research Center, University of California, Berkeley

P. Christie
Philips Research Leuven
p.christie@philips.com

Stephane Donnay
IMEC, Leuven

G. Doornbos
Philips Research Leuven
gerben.doornbos@philips.com

Gerhard Fettweis
Technische Universität Dresden
fettweis@ifn.et.tu-dresden.de

Bert Gyselinckx
IMEC, Leuven
bert.gyselinckx@imec-nl.nl

Rick Harwig
Philips Research Eindhoven
rick.harwig@philips.com

A. Heringa
Philips Research Leuven
anco.heringa@philips.com

Chris Van Hoof
IMEC, Leuven

Henk van Houten
Philips Research Eindhoven
henk.van.houten@philips.com

Ralf Irmer
Technische Universität Dresden
irmer@ifn.et.tu-dresden.de

Richard Kleihorst
Philips Research Eindhoven
richard.kleihorst@philips.com

Joachim Knoch
Institute for Thin Films and Interfaces, Forschungszentrum Jülich
j.knoch@fz-juelich.de

A. Kumar
Philips Research Leuven
aatish.kumar@philips.com

Willem Laflere
ESAT-MICAS, KULeuven
willem.laflere@esat.kuleuven.be

Geert Langereis
Philips Research Eindhoven
geert.langereis@philips.com

Steffen Leonhardt
Helmholtz Institute of Biomedical Engineering, RWTH Aachen University
medit@hia.rwth-aachen.de

Hugo De Man
Katholieke Universiteit Leuven, IMEC
deman@imec.be

Patrick P.E. Meuwissen
Philips Research Eindhoven
patrick.meuwissen@philips.com

Satyen Mukherjee
Philips Research North America
satyen.mukherjee@philips.com

Rick Nas
Philips Research Eindhoven
rick.nas@philips.com

R.K.M. Ng
Philips Research Leuven
ranick.ng@philips.com

V.H. Nguyen
Philips Research Leuven
viet.nguyenhoang@philips.com

P.H.L. Notten
Philips Research Eindhoven, Eindhoven University of Technology
peter.notten@philips.com

Brian Otis
Electrical Engineering Department, University of Washington, Seattle
botis@ee.washington.edu

Denis Petrovic
Technische Universität Dresden
petrovic@ifn.et.tu-dresden.de

Nathan Pletcher
Berkeley Wireless Research Center, University of California, Berkeley

Huifang Qin
Berkeley Wireless Research Center, University of California, Berkeley

Jan Rabaey
Berkeley Wireless Research Center, University of California, Berkeley
jan@eecs.berkeley.edu

Padmakumar R. Rao
Department of Microelectronics/DIMES, Delft University of Technology
p.rao@tudelft.nl

Derk Reefman
Philips Research, Eindhoven
derk.reefman@philips.com

Elizabeth Reilly
University of California at Berkeley

Shad Roundy
LV Sensors
sroundy@lvsensors.com

Curt Schurgers
University of California, San Diego
curts@ece.ucsd.edu

Mike Sheets
Berkeley Wireless Research Center, University of California, Berkeley

Boris Škorić
Philips Research Eindhoven
boris.skoric@philips.com

Martijn F. Snoeij
Department of Microelectronics/DIMES, Delft University of Technology
m.f.snoeij@tudelft.nl

Michiel Steyaert
ESAT-MICAS, KULeuven
michiel.steyaert@esat.kuleuven.be

V. Sundararajan
University of California at Riverside
vsundar@engr.ucr.edu

Albert J.P. Theuwissen
Department of Microelectronics/DIMES, Delft University of Technology
a.j.p.theuwissen@tudelft.nl

Pim Tuyls
Philips Research Eindhoven
pim.tuyls@philips.com

Wim Vereecken
ESAT-MICAS, KULeuven
wim.vereecken@esat.kuleuven.be

Alice Wang
Texas Instruments
aliwang@ti.com

X. Wang
Department of Microelectronics/DIMES, Delft University of Technology
x.wang@tudelft.nl

Marcus Windisch
Technische Universität Dresden
windisch@ifn.et.tu-dresden.de

Paul Wright
University of California at Berkeley

Peter Zillmann
Technische Universität Dresden
zillmann@ifn.et.tu-dresden.de

Preface

The term Ambient Intelligence (AmI) has come to symbolize systems, designed to serve us, that are embedded in our surroundings and are context aware, adaptive and anticipatory. Ambient Intelligence systems have been around for much longer than many of us would imagine. If we look at our surroundings—our homes, offices, automobiles or public places—we might realize that evolving technologies have been making us more comfortable, safe and content for some time. Fire or burglar alarm systems are examples of early AmI technologies. These early systems were followed by basic devices such as motion detectors coupled with lighting controls in homes and offices. Considered elementary today, these early applications are often not classified as AmI due to the limited intelligence involved in these systems.

Typically, the process of application innovation is driven by a two-pronged effort–the application pull from the top and technology capability push from the bottom. High performance and high functionality technologies trigger creative ideas leading to innovative applications. Or, creative minds conceive of new ideas and then seek practical technological solutions. In either approach, a close interaction between the two levels (top and bottom) is conducive to meaningful innovation.

AmI applications can be classified into four broad categories based on the maturity level:

I. Applications that already exist in the marketplace.
II. Applications in the development phase that primarily involve technology integration. Here system requirements are well defined and practical integration is the major challenge.
III. Applications requiring a basic technology breakthrough to accomplish the requirements. The requirements are mostly known but implementation technologies are not fully available.
IV. Applications which are not completely defined but are more in the idea phase. These typically have to do with providing cognitive

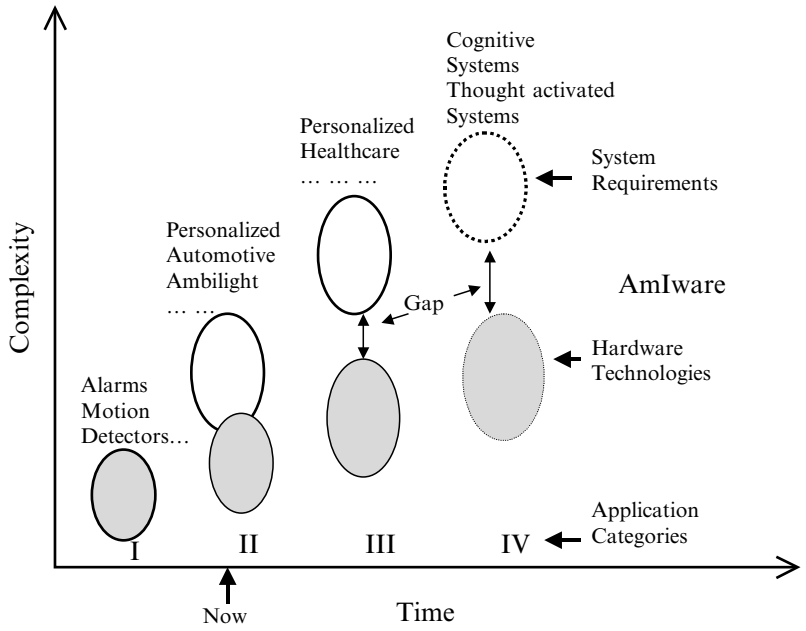


Figure 1. The AmI Application categories and the capability gap.

functionality and the like. The technology requirements are not clearly defined but expectations exist.

From a research perspective the last three categories are of interest. One can map ideas in the appropriate categories to determine requirements and challenges. The number order reflects the degree of challenge which consequently is related to the time frame for practical implementation as depicted in figure-1. Also depicted in the figure is the increasing gap between system level ideas and practical technology capabilities.

The process of top down innovation in AmI has been addressed in numerous conferences, seminars and publications in the past. To complement this effort, a symposium was held to emphasize the bottom-up process to help bridge the gap between the top-down thinkers and the bottom-up innovators. This book is based on the contributions to the symposium "Hardware Technology Drivers of Ambient Intelligence (AmI)", held on Dec 9-10, 2004 at The Koningshof, Veldhoven, the Netherlands. Technical experts from around the world were invited to present developments in various hardware technologies, relevant to AmI. Defining the term AmIware for these hardware technologies, the authors were asked to identify the drivers of their respective technologies with

regard to AmI applications. Key hardware technologies were selected based on their relative importance in AmI applications.

The book begins with a broad overview of the physical layer of AmI applications and is followed by five sections covering key technologies that are relevant to AmI applications:

- Wireless communications
- Smart sensors
- Low power electronics
- Energy supply and management
- Enabling technologies and devices

Papers in each section cover different elements of hardware technologies, describe the state of the art and identify challenges that need to be addressed to bridge the capability gap shown in figure-1. A concluding chapter summarizes and adds to the key hardware challenges that need to be addressed in the future.

It is hoped that this bottom-up effort to identify some of the main hardware technologies (AmIware) and the challenges involved will inspire top-down thinkers to devise meaningful AmI applications and solutions. Bridging the existing gap between the system architects and the hardware technologists will allow us to create truly rewarding consumer applications that will improve our quality of life.

Satyen Mukherjee
Philips Research

Acknowledgements

Aside from the contributing authors of the chapters we would like to thank the management of Philips Research for their inspiration and support in putting together this book. We would like to especially thank Audrey Tobin for her assistance in coordinating the editing process.

Satyen Mukherjee, Philips Research North America
Emile Aarts, Philips Research Laboratories Eindhoven
Raf Roovers, Philips Research Laboratories Eindhoven
Frans Widdershoven, Philips Research Leuven
Martin Ouwerkerk, Philips Research Laboratories Eindhoven

Foreword

Hugo De Man
Professor Katholieke Universiteit Leuven
Senior Research Fellow IMEC

Ambient intelligence (AmI) refers to the vision of a world in which secure, trustworthy computing and communication will be embedded in everything and every-body. This will create a pervasive, context aware electronics ambient that is adaptive and sensitive to the presence of people. It is all about cooperating electronic systems that will enrich human experience and improve the quality of life in an unobtrusive way.

So far much attention has been paid to application scenarios that often look like dreams in a far future. It is now time to pay attention to the realities of hardware technologies and engineering art needed to realize these dreams. This book is one of the first to fill the gap between scenario dreams and hardware reality. It is a compilation of contributions to an international workshop on state-of-the-art hardware suited for ambient intelligent systems organized by Philips NV in December 2004.

Ambient intelligent systems are very heterogeneous in nature and require the smooth cooperation between a wide variety of hardware technologies. Design of such systems is interdisciplinary by nature and requires an intense dialog between specialists of multiple domains. This book reports on the dialog between such specialists working on hardware issues that fit within the framework of AmI systems.

Personal healthcare and wellness, smart home systems and the progress of communication networks are taken as the system drivers in this book.

These driving applications define the AmI framework, which consists of tether less networking between people and objects, of transducer networks to sense and actuate the ambient and the human body, of interfacing to the human senses, and of wearable or stationary embedded computing to provide intelligence, interpretation and smart interaction when necessary.

Pervasive to all of the above is the extreme requirement on the energy efficiency of all hardware components either because they are nomadic (small size battery or fuel cell powered) or because they are cheap such that cooling cost and packaging must be minimized or, last but not least, because sensors and actuators must be autonomous also from an energy standpoint. This leads to the concept that they must scavenge their energy from the ambient itself. All of this introduces great challenges that cover a wide spectrum of hardware technologies.

Wearable computing and communication systems operating on huge databases of audiovisual data will require "More Moore" i.e. further relentless scaling of digital devices down to the nano-scale dimensions to provide the giga-scale complexity required by AmI computing and data-storage within the energy constraints discussed above, while keeping embedded software programmability to provide flexibility and adaptivity. As all AmI systems have to communicate with the analog nature of human interfaces particular challenges occur to analog electronics in a nano-scaled device world. These "More Moore" issues are covered in the section on Low Power Anaolog and Digital Electronics and Technology.

Next to the above world of "More Moore" there is the world of "More than Moore". This refers to the plethora of hardware technologies on top of CMOS that are needed for interfacing to the real world and to the human body and senses. Sensors are a crucial part of all AmI systems leading to exciting requirements on MEMS, on cheap but ever more powerful optical sensors as well as on biosensors to couple bionics to electronics while no sensor is complete without appropriate packaging technology. These issues are covered in the section on Smart Sensors. But AmI is also about providing experience to people and hence actuators must provide natural light, sound and image experience to people. The section on Actuation provides us with state-of-the art realizations in solid-state lighting, flexible displays and the latest in high quality sound reproduction within an extremely small form factor.

Last but not least, the section on power supply and energy scavenging pays attention to one of the key technologies of nomadic devices and transducer nodes: power supply and power management. Opportunities and limitations on energy scavenging are reported and recent progress in flexible batteries with highly improved efficiencies are reported. This book is one of the first to bring all these hardware aspects together under one cover. It is recommended to all engineers involved in the realization of the AmI dream and is an inspiring source to create a dialog between the many hardware disciplines that must become interoperable to create a working AmI system. It is precisely that global system aspect that makes AmI such

an exciting discipline that promises to have a great impact on society and on economical progress.

Leuven, March 2005

Section 1

Introduction

Chapter 1.1

AMBIENT INTELLIGENCE: THE NEXT WAVE IN CONSUMER ELECTRONICS

Rick Harwig
Philips Research Eindhoven
rick.harwig@philips.com

The history of consumer electronics has been marked by major advances in each of the past three decades, the 1970s, 1980s, and 1990s. Our present era (2000–2010) holds the promise of even more exciting technological progress. The seventies (1970–1980) could be called the era of the ''hardware wave'' when major consumer electronic (CE) devices, such as TVs and VCRs, were primarily hardware-based and product differentiation came from component technologies and system architectures. As component technologies matured and became commoditized, software became the key factor in product innovation, contributing to differentiation in device applications such as the CD player and the PC. Thus, 1980–1990 could be called the ''software decade.'' With the nineties (1990–2000) came the ''network wave'' when markets became flooded with devices such as the mobile phone. The Internet took off in a major way and the world became increasingly more connected through all kinds of information systems. The networks and their performance were the differentiators. The present era (2000–2010) belongs to a new class of technologies called ambient intelligence (AmI). AmI, along with commodity hardware, standardized software, and networks that are becoming ubiquitous, is transforming the way that we utilize and experience traditional products and services. AmI will manifest itself in many ways and affect our lifestyle in a fundamental manner. Characterized by invisible electronics embedded in our environment (ambient) the AmI system will be capable of automatically responding to our needs and even desires (intelligence). This vision has been embraced by the European Union with a 3.6 billion Euro ICT program until 2006.

S. Mukherjee et al. (eds.), AmIware, 3–7.

The need for AmI can be traced back to the dark ages when people were at the mercy of nature with little control over their environment. Improvements came gradually until the era of broadcast radio and television which represented a major leap toward allowing ready access to information. Today with mobile phones and the internet, information access is unrestricted in space and time. In the coming era AmI will bring us a step further in enhancing our lives and simplifying control of our environment.

From the consumer electronics point of view the most striking change in the living room between the last century and the next will be the disappearance of the entertainment boxes (TV, VCR, DVD player, and PC) and the appearance of an ambience (characterized by unobtrusive displays and sound reproduction systems) that surrounds the user in a simple manner but makes a profound difference in the user's overall living experience (Figure 1.1–1).

Philips has been engaged in the revolutionizing technology of AmI for some years now and has introduced a number of products to the market. The ambilight TV, for example, automatically adjusts the ambient light around a TV to extend the image viewing experience beyond the boundaries of the physical display as shown in Figure 1.1–2.

Electronic paper is another example of AmI where information is made conveniently available everywhere in a flexible manner. In the healthcare domain an example of particular interest to the physician is the 3D image rendering of vital organs, such as the human heart in real time as shown in Figure 1.1–3.

There are a number of key hardware technologies in the semiconductor arena that contribute to AmI applications. Important among them is the systems-in-package (SiP) that allows multiple functions to be integrated in one component module for a variety of applications. The celebrated Moores's law has allowed increasing functionality in system-on-chip (SoC) using digital CMOS technologies and primarily addressing computation applications. With the addition of components such as RF circuits, high voltage circuits, different types of sensors and even microfluidics and solid state light sources in the same SiP, implementation of AmI systems is moving closer to reality. The range of technologies and devices that can be integrated in a SiP are shown in Figure 1.1–4.

The near field communication (NFC)-enabled mobile phone is one of the recent applications introduced by Philips for a broad range of uses including mobile banking, money transactions, building access, public transport access, information display anywhere on the move, micropayment, and e-business cards as shown in Figure 1.1–5. This product involves a concerted effort between Philips, Sony, and Nokia.

Figure 1.1–1. The change in the living room from the last century (above) to the next (below) enabled by ambient intelligence (AmI).

AmI now represents embedded systems that are context aware, personalized, adaptive, and anticipatory to our needs and desires. Several institutions and companies around the globe are engaged today in various aspects of this exciting development, and a technology ecosystem is developing to address this field. Hardware advances continue to be one of the major drivers towards commercialization.

Figure 1.1–2. Philips ambilight TV.

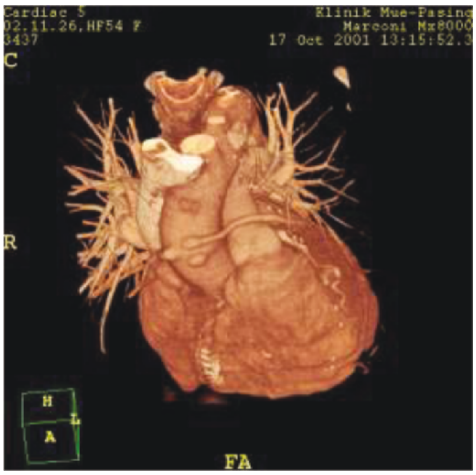


Figure 1.1–3. 3D image rendering (of the human heart) for professional and consumer applications.

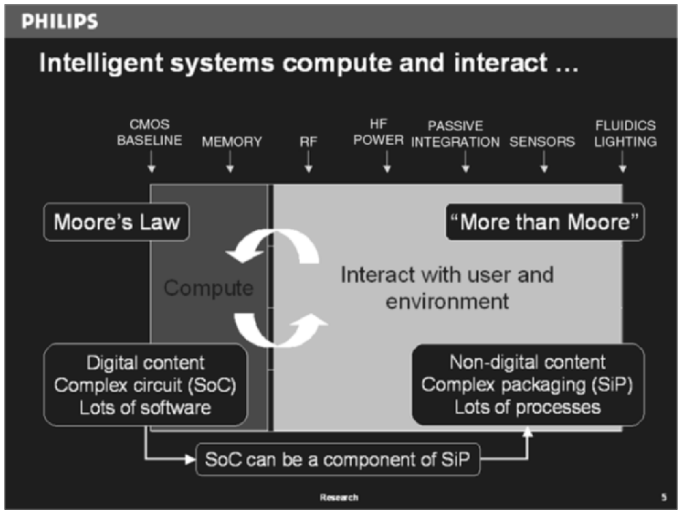


Figure 1.1–4. The domains of SoC and SiP.

Figure 1.1–5. NFC-enabled mobile phone and various application opportunities.

Chapter 1.2

THE PHYSICAL BASIS OF AMBIENT INTELLIGENCE

Henk van Houten
Philips Research Eindhoven
henk.van.houten@philips.com

Abstract Ambient intelligence (AmI) refers to future digital environments that are sensitive and responsive to the presence of people. This chapter addresses some challenges at the device level that need to be overcome to be able to realize a physical layer for AmI. As we will show, the physical layer of AmI cannot be realized by Moore's law type of innovation alone. The nontraditional functionality and form factors of the required devices call for heterogeneous integration and miniaturization approaches. The key enabling technologies are system-in-package(SiP) and large-area electronics. The soaring complexity of innovation in this field calls for open innovation approaches.

Keywords ambient intelligence; autonomous microdevices; displays; large area electronics; open innovation; sensors and actuators; solid state lighting; system-in-package

1. INTRODUCTION

Ambient intelligence (AmI) refers to future digital environments that are sensitive and responsive to people [1]. We note that many other companies and institutes are studying concepts similar in spirit to AmI [2]. The concept of AmI has been studied primarily from a systems perspective, combining visionary insights from social science and computer science, and stressing the new experiences that will be enabled by future technologies. The AmI vision can be applied to very diverse application environments, varying from homes, offices, or cars to homes for the elderly and hospitals. It refers to all human senses: sight, hearing, touch, smell, and perhaps even taste. It also refers to a wide range of human

S. Mukherjee et al. (eds.), AmIware, 9–27.

emotional and intellectual needs, from comfort, pleasure, and entertainment to safety, security, and health. In this chapter, we present an overview of the challenges that this vision poses to innovation in the field of devices and microsystems, which constitute the building blocks for the physical layer of AmI.

As we will discuss, the physical layer of AmI cannot be realized by Moore's law type of innovation alone. Next to mainstream system-on-chip (SoC) solutions for signal processing, solid state storage, and control, we expect that the nontraditional functionality and form factors of the required devices for ambient intelligent environments will require heterogeneous integration and a variety of miniaturization technologies. This inspires new paradigms in microelectronics, such as system-in-package (SiP) solutions, large area electronics, and textile electronics.

The SiP concept applies to quite diverse technology and application areas, ranging from sensors and actuators, RF modules for mobile communication devices, solid state lighting to health care devices, such as biosensors or cardiac pacemakers. The challenges include miniaturization and special packaging, allowing for a variety of input–output functionalities. Autonomous wireless sensor nodes, accepted to be essential components of ambient intelligent environments, pose the additional challenge of operation without having to replace batteries regularly, which requires ultra low power operation, and energy scavenging techniques.

Large area electronics has been developed in particular with active matrix displays as application carrier, and with amorphous or polycrystalline silicon on glass as the technology of choice. Polymer electronics is an emerging technology, which will allow large area flexible devices to be manufactured on plastic substrates, eventually using roll-to-roll technology. In addition to flexible displays, applications may include low cost RF tags, solar cells, or electronics embedded in paper. Light emitting polymers are being explored for display applications, but increasingly also as large area light sources. We predict that a cross fertilization will occur between the display and the lighting application domains, leading to new products such as pixel lamps. These can be used to generate special lighting effects, and novel user interfaces.

Finally, it will be a characteristic of ambient intelligent environments where electronics will be integrated in almost everything, including soft or deformable objects, such as furniture, curtains, or clothing. This field is known as textile electronics, which can be approached in quite different ways. One approach is to make functional fibers, which are subsequently woven into a fabric. Another is to integrate discrete electronic building blocks (made by SoC, SiP, or large area electronics technologies) into textile fabrics.

In the following sections, the physical layer of AmI is discussed while focusing on examples and concluding with remarks on the necessity of open innovation approaches in order to manage the soaring complexity associated with this field.

2. THE ROOTS OF AMBIENT INTELLIGENCE

The concept of AmI builds upon the pioneering work by Mark Weiser on ubiquitous computing. A computer has already progressed enormously from the 1970's powerful mainframe, enshrined in a dedicated computing building, to the present day's almost equally powerful personal desktop computers and portable notebooks. The usage paradigm has remained essentially the same: input of data through typing on a keyboard, and output of data through a display. Weiser's vision was to make computers effectively *invisible* to the user, by having them distributed throughout the physical environment [3]. The display remains as the device to interact with.

Weiser describes three such devices, ranging from a wall-sized interactive *board* (like an office whiteboard), a much smaller note *pad*, and really small devices called *tabs*. The tabs are of the size of the text on the spine of a book, or a post-it note. In a typical environment, the user would interact with one or two boards, tens of pads, and a swarm of tabs.

The essential point is that these computing devices should be unobtrusive: their size is determined by the display, and the rest is invisible. Many more unobtrusive devices are conceivable that do not have a display, for example sensor and actuators. These can be even smaller, leading to the concept of "smart dust" [4].

A second major pillar behind the concept of AmI is the work on intelligent social user interfaces by Nass and Reeves [5]. They wrote an influential book "The Media Equation", subtitled "How people treat computers, television, and new media like real people and places". Nass and Reeves found that human attributes like emotion, and aspects of human behavior like politeness, equally apply to the interaction of man with modern machines. Designing an effective user interface calls for an understanding of such psychological phenomena.

3. THE VISION OF AMBIENT INTELLIGENCE

Philips has defined AmI as *digital environments that are sensitive and responsive to the presence of people*. It is not a purely technical vision but a people oriented vision. The emotional dimension is crucial. In this sense it

can be seen as a marriage of the unobtrusive computing world of Mark Weiser, and the sociological vision of human–media interaction of Nass and Reeves.

Putting emphasis on the emotional aspects is fully in line with an understanding of how the economic added value progresses through the stages of extracting commodity goods or materials (such as silicon or glass) to producing finished products out of the basic materials (such as an IC or a compact disc player) to delivering content or services (like music on a CD or broadcasting a movie) to staging experiences (like attending a live opera at La Scala in Milan). Putting it in a different way—people are prepared to pay substantially more for a cup of coffee at the Champs Elysées in Paris than at the office coffee machine. This is why in commercials a photo camera is not presented as a box one can shoot pictures with but as a tool to capture valuable or dear memories.

In a world where computing devices are ''invisible,'' because they disappear into the background, it becomes even more important for companies to find innovative ways to capture the value. This must be done through selling experiences rather than products.

This vision requires electronic functions that are seamlessly integrated with the home, car, or office environment. These functions will be integrated in lifestyle products. Appliances will be smart, forming self-configuring networks. Wearable electronics, artificial intelligence, and sophisticated user interface functionality will be required.

Studies of human behavior will have to reveal whether people actually like to give orders to a picture on the wall, whether they would like to have their friends being ''telepresent'' while they are watching a soccer game in the living room. Will people appreciate having a wearable personal assistant that continuously monitors a number of biological functions, such as blood pressure or heart rate variability? Will they want the music and lighting atmosphere in their living room to adapt automatically to their unarticulated preferences? Philips Research has recently opened a ''Home Lab'' where such questions are being addressed by interdisciplinary teams through carefully designed and monitored experiments.

What does it take for an environment to be truly ''ambient intelligent''? A key requirement is that many invisible or unobtrusive devices should be distributed throughout the environment. These should be *context aware*, in that they should know about their situational state. These devices should also be *personalized*, so that their function is tailored towards a specific user's needs. They should be *adaptive*, able to learn and recognize people. Ultimately, they should be *anticipatory*, wherein the user's desires are anticipated without the need for input by commands.

The scientific community tends to address these requirements from the perspective of the higher architectural layers of the system. Indeed, here one may find many interesting computer science challenges, such as data and content management, and dealing with data stored onto many different storage devices incorporated into a network. There is a strong link with connectivity problems, such as pervasive wireless networks, ubiquitous communication, or peer-to-peer computing. At a more subtle level, issues such as trust and privacy are very important, with associated disciplines, such as digital rights management, encryption, and biometric identification. User interface technologies, such as speech recognition and synthesis, or computer vision will be highly relevant as well. The ultimate system level challenge will be to design algorithms and programs for computational intelligence and contextual awareness.

4. PHYSICAL LAYER CHALLENGES

Much less attention has been devoted so far to the physical layer of AmI, probably because of the implicit assumption that the Moore's law type of progress will provide all the technology that will be required. Weiser has mentioned some generic hardware issues: ultra low power computing, low-power small range communication (tiny cellular systems, with cells measuring just a cubic meter), and the realization of interaction devices, such as special pens for his digital whiteboards.

However, as discussed in this chapter, the physical layer of AmI is much richer than this. For example, we will need flexible displays that can be integrated into a variety of objects. We will have to invent soft controls and wearable displays for user interfaces integrated into garments or furniture. Smart RF identification tags, able to measure a variety of environmental parameters, will have to be incorporated into objects. A range of sensor and actuator functions will be required, with CMOS-enabled intelligence, integrated in miniature packages. These will also have to have a wireless communication function. And ideally, they would need to "scavenge" their energy from wherever and what is available in their environment. In the following paragraphs, a number of such physical layer challenges are discussed, without any claim to completeness.

4.1. Displays Everywhere

In an ambient intelligent world, displays will be virtually everywhere. First of all, we will need large flat screens integrated in the walls of our living rooms or on our desks. For this purpose, liquid crystal displays with thin film

transistors are the current winning technology. LCD panel sizes are increasing steadily, and costs are coming down. Many of the performance issues, such as a limited viewing angle have been resolved. LCD panels are rapidly penetrating in the monitor market, and LCD-TV will be the next wave.

Manufacturing larger and larger LCD panels is requiring huge capital investments. It is intriguing to think about new ways to manufacture LCD panels, with the aim of achieving a cost breakthrough. For example, one may think of printing technologies to define the active matrix. It has been demonstrated by Philips Research that displays can be made using a bottom-up technology. Starting with a single substrate containing a pattern of interdigitated electrodes, this so-called stratified LCD technology [6] allows the formation of LCD cells using photoinduced phase separation and polymerization. Such a technology would eliminate one of the very large glass plates, and the associated complex handling issues.

LCD technology will clearly be an enabler for the realization of the vision of ubiquitous displays. A recent example is the so-called ''Mirror TV'' launched in 2003 by Philips. By integrating an LCD display behind a semitransparent mirror, display functionality is added to the bathroom mirror. This enables entirely new ambient intelligent applications.

LCD panels have various limitations as well. For example, they require a backlight. An emerging emissive display technology that does not require a backlight and offers very high quality color images, are active matrix panels with light emitting organic materials. There are currently two routes being pursued: conjugated polymers and oligomers. The small molecules require deposition by evaporation in vacuum, which is a relatively expensive method. The polymers are soluble and can be processed

Figure 1.2–1. A rollable display using a polymer electronics active matrix.

by spin coating or ink-jet printing. Full color inkjet-printed polymer-matrix displays suitable for high quality television applications have been shown to be possible. With emissive polymers, new types of displays can be made in principle, such as dual sided displays, or a transparent display integrated onto a window's pane.

Eventually, one would like to develop a reflective display technology that would offer high contrast images even under conditions of bright sunlight. Both LCD and polymer LED displays suffer from poor daylight contrast. A contender in the race towards developing real "electronic paper" is the electrophoretic display technology that is being developed jointly by e-ink and Philips. The principle is based on electrophoretic separation of white and black particles that are suspended in a liquid cell. It has now been demonstrated that gray scales are possible using sophisticated electrical switching strategies. The realization of full color electronic paper remains a challenge. A recent contender in this race is the electrowetting display, invented at Philips Research Laboratories [7].

One would like to have flexible displays as well, ranging from weakly curved displays that can be made conformal to the shape of an object or that can be integrated onto non-flat surfaces, such as car dashboards to displays that can be rolled up into a cylindrical container to paper like displays. Flexible displays are being investigated based on many of the known flat display principles. The manufacturing of the active matrix is one of the real challenges. LCD and e-ink panels can be switched with relatively low currents, and can probably be based on an amorphous silicon active matrix on flexible substrates. Organic LED displays demand larger driving currents, and will probably require a polycrystalline silicon active matrix. Polymer electronics is also a candidate technology for driving flexible displays. A desired breakthrough would be the invention of semiconducting organic materials with a much higher mobility than the conjugated polymers used so far. This would enable the realization of a fully organic display technology optimal for flexible displays. However, already today it is possible to make reflective flexible displays based on an electrophoretic display principle, in combination with a polymer electronics backplane (see Figure 1.2–1, courtesy of Polymer Vision).

4.2 Sensing and Control

AmI calls for a multitude of different sensors to be embedded in the environment and into objects. As stated by Estrin et al, "interfacing to the physical world is arguably the single most important challenge in computer science today" [8].

The main physical layer challenges in this domain are integration and miniaturization, and function extension beyond measuring simple parameters, such as temperature, or pressure.

For known sensor principles, integration and miniaturization are the key challenge. This is needed, because it will be necessary to have very low cost sensors that can be unobtrusively integrated into objects. They will need to be integrated with a certain amount of silicon electronics, for preprocessing the data, for identification purposes, and for adding communication functionality. Microsystems and SiP technology are rapidly emerging as enablers for miniaturization of functions that cannot be realized using mainstream SoC technology. Whereas most analog and digital signal processing functions can now be integrated in standard IC processes, this is not true for many sensor and actuator functions or for discrete components, such as inductors, transformers, capacitors, and antennas, which require specific physical materials properties for their proper functioning.

A wide range of physical parameters is easily accessible to measurement, such as temperature or heat flow. The orientation and acceleration of objects can be measured. Image capture is becoming a standard feature on mobile phones. Miniaturization of mechanical functions, such as zooming and focusing, will call for new types of adaptive optics. Philips has invented a miniature camera, comprising a CMOS image sensor and a tunable lens based on electrowetting (called Fluid Focus). Figure 1.2–2 illustrates various prototypes of this lens, showing the progression from a research demonstrator to a solution that lends itself to mass manufacturing. The phenomenon exploited in electro wetting lenses is the electric field dependence of the contact angle of a liquid at the walls of a container [9].

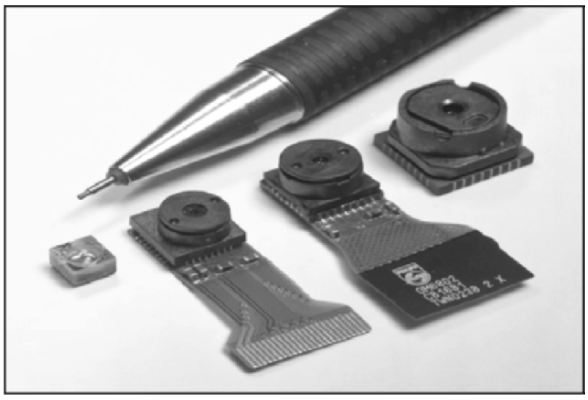


Figure 1.2–2. Successive prototypes of the Fluid Focus lens.

A rugged lens is made by adding a layer of oil on top of a layer of water. The two liquids have a different refractive index, but the same specific density, so that there are no undesired side effects of gravity or acceleration.

Also in the domain of function extension, microsystems technologies are very promising. For example, acoustic MEMS microphones and loudspeaker arrays in combination with phased-array technology and dedicated signal processing will enable miniature adaptive highly directional acoustic systems with obvious advantages for the suppression of background noise in mobile phones for speech recognition applications, and for directed speech output.

The chemical composition of the air in a room will give useful information that can be fed into air purification systems or in security systems (think of the Anthrax scare). Analysis of breath can provide information with a high clinical relevance (for example for patients suffering from asthma). But one can also think of sports coach systems requiring information on the lactate concentration in sweat. These applications will need chemical sensors, or biosensors. Such sensors can be made using a variety of different microsystems technologies, enhanced by surface modification with suitable capture molecules. MEMS like technologies can be added to manipulate gases or liquids.

The multitude of sensors embedded in ambient intelligent environments will have to communicate wirelessly to control systems, and possibly to each other. This calls for low power short-range communication systems, with appropriately designed miniature antennas.

4.3 Autonomous Devices

The ultimate challenge in the field of AmI is to invent technologies for smart and *autonomous* microsystems. Such microsystems would typically need to incorporate sensor and communication functions. For applications where many autonomous devices are distributed throughout the environment, it would clearly be impractical to replace batteries all the time. Autonomous devices would therefore preferably be batteryless, taking their energy from the environment (visible or RF radiation, thermal gradients, vibration ... etc.). This is called energy scavenging.

An attractive practical example of an autonomous microdevice is the wireless wall switch marketed by EnOcean [10]. This "push button" generates 50 μJ of electrical energy by converting the mechanical energy associated with pushing the button on the switch, using a piezoelectric transducer. This is enough to meet the power demands of the electronic circuitry needed to transmit a payload of 250 bits using an RF

link at a bit rate of 100 kbit/s, at an activation lifespan of 2.5 ms (20 mW average power)

Familiar devices that can operate without a battery are the RF identification tags, that are rapidly becoming ubiquitous. These are *passive* devices, wherein the energy is provided from an external RF field generated by the reader device. Such RF devices enable intuitive interaction between devices and users through near field communication. Figure 1.2–3 shows a recently developed paper-thin RF tag made using "Silicon on Anything". In fact, this tag is so thin that it can be integrated into value paper.

Other wireless autonomous microsystems that are on the market today are tire pressure sensors. A practical example of an *active* autonomous microsystem is the quartz wristwatch powered by the irregular motion of the wrist. Also motion powered pace makers have been developed based on similar principles.

Various research laboratories have begun to study the design and system requirements for networks of smart and autonomous microdevices that operate in a collective fashion. The realization of such electronic dust [4] will be essential to make the vision of AmI come true. So far, most groups have focused on system design issues. Prototypes have been based on off-the-shelf components with a limited level of integration. Typically, a standard battery is used to power these early prototypes, and the battery tends to determine the size of the system.

We expect that (wireless) energy supply will remain the key limiting factor in the realization of ambient intelligent environments. Although there are some interesting approaches to integrated microbatteries and

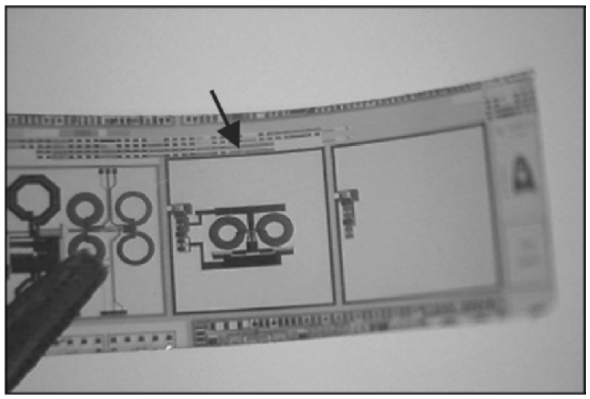


Figure 1.2–3. Paper-thin RF powered tag made using the "Silicon on Anything" technology.

super capacitors, it is fair to say that progress in the energy storage density of batteries is by far not as impressive as that of the digital technologies mentioned above [11].

The amount of power that can be scavenged from the environment is a function of system size and of the strength of the driving forces. Rabaey has provided some estimates of attainable power levels [12] (Table 1.2–1):

From these numbers we can conclude that energy scavenging can provide typically 10–1000 $\mu W/cm^2$ for a device of cm dimensions under normal conditions. It is of interest to contrast this to a typical battery, with an energy density of 400 Wh/l. Assuming continuous operation during 10 years without any recharging, such a battery can provide a power of 4 $\mu W/cm^3$. Of course, this number would rapidly go up in case recharging is allowed or if the operational lifetime could be shorter. For ambient intelligent environments with swarms of smart sensors it will clearly be impractical to exchange or recharge batteries.

Metrics from the field of IC design can be used to estimate what can be done with limited power. We should consider both computation and communication [13]. From the computational point of view, a power range of 10–1000 $\mu W/cm^2$ would enable 1–100 k operations per second for the 0.18 μm CMOS technology node, neglecting leakage power. This would be just enough for low data rate sensor applications. Moore's law type technology scaling would enable successively more complicated computing applications, such as audio processing. Unfortunately, devices optimized for high performance computing have a rather large leakage power, typically 0.01 μW per device. For low complexity sensor applications, with 10000 devices, leakage power alone would already take as much as 100 μW, and this problem gets worse with scaling.

From the communications point of view, we should focus on the short range (on the order of meters). This is the realm of systems such as Bluetooth (1 Mbit/s), Zigbee (80 kbit/s), or PicoRadio (1 kbit/s). From these, only PicoRadio falls in the power consumption range of 10–1000 $\mu W/cm^2$. Typically, the energy required to transmit a single bit over short ranges is in the order of 10–30 nJ. This energy does not scale

Table 1.2–1. Energy scavenging [12].

Energy source	Power density
Vibrations	0.05–0.5 mW/cm^3
Sound pressure (100 dB)	1 $\mu W/cm^2$
Temperature gradient (thermoelectric)	1–10 $\mu W/K/cm^2$
Photovoltaic: direct sunlight	10 mW/cm^2
Photovoltaic: indoor	10 $\mu W/cm^2$

with Moore's law. For the lowest bit rates, standby power (typically 50 μW) is again a limiting factor.

Of course, a lot can be gained by proper system optimization. For example, microsystems can be provided with a "sleep mode", so that they consume power only intermittently, whereas energy scavenging is functional continuously. Yet, it can be concluded safely that with real smart dust type devices relying on energy scavenging or on a single non-replaceable battery, very limited computing and communication is possible. The realization of simple wireless sensor functions will undoubtedly be feasible, but ubiquitous computing and communication based on wireless and batteryless autonomous microdevices is likely to remain a dream.

4.4. Textile Electronics

We are used to thinking of consumer electronics in terms of self-contained appliances, or "boxes." There are circumstances where such boxes are not the most convenient ways to package electronics. Think of a jogger carrying an MP3 player. It is not convenient to operate the tiny controls of such a device while jogging. Sportsmen, as well as elderly people, or individuals with an identified health risk may well want some of their body parameters to be monitored continuously. Examples are the electrocardiogram for sportsmen, or for heart patients, or the glucose concentration in the blood for diabetic patients.

Such cases call for electronics integrated into clothing. A first step could be to integrate controls and sensors in apparel. Systems researchers are already going a step further, and are investigating wireless "body area" sensor networks.

Device physicists and materials scientists need to come up with the enabling technologies for this new field of "textile electronics" [14]. What needs to be invented is soft controls, wiring and interconnect solutions, flexible segmented displays for signaling functions, and integrated sensors. One may also think of apparel woven from light emitting, or color changing fibers. By incorporating suitable fibers, also the textile itself can have sensorial functions, for example to measure the degree of stretch [15].

The application of textile electronics is not limited to personal apparel. Electronics can be incorporated in furniture, or soft furnishings (curtains and wall paper) as well. A research group working at Infineon has developed a concept for "intelligent carpets"[17]. These carpets are woven fabrics containing conducting wires, connecting and powering a two-dimensional array of (conventional silicon) chips. The chips form a self-organizing network: it can be cut into any shape, without losing its functionality. The carpet is provided with sensors, able to monitor

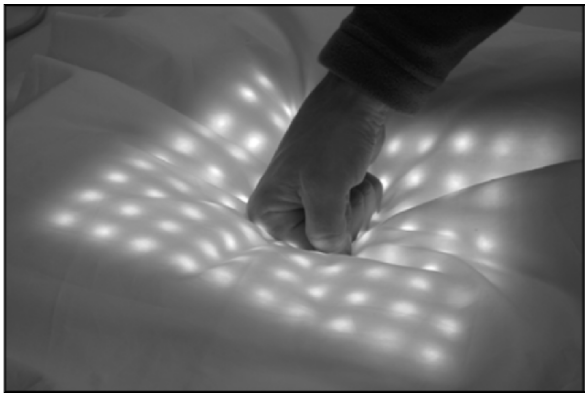


Figure 1.2–4. Textile photonics: LED's integrated into a pillow.

temperature, pressure, or vibrations. This approach may also be used to solve the energy supply problem of "smart dust" type of autonomous microdevices, allowing truly ubiquitous computing.

Energy scavenging looks like a viable option in the domain of wearable electronics. A human typically consumes 2500 kcal a day, or 10 MJ. This corresponds to an average power of about 100 W, which roughly equals the power consumption of a desktop computer. Starner [16] estimated that while walking about 5 W may be recoverable using appropriate transducers. This may be enough for wearable computing.

4.5. Solid State Lighting

Light bulbs have provided the basic technology for the invention of electronics based on vacuum tubes. Yet, illumination as an application domain has remained relatively unaffected by electronics. Exceptions are electronic ballasts for compact fluorescent lamps, and electronic dimmers. However, we can expect that the emerging solid state lighting technologies based on light emitting diodes will change this state of affairs [18]. There are two major solid-state lighting technologies. The first is LEDs based on heteroepitaxial growth of III-V semiconductors that have shown an exponential increase in luminous efficacy, and they now offer very bright point sources. An LED lamp containing red, green, and blue LEDs can be controlled electronically to span a large part of the color triangle. In its most complex manifestation, one may expect that LED lamps of the future will encompass various LED dies, photo detectors and control electronics for active color control, passive or active microoptics for beam control, passive or active thermal management, and a wireless link.

Such lamps will be made using SiP technologies. An early prototype is shown in Figure 1.2–5.

The second emerging solid state lighting technology is based on electro luminescent amorphous organic semiconductors, which can be deposited on large area substrates by evaporation (small molecules) or wet processing technology (polymers) — just as in the organic displays described above. Organic LED light sources will be ideal for large area diffusive light sources (see Figure 1.2–6). They can be flexible, and of any shape. They can be made to emit white light through the appropriate layer composition or red, green and blue pixels can be made in combination with some kind of a matrix-addressing scheme. It is also possible to make transparent light emitting devices on glass, enabling new applications, such as privacy or atmosphere providing windows.

It is expected that these LED technologies will show rapid growth in special lighting markets, followed by a penetration in the general lighting market. This will probably start with halogen lamp displacement. But eventually, one can envision a more dramatic impact. They may not require a socket because of the long lifetime of LEDs. In combination with the increased design freedom in tailoring the shape of organic LEDs, one can anticipate that entirely new types of luminaire will be designed.

Because of the added control over color and illumination intensity and distribution, the distinction between the lamp of the future and a display may become less sharp than we are used to. Perhaps the lamp of the future will be a flat light-emitting surface that shall also function as a touch screen, just like an interactive display. This way one could freely adapt parts of the surface that would emit light. The added degrees of freedom in

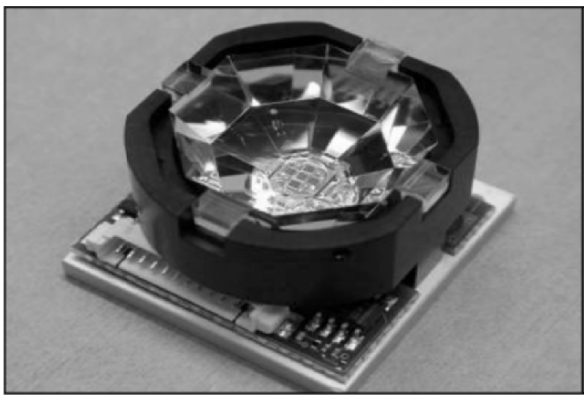


Figure 1.2–5. Color tunable solid state lamps will be based on System–in-Package technology.

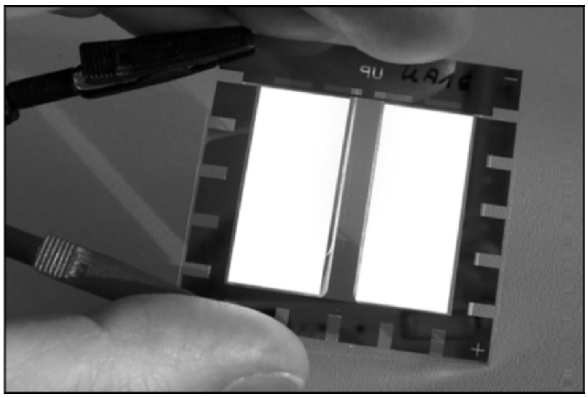


Figure 1.2–6. An organic light emitting diode.

controlling the light output of solid-state lighting will also call for an increased number of (light) sensors and control functions in the environment. Digital control and wireless networks will play a role in this application domain as well. Solutions will eventually be found to provide lighting tailored to the specific mood of the people present. This will call for context awareness. Thus, the expected impact of solid-state lighting fits very naturally with the vision of AmI.

5. MORE OF MOORE AND MORE THAN MOORE

5.1. System on Chip and System in Package

Microelectronics has become the most pervasive global industry with products in almost any application domain, ranging from mobile communications, computing, and automotive electronics to personal healthcare, security and identification. Innovation in mainstream silicon microelectronics is following the path of Moore's law all the way down to the nanoelectronics size regime. In formulating his famous law, Gordon Moore not only made a lucid prediction but also provided a clear direction and time line for innovation in the semiconductor industry, which has enabled the definition of internationally agreed roadmaps. Because of this, the various actors in the industry ranging from materials suppliers to toolmakers, IC manufacturers, and software houses can be fairly confident that their expenditures on R&D are timely, well directed and profitable. On an architectural level, the winning solution is the SoC approach. Based on generic and standardized technology generations, a SoC

provides a solution based on modular IP blocks, which can be designed using well-defined design rules with compact models describing the behavior of the individual devices. Because of its programmability, a SoC can be manufactured in large numbers for a range of applications, offering maximum reuse of costly design and testing effort.

Meanwhile, as described in this chapter, we are witnessing a diverse branching off of progress in microelectronics beyond the one-dimensional highway of Moore's law. Further miniaturization of electronics requires extensive integration of passive components, such as inductors and capacitors, which nowadays take up the majority of PCB area in, for example, a mobile phone. This requires new approaches in materials, design and processing as the traditional scaling rules for CMOS circuits do not apply here. Another trend is the increasing incorporation of a rich variety of novel functions into personal electronic systems. One may think of miniature camera modules, a personal weather station, GPS, accelerometers, biometric identification, and health monitoring systems. Apart from dedicated non-CMOS semiconductor process features, such as high-voltage, low power, analog, and radio frequency devices, non-semiconductor technologies too are needed to realize functions like integrated passives, or mechanical and optical sensors (MEMS). Nanotechnology and biotechnology are around the corner, and they will have a dramatic impact on this new class of sophisticated elements.

At an architectural level, the complexity can be addressed at least partially, by the concept of SiP. Where a SoC minimizes the cost per switch, a SiP focuses on achieving the highest value for a single packaged microsystem. Added value in a SiP is achieved by integration of several functions into a single module or package thereby combining electrical as well as nonelectrical elements. The challenge for SiP will be to develop a degree of standardization and reuse, so that an economically viable industry can be built. Of course, the economics will be quite different for high volume consumer applications and low volume professional applications, for example health care.

5.2. Open Innovation Approaches

In the world of "more of Moore", the exploding costs of realizing successive technology generations have become almost prohibitive. Dealing with the *soaring costs of innovation* is today the number one challenge for silicon microelectronics. The solution is offered by open innovation approaches [19]. As an early adopter of this approach, Philips Research has pioneered intensive partnerships in R&D. Thus, we are carrying out our advanced silicon process technology research together with the

Interuniversity Microelectronics Center (IMEC) in Leuven, Belgium. Philips semiconductors and its partners STMicroelectronics and Freescale carry out the subsequent development and pilot manufacturing in the Crolles 2 Alliance near Grenoble, France. By using advanced wafer foundries, such as TSMC of Taiwan, cost sharing and risk reduction is achieved in mass manufacturing as well.

The key challenge to be addressed in the new field of ''More than Moore'' is managing the *soaring complexity*, caused by the bewildering variety of applications and emerging technical solutions. Roadmaps are virtually nonexistent, and the pursuits of material suppliers, toolmakers, device makers, designers, and OEM customers are not yet well aligned. Added to this are the pitfalls of the innovator's dilemma: how to balance investments in a new application based on unproven technology against investing in the next generation of a predictable technology development?

In our view, the world of ''More than Moore'' can only become as successful as silicon microelectronics if we succeed also in this field in defining shared roadmaps, with attention for modular and platform solutions, common modeling, simulation and design tools, and shared testing strategies. This will call for intensive technology partnerships. Philips, being a strong believer in open innovation, is pursuing this actively by articulating and sharing our long-range application and technology visions with technology partners and key customers. We are also actively engaged in a variety of joint R&D programs with our technology partners.

As a concrete step toward open innovation in the new field of SiP, Philips Research has recently opened its new MiPlaza cleanroom at the High Tech Campus in Eindhoven, the Netherlands. Facility sharing is actively promoted here. Several third party start-up companies are already using MiPlaza today. Also, we are working together with toolmakers and with leading universities. It is our intention that MiPlaza will become a highly interdisciplinary meeting place for academic and industrial researchers in the field of microsystems and nanotechnology working together in a spirit of open innovation [20].

6. CONCLUSION

Summarizing, we hope to have provided an idea about the type of physical layer challenges posed by vision of AmI. The major topics to work on are: ubiquitous, flexible and reflective displays, extreme miniaturization, and integration of new functions, such as sensors and actuators with silicon logic, textile electronics, and solid-state lighting technologies. A key bottleneck in the development of autonomous devices and

ubiquitous computing is the energy supply of the distributed device nodes. Open innovation approaches will be required to manage the soaring complexity of this field.

ACKNOWLEDGEMENT

The author wishes to acknowledge the contributions of his colleagues at Philips, in particular Emile Aarts, Gerjan van de Walle, and Fred van Roosmalen.

REFERENCES

[1] Emile A. and Stefano M., (eds), *The New Everyday, Views on Ambient Intelligence*, Koninklijke Philips Electronics N.V., 010 Publishers, Rotterdam (ISBN 90–6450–502–0).
[2] Without claiming completeness: Sony is talking about augmented reality, IBM and MIT about pervasive computing, AT&T about sentient computing, Xerox about ubiquitous computing, Microsoft about intelligent environments, Georgia Institute of Technology about aware computing, and Stanford university about interactive workspaces.
[3] Mark W., 1991, *The Computer for the Twenty-First Century*, Scientific American, pp. 94–110 September 1991. See also: Mark W., 1993, Some computer science problems in ubiquitous computing, *Communications of the ACM*, July 1993. (reprinted as *Ubiquitous Computing*, Nikkei Electronics, pp. 137–143.)
[4] Hsu, V, J. M. Kahn and K. S. J. Pister, 1998, *Wireless Communications for Smart Dust*, Electronics Research Laboratory Technical Memorandum Number M98/2, February 1998.
[5] Byron R. and Clifford N., 1998, *The Media Equation*, CSLI Lecture Notes, University of Chicago Press, Chicago.
[6] Penterman, R., Klink, S., de Koning, H., Nisato, G., Broer, D., 2002, Nature, **417**, 55.
[7] Hayes, R. A. and Feenstra, B., 2003, Nature (London), **425**, 383.
[8] Estrin, D., Culler, D., Pister, K. and Sukhatme, G., 2002, Connecting the physical world with pervasive networks, *PervasiveComputing*, January 2002.
[9] Kuiper, S. and Hendriks, B. H. W., 2004, Variable focus liquid lens for miniature cameras, *App. Phys.Lett.*, **85**, 1128.
[10] Pistor, K., Schmidt, F., patent application WO 01/913115 A2. See also www.EnOcean.com.
[11] West, W. C., Whitacre, J. F., White, V. and Ratnakumar, B. V., 2002, Fabrication and testing of all solid-state microscale lithium batteries for microspacecraft applications, *J. Micromech. Microeng.*, **12**, 58.

[12] Roundy, S., Wright, P. K. and Rabaey J., 2003S A study of low level vibrations as a power source for wireless sensor nodes, *Computer Communications*, **26**(11), 1131–1144 July 2003.
[13] Pelgrom, M., Roovers, R. and Bergveld H. J., *Smart Dust*, *in* Ref. 1 of Chapter 2.1. See also Bergveld, H. J., Kruijt, W. S. and Notten, P. H. L., 2002, *Battery Management Systems – Design by Modelling*, Kluwer Academic Press.
[14] Farrington, J., Moore, A. J., Tilbury, N., Church, J. and Biemond, P. D., *Wearable Sensor Badge and Sensor Jacket for Context Awareness*, J Farrington, AJ Moore, N Tilbury, J Church, PD . . . The Third International Symposium on Wearable Computers, 1999.
[15] Gough, P., *Wearable Technology*, *in* Ref. 1 of Chapter 4.4.
[16] Starner, T., 1996, Human powered wearable computing, *IBM Syst. J.*, **35**, 618.
[17] Jung, S., Lauterbach, C., Strasser, M., Weber, W., *Enabling Technologies for Disappearing Electronics in Smart Textiles*, Solid-State Circuits conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International, **1**, 386–387.
[18] Zakauskas, A., Gaska, R., Shur, M., Shur, M. S., Zukauskas, A., 2002, *Introduction to Solid State Lighting*, John Wiley, New York.
[19] Chesbrough, H., 2003, *Open Innovation*, *The New Imperative for Creating and Profiting from Technology*, Harvard Business Press, Boston.
[20] Information about the MiPlaza cleanroom can be found at http://www.research.philips.com/institutes/index.html.

Section 2

Wireless Communication

Chapter 2.1

CIRCUITS AND TECHNOLOGIES FOR WIRELESS SENSOR NETWORKS

Brian Otis
Electrical Engineering Department
University of Washington, Seattle, Washington
botis@ee.washington.edu

Mike Sheets, Yuen-Hui Chee, Huifang Qin,
Nathan Pletcher, and Jan Rabaey
Berkeley Wireless Research Center, University of California, Berkeley
jan@eecs.berkeley.edu

Abstract Successful deployment of wireless sensor and actuator networks in sufficient numbers to provide true ambient intelligence requires the confluence of several disciplines including networking, low power RF and digital IC design, MEMS techniques, energy scavenging, and packaging. Progress in each of these areas has been documented and proof-of-concept prototypes have been tested. Research in RF transceiver design utilizing bulk acoustic wave resonators has yielded fully integrated, ultralow power transceivers. Novel digital circuit design techniques, including aggressive power management, robust subthreshold logic operation, and ultralow voltage SRAM with data retention enable efficient computation. These technological advances should be accompanied by novel opportunistic networking and media access techniques to provide robustness and decrease the duty cycle of the node. Future challenges include the integration of a sub-50 μW carrier sense detector for asynchronous and non-beaconed receiver wake-up, efficient hybrid energy scavenging power generation, and cheap, robust three-dimensional packaging techniques.

Keywords leakage reduction; MEMS; RF CMOS; SRAM

1. INTRODUCTION

The successful large-scale deployment of wireless sensor networks (WSN) will yield large benefits for environmental, medical, and security

S. Mukherjee et al. (eds.), AmIware, 31–50.

monitoring applications [1]. A truly ubiquitous deployment is economically feasible only when the cost and size of the individual elements become negligible. Achieving such a diminutive stature requires a minimal number of components, favoring those that are inexpensive, have a high level of integration, and have simple packaging and assembly requirements. In addition, the cost of deployment and maintenance of the network should be negligible. The physical implementation of an individual network node is constrained by three important metrics: power, cost, and size. These requirements are outlined below.

- **Power:** To reduce installation cost and to allow flexible deployment, most nodes must be untethered and have their own energy source. Frequent replacement of the energy source is infeasible because it would result in a high maintenance cost. Thus, the nodes must scavenge their energy from the environment. This requires the average power dissipation of the node to be extremely low ($< 100\,\mu W$).
- **Low Cost Implementation:** For commercial viability, the cost/area of the wireless sensor network mesh must be very small. For network reliability, the node density (nodes/area) must be high. Thus, the cost of each node must be extremely small (<$1US).
- **Small Form-Factor:** Embedding the components into the infrastructure of daily life (walls, furniture, lighting, etc.) requires a small form factor for each node. For many scenarios, sizes smaller than $1\,cm^3$ are necessary. A very high level of integration is mandatory if such small dimensions are to be achieved.

This chapter discusses circuits and technologies aimed at realizing the ubiquitous deployment of WSNs. We begin with an exploration of advancements in digital computation that reduce active and standby power requirements. Recent and future research in RF communications links is then described.

2. EFFICIENT DIGITAL COMPUTATION

As process dimensions and supply voltages decrease, the CMOS leakage power becomes a significant component of the total power budget. This is particularly true for sensor nodes because the low duty cycle means that a large portion of the total time is spent idling. This section describes techniques to reduce the leakage power consumption of the digital computational blocks and embedded memory.

2.1 Advanced Power Management

Power management is critical in ambient, wireless devices to maximize battery life or enable operation on scavenged energy. Numerous leakage reduction techniques for logic have been explored, including using non-minimum device length, stack effect, reverse-body-bias, sleep transistors, and MTCMOS [2,3]. The most dramatic savings are attained by the latter three techniques, which are run-time strategies that have an active mode and a low-leakage sleep mode. A power manager is thus required to control the mode of the various subsystems in the node.

To make an informed decision about when to switch modes, the power manager must collect data from the various subsystems. Each subsystem that can be controlled independently forms a power domain (PD). A common interface for each PD facilitates composition of the PDs into a complete low-power system. Within each PD, related input/output (I/O) signals are bundled into ports that can be either open or closed. When the port is closed, a signal wall forces all I/Os to a known state. This prevents spurious external signaling from corrupting a PD that is in sleep mode [4]. A PD can influence the port states by issuing power control messages through the common interface. Compatible ports on different PDs are directly connected, but the ability to actually communicate is controlled by the power manager.

An example of this power management architecture is the PicoRadio Charm digital protocol processor. The chip integrates a synthesized 8051-compatible embedded microcontroller with 64 kB of RAM, two 1 kB packet queues, a custom data-link layer, a neighborhood management subsystem, digital baseband logic, a location computation subsystem, and several standard interfaces enabling connection of arbitrary sensors and actuators. These functions are divided into eight PDs, all controlled by a centralized power manager (system supervisor). This supervisor is programmed with the connectivity between all the ports in the system and keeps track of which ports are currently open. Its basic power policy permits a PD to sleep when a port is closed on both sides and the PD is idle. A PD is reactivated when another PD opens a connected port. This method is particularly suitable for event-driven systems like sensor nodes. False wake-ups never occur because the power manager simply reacts to pending events. However, it is possible for the supervisor to put a block to sleep that will be awoken before the minimum idle time is reached (on the order of hundreds of microseconds). Since the expected time between external events is on the order of milliseconds, minimal false alarms are expected.

The leakage suppression method used in the Charm chip is a switched virtual supply rail that retains the state of the PD during sleep mode. It is implemented as a power control switch (PCS) cell that is added to the

standard cell library. When a PD is active, a power switch in the PCS connects the active voltage ($V_{DDHI} = 1.0\,V$) to the virtual power rail. Since the control signals have sufficient swing to prevent a threshold-drop, a smaller NMOS switch connects a lower retention voltage ($V_{DDLO} = 0.3\,V$) when the PD is sleeping. The cells are placed using a script that easily integrates into an industry standard place and route design flow. No changes to the standard cell library are required. The retention voltage is generated by an on-chip, switched-capacitor voltage converter.

The Charm chip, shown in Figure 2.1–1, is implemented in a 0.13 μm triple-well, bulk digital CMOS process with six metal layers.

The chip is $(2.7 \times 2.7)\,mm^2$ and integrates 3.2 M transistors. The power manager core utilizes 1574 gates. The next section describes how the embedded SRAM memory leakage for this chip can be reduced.

2.2. Ultralow Data Retention Voltage SRAM

A large percentage of the leakage power in the Charm chip is dissipated in the embedded SRAM. Thus, this section contains a detailed analysis of how to reduce the SRAM leakage while maintaining the state. The most effective circuit level techniques are to lower the supply voltage and increase the transistor threshold voltage. Dynamic techniques include putting inactive sections into standby mode, such as lowering the supply to a minimal retention voltage. The idle time between active bursts is usually long ($> 1\,\mu s$) so switching the entire SRAM module to standby mode is feasible. This scheme can effectively suppress the leakage of entire SRAM block as a whole. To preserve the memory contents while maximizing power savings during the low voltage standby mode, the minimum reliable data retention voltage (DRV) should be used. We now present an analysis of the minimal DRV for a 6T SRAM cell.

2.2.1. Ultralow voltage data retention analysis

The circuit structure of a common 6T SRAM cell is shown in Figure 2.1–2. During standby mode, the bitline voltages are set to V_{DD}. When

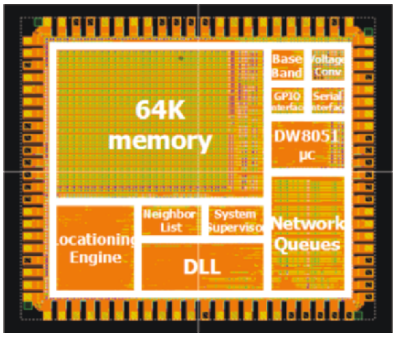


Figure 2.1–1. Charm protocol processor.

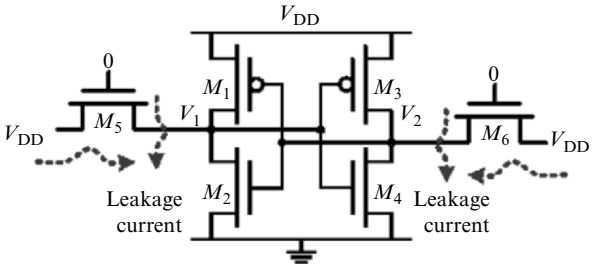


Figure 2.1–2. Standard 6T SRAM cell structure.

V_{DD} is reduced to the DRV, all six transistors in the SRAM cell operate in the subthreshold region. Thus, the data retention strongly depends on the subthreshold current conduction behavior of the transistors.

To reliably preserve the state of the cell, the feedback loop must be regenerative and have high stability. Regeneration requires the cross-coupled inverters to have a loop gain greater than one. Stability of an SRAM cell can be measured with the static noise margin (SNM) [5]. As shown in Figure 2.1–3a, the SNM is graphically represented as the area between the voltage transfer characteristic (VTC) curves of the cross-coupled inverters. As V_{DD} is lowered, the loop gain is reduced and the SNM falls to zero. Note that the VTC curves in this figure represents worst case variations in both inverters.

Thus, if V_{DD} is reduced below the DRV, the inverters can switch to the other biased state determined by the deteriorated inverter VTC curves,

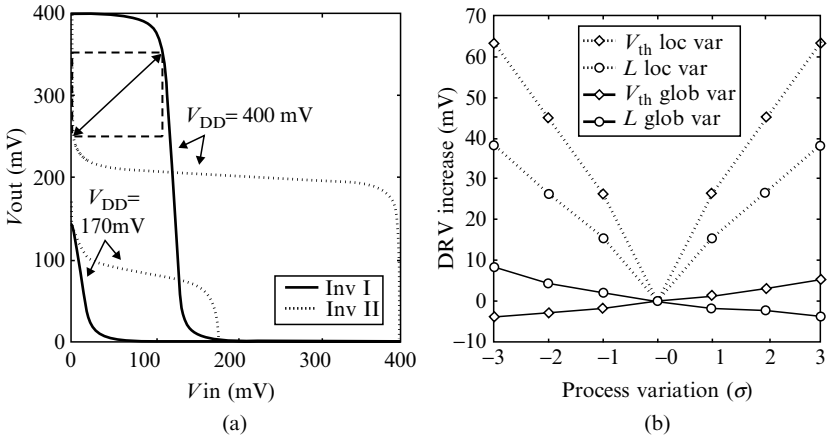


Figure 2.1–3. (a) Inverter VTC under low V_{DD}. (b) DRV sensitivity to process variations.

corrupting the stored data. The DRV of an SRAM cell can be determined by solving the subthreshold VTC equations of the two internal data-holding inverters. Transistor currents involved in DRV analysis are dominated by the drain–source leakage in current technologies.

The DRV is dependent upon process and temperature variations. Mismatch between the two internal inverters has a strong impact on its DRV. Simulation data in Figure 2.1–3b plots the change in DRV value versus the magnitude of variations in an SRAM cell. Local mismatch among transistors results in substantial DRV increase. Based on a 0.13 μm technology model, with 3σ local V_{th} mismatch, the DRV is 70 mV higher than the ideal case with perfect matching. Global parameter shifting has a much weaker impact on DRV.

2.2.2. SRAM test chip implementation and measurements

A 4 KB SRAM test chip with dual rail standby control was implemented in a 0.13 μm technology (Figure 2.1–4b) to verify the DRV model and explore the potential of leakage suppression using this method. As shown in Figure 2.1–4a, the SRAM supply rails are connected to the standard and standby V_{DD} through power switches. The test chip contains a 4 KB SRAM module and a switched-capacitor converter that generates the calculated DRV with 85% conversion efficiency [6].

The DRV is measured by monitoring the data retention capability of an SRAM cell with different values of standby V_{DD}. The actual DRV for each of the 32K SRAM cells is measured by monitoring the data retention with different values of standby V_{DD}. Figure 2.1–5a shows that the distribution of the results ranges from 60 to 390 mV with a mean of 122 mV.

Such a wide range of DRV uncertainty reflects the existence of considerable process variations or noise. Due to global variations, the lower end

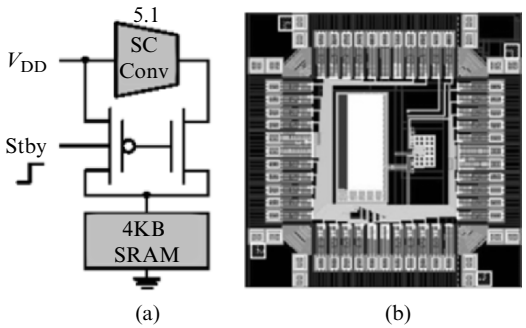


Figure 2.1–4. (a) Standby leakage suppression scheme. (b) 0.13 μm test chip.

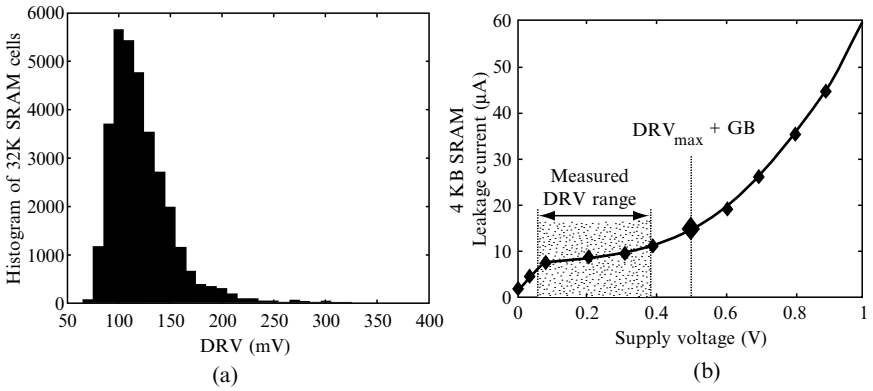


Figure 2.1–5. (a) Measured DRV distribution of a 4 kB SRAM chip. (b) Measured SRAM leakage current.

of measured DRV is slightly lower than the 78 mV ideal DRV, assuming perfect process matching. The long distribution tail reduces the leakage achievable reduction using this method.

Leakage measurement results of the 4 KB SRAM are shown in Figure 2.1–5b. The shaded area in this figure indicates the range of measured DRV (60–390 mV). Although the memory states can be preserved at sub-400 mV V_{DD}, the robustness of state preservation can be increased by adding an extra guard band of 100 mV to the standby V_{DD}. With the resulting 490 mV standby V_{DD}, SRAM leakage current can still be reduced by over 70%. Subsequently the leakage power, as the product of V_{DD} and leakage current, is reduced by about 85% compared to 1 V operation.

2.3. Ultralow Voltage Digital Design

Reducing the power dissipation of future digital circuit and system designs requires the adoption of drastic new design approaches, such as operating under very low supply voltages (300 mV and below). Design challenges for such ultralow voltage operation include performance degradation and exaggerated variation effects, which can cause dramatic losses in yield. However, innovative techniques at all levels can be applied to enhance the system operation speed and robustness. For example, effective error-correction schemes can help reduce the standby V_{DD} of an SRAM below the DRV, achieving a dramatic reduction in memory leakage power. Other techniques include adaptive design methodologies and yield-aware logic optimizations.

3. LOW POWER RF COMMUNICATION LINKS

Experience and experimentation have shown that the radio frequency (RF) communication link constitutes a majority of the power consumption in wireless sensor nodes. Commercial low power transceivers typically consume over 10 mW in both transmit and receive modes. In addition, numerous external components are typically required for these implementations. The unique constraints placed on the transceiver by the wireless sensor network application can be broken into three categories:

(1) **Power Dissipation:** Assuming a 100 μW node power budget, a 1% transceiver duty cycle, and a 10% power allocation to the RF link, the transceiver must consume approximately 1 mW when active.
(2) **Integration:** There exists a well-documented power/integration tradeoff in the design of RF transceivers. External filters, surface mount inductors, low carrier frequencies, and special IC processes enable low power designs. However, the 1 cm^3 volume constraint eliminates the luxury of multiple external components, or low carrier frequencies to relax the transceiver specifications.
(3) **Agility:** In an ad-hoc network scenario, the radio is idle or off mode most of the time, data communications are rare, and packets are short. Thus, it is essential that the radio start-up and acquisition be very short to represent a small overhead in the overall packet duration.

This section describes design strategies and examples for how these goals can be achieved.

3.1. MEMS/CMOS Codesign

The relatively new field of radio frequency microelectro-mechanical systems (RF-MEMS) provides unique opportunities to RF transceiver designers. This section provides background on RF MEMS and gives insight into the opportunities presented by these new technologies.

The field of RF MEMS includes the design and utilization of RF filters, resonators, switches, and other passive mechanical structures constructed using integrated circuit fabrication techniques. These devices are already available as discrete board-mounted components, primarily used to enhance the miniaturization of mobile phones [7]. RF-MEMS components have the potential to be batch fabricated using existing integrated circuit fabrication techniques. New capacitively driven and sensed structures offer the potential of integration on the same substrate as the CMOS circuitry. In addition, since the resonant frequency is set lithographically,

and not by a deposition layer thickness, it is possible to fabricate devices with many unique resonant frequencies on the same wafer.

A proof-of-concept, 2 GHz bulk acoustic wave (BAW) resonator was codesigned with a 0.18 μm CMOS oscillator [8]. The completed system is shown in Figure 2.1–6. The BAW resonator die is wirebonded directly to the CMOS chip. The elimination of surface mount quartz crystals, the low power consumption, and the fast start-up time makes this a good candidate to replace the quartz-based PLL in low power, highly integrated, agile transceivers.

3.2. Ultralow Power Receiver Architectures

The first BAW-based transceiver was shown in [9]. A two-channel tuned radio frequency (TRF) architecture was chosen to demonstrate the effectiveness of RF-MEMS resonators in low power transceivers. By shifting the burden of channel selection from active components to passive resonators, the transceiver size and power consumption was decreased. See the transceiver block diagram in Figure 2.1–7.

The antenna feeds a 50 Ω impedance presented by the LNA. The LNA drives a tuned LC load absorbing the capacitive input of two channel select amplifiers (CSAs). Each CSA is tuned by a BAW resonator, which performs receiver channel selection. Channel selection with passive high Q filters eliminates the need for a PLL or mixers, reducing the start-up time and power consumption of the receiver. Although two channels were used in this implementation, the architecture is scalable to larger numbers of channels (limited by the 200 MHz LNA bandwidth). In the future, advanced RF-MEMS technology will allow lithographically defined resonant frequencies and integration of MEMS onto CMOS wafers. An RF detector performs self-mixing of the signal to perform downconversion

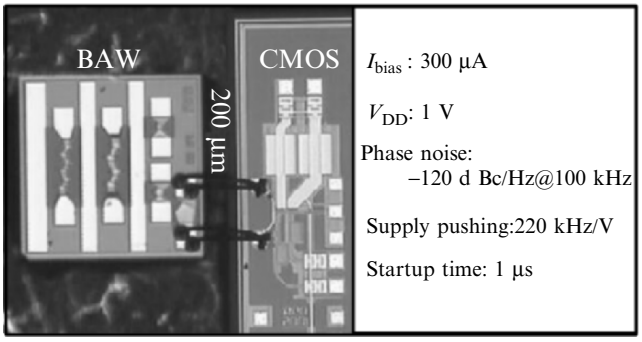


Figure 2.1–6. 300 μW BAW-based 2 GHz oscillator.

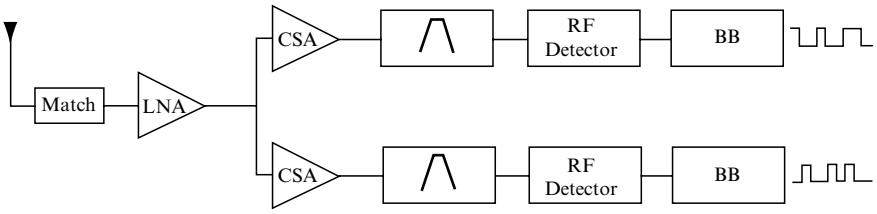


Figure 2.1–7. Two-channel transceiver block diagram.

without a local oscillator (LO). See the receiver front-end schematic in Figure 2.1–8.

The channel select amplifier provides gain to reduce the noise contribution of the detector as well as to interface the electrical signal with the acoustic resonator to perform high-Q filtering of the signal. The amplifier must exhibit high RF gain with low power consumption while limiting the extent to which the resonator is detuned. The BAW, used in the parallel resonance mode, performs high Q filtering by presenting an high impedance for a very narrow bandwidth (~ 3 MHz) about its parallel resonant frequency. However, for off-resonance frequencies, the resonator presents

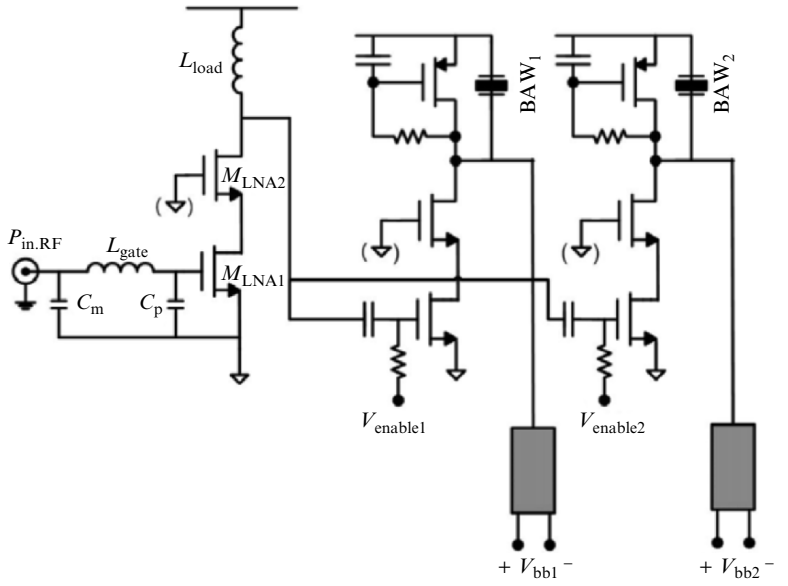


Figure 2.1–8. Two-channel transceiver front-end schematic.

a $1/j\omega C$ impedance (mechanically, off-resonance, the resonator is modeled two parallel plates filled with an AlN, $k = 9$ insulator). To provide bias current to the active devices and reduce LF gain, a low load impedance at DC is required.

However, at the signal frequency, the load impedance must be high to avoid detuning the BAW resonator. Thus, an inductive structure is a natural candidate. To avoid detuning of the BAW at 1.9 GHz, the bias device must exhibit an impedance greater than 1500 Ω. This corresponds to an inductance of 125 nH, which is marginally feasible to fabricate on-chip. Even if possible, it would resonate with the parallel plates of the BAW (2 pF) at 318 MHz, adding an out-of-band response to the receiver input. It would also consume $> 10{,}000\,\mu m^2$ of silicon area. The use of an active inductor, however, allows the realization of very high inductance values and accurate control over the inductor Q.

The two-channel embodiment displays flexibility in terms of modulation schemes; the receiver can detect two unique on–off keying (OOK) data streams at two carrier frequencies or it can detect FSK. For dense wireless sensor networks, it is anticipated that two separate OOK channels will be used, with one reserved for beaconing. Changing between these two modulation schemes can be accomplished with no receiver modifications, and can be performed dynamically in either the analog or digital baseband detection circuitry. The transceiver was implemented in a 0.13 μm standard CMOS process (Figure 2.1–9).

The $(4 \times 4)\,mm^2$ CMOS die is wirebonded to the four-BAW resonator die: two for transmit and two to receive. Both receive channels exhibit a

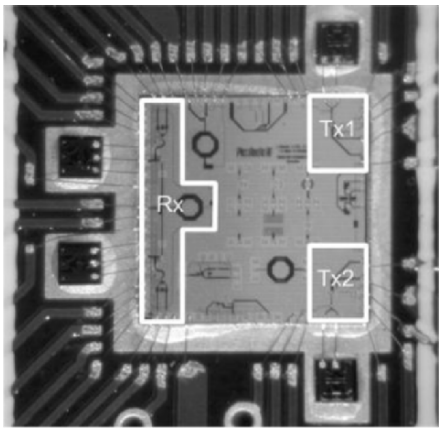


Figure 2.1–9. Low power two-channel transceiver implementation.

3 MHz bandwidth and close gain matching. The loaded Q of the BAW resonators in the receive chain is approximately 600. Receiver sensitivity for a 12 dB SNR was measured at −78 dBm up to 40 kbps. As discussed above, the agility of the receiver is crucial in a wireless sensor network application. The enable-to-data latency of the receiver was measured to be 10 μs, much less than one symbol period. The total receiver current consumption is 3 mA with both channels enabled.

To further reduce the power consumption of the receiver, a super-regenerative architecture was explored. The design and implementation is beyond the scope of this chapter. The super-regenerative receiver consumes 400 μW and is implemented in 1 mm^2 of standard 0.13 μm CMOS.

3.3. High-Efficiency Transmitter Architectures

Wireless sensor networks present new challenges to the RF transmitter. This application requires (1) power shutdown of the transmitter when idle, (2) a short turn-on time of the transmitter to minimize overhead power, (3) a low transmit power of about 0 dBm (1 mW), and (4) a high global transmitter efficiency.

We first examine the direct conversion transmitter, which is the workhorse of today's wireless systems, and later propose the direct modulation transmitter, which is more suitable for low power ad-hoc wireless sensor networks. Finally, we present the implementation of the direct modulation transmitter.

3.3.1. Direct conversion transmitter

The direct conversion transmitter is the most common architecture in today's wireless transceivers. See Figure 2.1–10.

In the direct conversion transmitter, the data stream is processed by the digital modulator to a suitable baseband signal before being converted

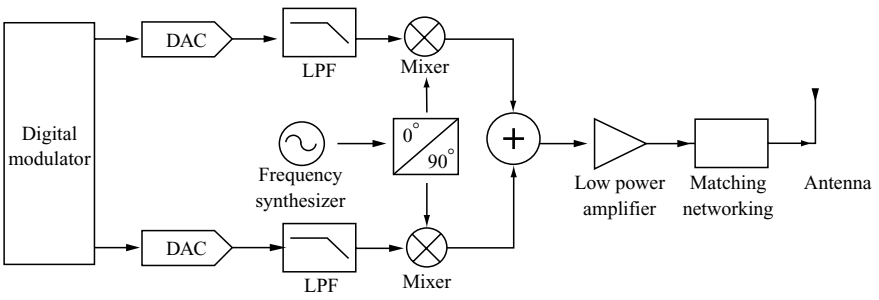


Figure 2.1–10. Traditional direct conversion transmitter.

into an analog waveform at the DAC. Next, the baseband signals are up converted and boosted to the required power level by the power amplifier. The matching network transforms the 50 Ω antenna to an optimal impedance for efficient operation of the power amplifier.

An implementation of the direct conversion transmitter for wireless sensor networks is reported in [10]. The transmitter delivers 0 dBm and consumes a total of 30 mW, resulting in only 3.3% efficiency. The breakdown of the transmitter's power consumption is as follows: frequency synthesizer, 40%; PA, 48%; modulator, 1.8%; mixer, 10%. A careful analysis reveals that the direct conversion transmitter is not suitable for wireless sensor network applications for the following reasons:

(1) **High overhead:** The power breakdown above shows that the frequency synthesizer, modulator and mixer consume more than 50% of the total power. This is a high proportion compared to cellular applications where the PA dominates the transmitter's power consumption. The radiated power in cellular systems is much higher (~30 dBm) than that in wireless sensor networks (0 dBm). Thus, the overhead of the synthesizer and upconverter becomes comparable or higher than the radiated power. This results in low transmitter efficiency.
(2) **High Complexity:** In typical cellular system or wireless LAN transceivers, complex modulation schemes are employed to maximize the spectral efficiency or data rate. However, in wireless sensor networks, the data rate is low and spectral efficiency is not paramount. In this case, the direct conversion transmitter is overkill and we could employ a much simpler transmitter which uses fewer circuit blocks to reduce the overall power consumption.

3.3.2. Directly modulated transmitter

Leveraging the unique characteristics of the wireless sensor network environment, the direct modulation transmitter shown in Figure 2.1–11 below is very attractive.

The transmitter is directly modulated by the baseband data with OOK modulation. The oscillator produces the RF carrier and the low power amplifier efficiently boosts the RF signal power. Direct modulation yields an ultralow power transmitter because of its simplicity. The mixers, digital modulator and DAC are eliminated. OOK obviates the need for quadrature channels, reducing the number of active circuit blocks. Furthermore, the transmitter is only active when transmitting a "one" symbol, resulting in a 50% energy savings if the long-term probabilities of sending a one or zero are equal.

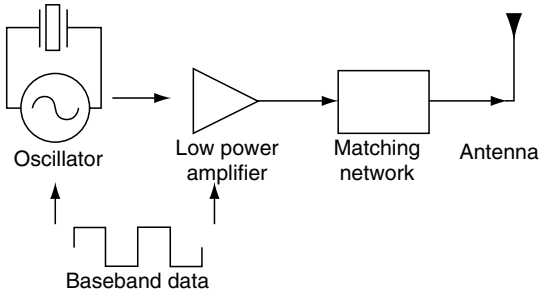


Figure 2.1–11. Directly-modulated transmitter.

3.3.3. Directly modulated transmitter implementation

The schematic of the direct modulation transmitter is shown in Figure 2.1–12.

The transmitter consists of a BAW-based Pierce oscillator and a class C low power amplifier. The baseband signal directly modulates the bias current of the oscillator and the low power amplifier, thus enabling OOK. Carrier generation is accomplished with an oscillator codesigned with a BAW resonator, eliminating the PLL and quartz crystal.

The nonlinear class C amplifier boosts the RF signal power. In this implementation, a capacitive transformer is used to transform the 50Ω antenna (R_L) to an optimal impedance for high efficiency operation. Capacitive transformation is preferred over LC matching networks because on-chip capacitors are less lossy and occupy smaller silicon area than integrated inductors. The cascode transistor ensures that the gate–drain voltage does not exceed the low gate breakdown voltage and increases the output/input isolation of the amplifier.

This transmitter was implemented in a 0.13 μm CMOS process [9]. The die photo is shown in Figure 2.1–9. The CMOS circuitry occupies about 1 mm^2 and the BAW is packaged with the CMOS circuitry using low cost chip-on-board technology. The measured peak efficiency of the transmitter is 16.5% when delivering 1.5 mW. The transmitter achieves a start up time of 2 μs, enabling a bit rate of 50 kbps if the start-up time is set to 10% of the symbol duration.

3.4. Opportunistic Routing and Efficient MAC Design

To ensure network-level efficiency, the transceiver cannot be studied in isolation because the RF transceiver cannot guarantee perfectly reliable com-

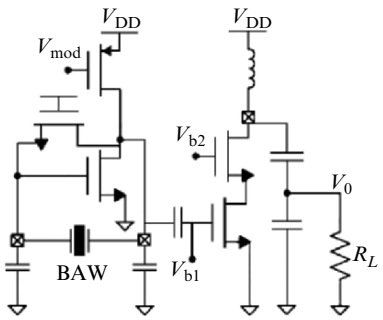


Figure 2.1–12. Schematic of directly-modulated transmitter.

munication. Low-power, narrowband transceivers in particular can succumb to deep fading, interference, and multipath effects. Thus, the protocol stack (routing and media-access control) is designed such that no end-to-end connection ever relies on the presence of a specific node. This approach is called "opportunistic routing" [11]. During packet forwarding, the recipient node is not based on routing tables, or predefined paths, but on the presence of a reachable node that is awake and is located in the direction of the final destination. This technique ensures reliable communication even in the case of deep fading or node energy shortages. To achieve network robustness even with unreliable, inexpensive hardware, system/hardware codesign techniques such as this are crucial.

In order for two nodes to communicate there must be a rendezvous scheme that ensures both will be simultaneously active to initiate the communication. Such schemes can be broadly categorized as purely synchronous, pseudo-asynchronous, and purely asynchronous.

Purely synchronous schemes require synchronized nodes across the entire network, which is difficult in ad-hoc networks. Pseudo-asynchronous methods employ a beaconing protocol where nodes wake up periodically to monitor the channel for data transmission. In a fully asynchronous rendezvous scheme, each node constantly monitors the channel with a carrier sense receiver, or wake-up radio. The wake-up radio monitors the channel for beacon signals from other nodes and turn on the node's main data radio for communication. Since the wake-up radio is *always* monitoring the channel, its power consumption must be minimal. Power modeling of several rendezvous schemes shows that the wake-up radio must consume less than 50 μW to save power over pseudo-asynchronous schemes [12].

3.5. Next Generation: Ultralow Voltage RF Design

As discussed in the previous section, a purely asynchronous communication scheme between nodes offers the lowest global power consumption. This necessitates an ultralow power carrier sense receiver to monitor for transmit beacons from neighboring nodes. This component must consume less than 50 μW to outperform pseudo-asynchronous schemes. This is an extremely aggressive target for any type of RF receiver. The implementation of the wake-up radio requires rethinking of traditional RF design techniques with power consumption as the limiting design constraint.

In addition to the new opportunities provided by RF-MEMS technology, the impressive advances in CMOS technology also open up new frontiers in low voltage and low current circuit design. Two design techniques, namely subthreshold device operation and exploitation of low supply voltages, offer a potential path to ultralow power carrier sense receivers.

3.5.1. Subthreshold design

In the past ten years, CMOS technology has become well-established in RF circuit design. Traditionally, high frequency transistors are operated in strong inversion to take advantage of the high device f_T in this regime. It has long been understood that subthreshold (weak inversion) device operation provides more transconductance g_m for a given bias current, enabling extremely low power CMOS circuits [13]. This increased transconductance efficiency comes at the expense of lower device f_T, so subthreshold design has historically been confined to low frequency applications. However, with recent CMOS technologies boasting maximum f_T above 100 GHz, it is no longer necessary to bias devices for the highest possible f_T. By carefully choosing the region of operation for critical transistors, the designer can trade off device bandwidth for transconductance efficiency to achieve a lower power design.

As presented earlier in the chapter, the device drain current in weak inversion is given by Equation (2.1–1). In this region, I_D becomes exponentially dependent on the gate voltage of the device, yielding a characteristic similar to a bipolar transistor. In order to gain an intuitive understanding of device performance across the various regions of operation, it is useful to define an *inversion coefficient (IC)* for a given device [14]. See Equation (a) in Figure 2.1–13. I_D is the device drain current, n is the subthreshold slope factor, $k' = \mu_0 C_{0x}$, and $U_t = kT/q$. The specific current I_0 is technology dependent and can be extracted from simulations or measurements. For a given device and bias condition, the inversion coefficient can be readily

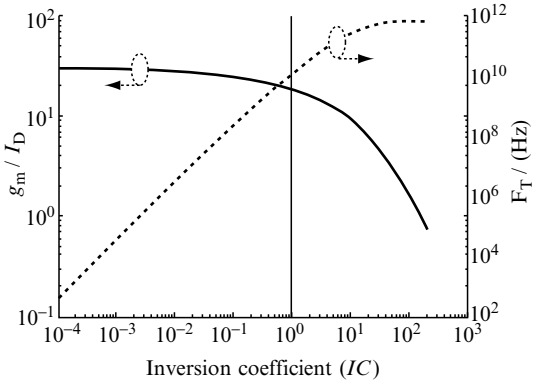
calculated (Equation (b) in Figure 2.1–13). If $IC \ll 1$, then the device is operating in weak inversion. $IC \gg 1$ indicates strong inversion, while $IC = 1$ is designated moderate inversion. For many applications, moderate inversion provides an attractive compromise between bandwidth and transconductance efficiency. A plot of simulated g_m/I_D versus IC is shown in Figure 2.1–13.

The vertical line in Figure 2.1–13 at $IC = 1$ indicates the center of moderate inversion. The roll-off of transconductance efficiency in strong inversion is evident in the region where $IC > 1$. Also shown is the trend of device f_T across all regions of operation. These concepts provide the designer with accurate, simulation-based charts that ensure efficient transistor biasing.

3.5.2. Design for low supply voltage

One of the difficulties with continuing technology scaling is the reduction in supply voltage for modern CMOS processes, causing reduced voltage headroom and dynamic range for analog applications. In some cases, however, it is possible to embrace this trend and reduce the supply voltage as low as possible to achieve minimum power consumption.

The standard differential LC oscillator shown in Figure 2.1–14a is able to operate with a very low supply voltage. The inductive load does not consume any voltage headroom, enabling the output to swing above V_{DD}. Furthermore, if the cross-coupled devices operate in weak inversion, V_{GS} and V_{Dsat} are minimized. Theoretically, the oscillator can operate on a supply voltage as low as $V_{Dsat1} + V_{Dsat3}$. The critical parameter for oscillator start-up is the g_m of the cross-coupled devices, which establishes a



a) $I_D = \frac{W}{L} I_t \exp\left(\frac{V_{GS} - V_t}{{}_nU_t}\right) * \left(1 - \exp\left(\frac{-V_{DS}}{U_t}\right)\right)$

b) $IC = \frac{I_D}{2nk' U_t^2 \frac{W}{L}} = \frac{I_D}{I_0 \frac{W}{L}}$

Figure 2.1–13. Simulated plot of g_m/I_D and f_T versus IC.

lower limit on the current consumption of the oscillator. Thus, the current consumption is minimized by operating in weak inversion.

The main drawback to such a low V_{DD} is reduced output voltage swing. Most VCOs for transceiver applications are designed for maximum swing in order to minimize phase noise [15]. For an ultralow power oscillator, phase noise performance is sacrificed for power savings. This results in a start-up limited design where swing is not limited by supply voltage.

3.5.3. Low power oscillator design

Employing the design principles outlined above, the oscillator shown in Figure 2.1–14a was designed and fabricated in a standard 0.13 μm CMOS process. The fully integrated design utilizes on-chip 10 nH inductors and includes 50 Ω output buffers to drive measurement equipment.

Nominally, the oscillator is biased at 400 μA from a 0.5 V supply, where the oscillation frequency is 1.35 GHz and the in situ differential output swing is 120 mV zero-peak. Under these conditions, the cross-coupled transistors operate in moderate to weak inversion with $IC \approx 0.4$. To verify performance for low supply voltages, the oscillator was tested across various bias currents and voltages. The resulting variations in oscillation frequency and output swing are shown in Figure 2.1–14b. The nominal DC power consumption is 200 μW, however the oscillator continues to operate down to an extremely low V_{DD} of 300 mV. Although the frequency changes due to nonlinear device capacitance,

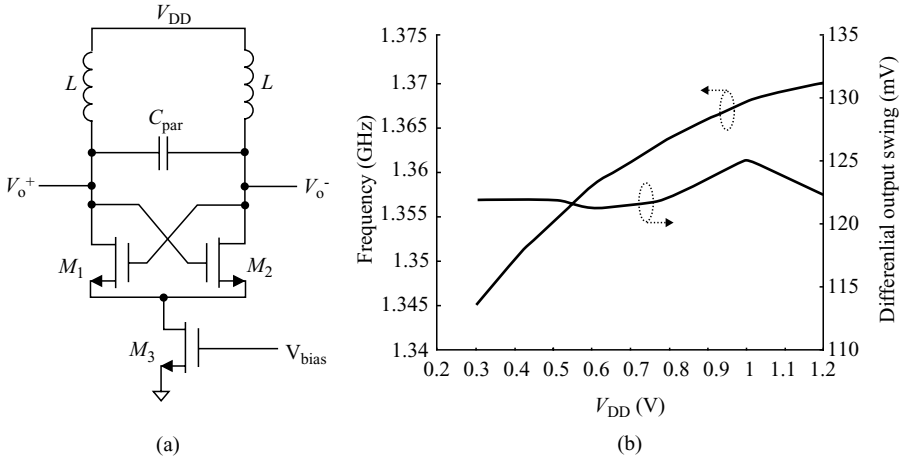


Figure 2.1–14. (a) Low voltage oscillator schematic (b) Measured oscillator performance.

output swing is nearly constant across the range, indicating that the low supply voltage is not limiting the swing.

This exploration into ultralow voltage RF design shows that RF circuits are indeed feasible with diminishing CMOS supply voltages. This is particularly convenient from a system integration point of view, since other components, such as the SRAM standby circuitry already require the availability of a low voltage supply to reduce leakage.

4. CONCLUSIONS

In this chapter we have described circuit design techniques that reduce the cost, decrease the power consumption, and increase the level of integration of wireless sensor nodes. Advanced power management and SRAM leakage suppression techniques are used to reduce digital power consumption. Subthreshold biasing and RF-MEMS technologies facilitate new RF-CMOS transceiver design techniques, enabling previously unattainably low power consumption and levels of integration.

REFERENCES

[1] Rabaey, J. M., Ammer, J., Karalar, T., Li S., Otis, B., Sheets, M., Tuan, T., 2002, PicoRadios for wireless sensor networks: The next challenge in ultra-low power design, *IEEE ISSCC Digest of Technical Papers*, 200–201, San Francisco, February2002.

[2] Chatterjee, B., et al, 2003, Effectiveness and scaling trends of leakage control techniques for sub-130 nm CMOS technologies, *Proc. ISLPED 2003*, Seoul, Korea.

[3] Kao, J., Narendra, S., Chandrakasan, A., 2002, Subthreshold leakage modeling and reduction techniques, *Proc. IEEE/ACM International Conference on Computer Aided Design 2002,* Piscataway, NJ, USA, 141–148.

[4] Mizuno, H. and Kawahara, T., 2001, Chip OS: Open power-management platform to overcome the power crisis in future LSIs, *IEEE Proc. ISSCC2001*, **463**, 344–345.

[5] Horiguchi, M., Sakata, T. and Itoh, K., 1993, Switched-source-impedance CMOS circuit for low standby subthreshold current giga-scale LSIs, *IEEE J. Solid-State Circuits*, **28**(11), 1131–1135.

[6] Qin, H., Cao, Y., Markovic, D., Vladimirescu, A., Rabaey, J., Standby supply voltage minimization for deep sub-micron SRAM, *IEEE Microelectronics Journal*, Aug 2005, **36**, 789–800.

[7] Ruby, R., Bradley, P., Larson III, J., Oshmyansky, Y., Figueredo, D., 2001, Ultra-miniature high-Q filters and duplexers using FBAR technology, *IEEE ISSCC Digest of Technical Papers*, pp.120–121, San Francisco, February 2001.
[8] Otis, B. and Rabaey, J., 2003, A 300 μW 1.9 GHz CMOS oscillator utilizing micromachined resonators, *IEEE J. Solid State Circuits*, **38**, 1271–1274.
[9] Otis, B.P., Chee, Y. H., Lu, R., Pletcher, N. M., Rabaey, J.M., An ultra-low power MEMS-based two-channel transceiver for wireless sensor networks,'' Digest of Technical Papers. 2004 Symposium on VLSI Circuits, 2004, pp. 20–23, *Honolulu*, 17–19 June 2004.
[10] Choi, P., et al, 2003, An experimental coin-sized radio for extremely low power WPAN (IEEE 802.15.4): Application at 2.4 GHz, *IEEE ISSCC Digest of Technical Papers*, 92–93, San Francisco, February 2003.
[11] Zorzi, M. and Rao, R., 2003, Geographic random forwarding (GeRaF) for ad hoc and sensor networks: Energy and latency performance, *IEEE Transactions on Mobile Computing*, **2**(4) Oct–Dec 2003.
[12] Lin, EA., Rabaey, J.M., Wolisz, A., Power-efficient rendez-vous schemes for dense wireless sensor networks, 2004 *IEEE International Conference on Communications*, **7**, 3769–3776, Paris, June 2004.
[13] Vittoz, E. and Fellrath, J., CMOS analog integrated circuits based on weak inversion operation, *IEEE J. Solid-State Circuits*, **12**, 224–231, June 1977 .
[14] Enz, C., Krummenacher, F., Vittoz, E., 1995, An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications, *Analog Int. Circ. Signal Proc. J.*, **8**, 83–114.
[15] Hajimiri, A., Lee, T.H., 1999, Design issues in CMOS differential LC oscillators, *IEEE J. Solid-State Circuits*, **34**(5), 717–724, May 1999.

Chapter 2.2

WIRELESS CONNECTIVITY FOR IN HOME AUDIO/VIDEO NETWORKS

Gerhard Fettweis, Ralf Irmer, Marcus Windisch, Denis Petrovic, and Peter Zillmann
Vodafone Chair Mobile Communications Systems, Technische Universität Dresden
{fettweis, irmer, windisch, petrovic, zillmann}@ifn.et.tu-dresden.de

Abstract At the advent of wireless connectivity of our audio/video entertainment equipment at home, we should reflect the advantages and use cases to understand the technology needed. The required data rate for high quality audio/video entertainment is 1 Gbit/s. As an example of the wireless gigabit system with advanced multimedia support (WIGWAM) system concept, the path to a system meeting these requirements is drawn. Technological challenges are shown as examples of dirty radio frequency (RF) effects.

Keywords audio and video connectivity; dirty RF; wireless short-range communications

1. MOTIVATION

1.1. Data Rate

Cables today provide sufficient data rates for audio and video applications in office, home, and public access scenarios. However, these cables will be replaced to a substantial amount by wireless connections, provided reliable standardized mass-market connectivity solutions with sufficient throughput will be available. Markets for in-house connectivity include office, entertainment, and productivity. From a technical point of view, A/V streaming for entertainment is almost equivalent to audio/video (A/V) conferencing. The net data rate requirement for high quality video is about 20 Mbit/s (e.g., with high quality H.264, or HDTV), and in ad-hoc net-

S. Mukherjee et al. (eds.), AmIware, 51–72.

works, up to three hops can be involved. With the medium access (MAC) overhead, the required gross data rate per user adds up to be 100 Mbit/s. With multiple users and applications in one room or area, in which the wireless medium has to be shared and also with the highly bursty nature of multimedia traffic, the required data rate for A/V applications is 1 Gbit/s.

This data rate is necessary for cable replacement and for new applications, which are just enabled by wireless connectivity. Other crucial requirements for home networks are self-configuration, zero-maintenance, and quality of service (QoS).

From a user requirement perspective, the breakthrough for A/V networks is at 1 Gbit/s. But when will technology provide this data rate? If we look at the development of user data rates in cellular systems in Figure 2.2–1, there is a factor of five in data rate increase every four years if we go from global system for mobile (GSM) to global packet radio service (GPRS), universal mobile telecommunication system (UMTS), high-speed downlink packet access (HSDPA) etc. The same applies to short-range wireless local area network (WLAN) data rates, as shown in Figure 2.2–2 for IEEE 802.11, 802.11b, 802.11a/g, and 802.11n. For A/V networks, WLAN-type networks are relevant. One Gbit/s is currently envisioned by several research initiatives, and a broad commercial application can be expected by 2010.

1.2. Future Seamless Indoor/Outdoor Broadband Connectivity

Seamless broadband connectivity is key to future consumer products. However, there will be no single dominating standard but many standards for different environments and applications. There will be devices, which will be capable to communicate via multiple standards. This is really a challenge because all these standards have to be integrated and approved. Also, these devices may have size and energy constraints. Additionally,

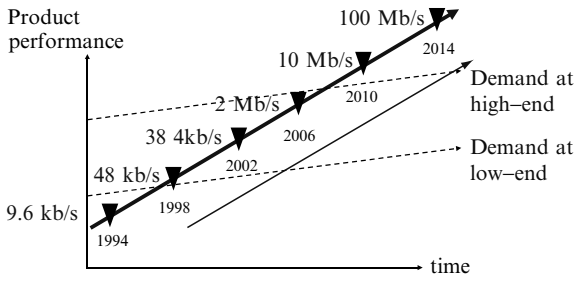


Figure 2.2-1. Data rate development for cellular networks.

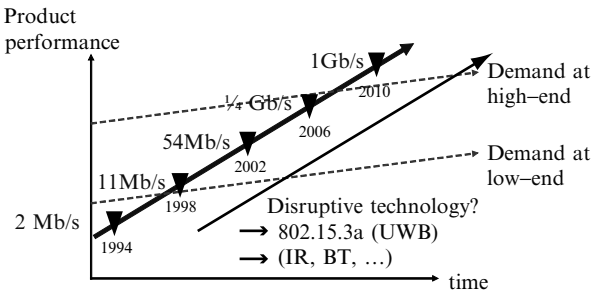


Figure 2.2-2. Data rate for wireless LAN networks.

seamless session handover, concurrent operation, and power management have to be handled. As an example, Figure 2.2–3 shows the integration roadmap of wireless standards into handsets. Currently, *only* GSM, UMTS, Bluetooth and IrDA are included in most handsets. Thus, the mobile handset can be seen as a *black hole of integration*. A similar constellation applies to set-top boxes and Internet home gateways.

One driver for this development is the advent of unlicensed mobile access (UMA).This technology provides access to GSM and GPRS mobile services over unlicensed spectrum technologies, including Bluetooth and 802.11. UMA will be an enabler for HSDPA and home networking.

Another challenge comes from the storage memory roadmap for handsets. In 2006, handsets will have 10 GB (= 80 Gbit) storage. This amount of data (e.g., audio/video content, games) has to be transferred to the handset somehow, ideally without any cables. With a data rate of 0.5 Gbit/s, the time to transfer this data will be 3 or 10 min without and with overhead, respectively. To upload a 100 GB hard disk of a handset, it would take about 1.5 h. Thus, 10 Gbit/s will be needed for local interconnect by 2009.

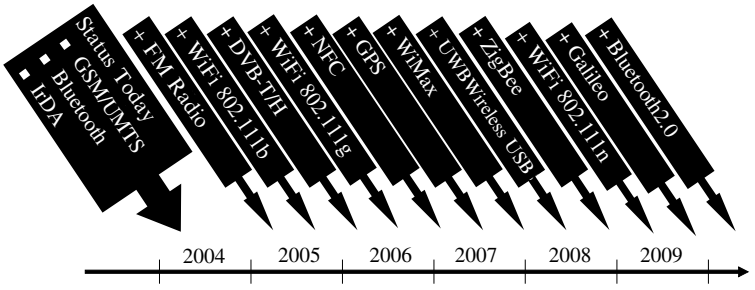


Figure 2.2-3. Integration roadmap of wireless standards into handsets.

Other challenges from a user perspective are real-time intra- and inter-standard handover for voice, data and A/V streaming applications

1.3. Technology Challenges

To meet the user requirements, various technology challenges have to be faced. The power consumption for mobile devices cannot be scaled with the growing data rate as shown in Figures 2.2–1 and 2.2–2. Thus, the power efficiency per bit has to be increased. This requires innovations in RF & mixed signal processing and in digital signal processing.

The driving trends of wireless standards and semiconductor components are:

- Higher carrier frequency
- Lower supply voltage
- Smaller scale and feature size.

This leads to new technology challenges, since physical effects, which used to be negligible become dominant. Figure 2.2-4 shows examples of these impairments, such as phase noise, nonlinearities, IQ imbalance, aperture jitter, sampling, ambiguity, and flicker noise. So far, transceivers have been designed in a way to keep the analog (RF) domain widely separated from the digital signal processing design. For wireless A/V connectivity with data rates at 1 Gbit/s and beyond, a paradigm shift is necessary. By the *Dirty RF* approach, digital signal processing algorithms are designed to cope with a new level of impairments, allowing leeway in the requirements set on future RF subsystems.

Digital signal processing for wireless A/V is also a technology bottleneck. The challenges are low cost, low power, small area, quick time-to-market, and reconfigurability. Before wireless standards are defined, proprietary solutions are necessary early in the market. Furthermore, standards are constantly enhanced and updated, thereby requiring flexibility.

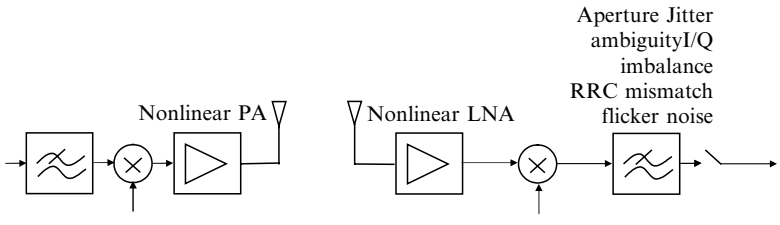


Figure 2.2-4. Picking up "Dirt".

For A/V applications, the sampling rate can be higher than the usual DSP processor clock speed. This requires either tailored application specific integrated circuit (ASIC) design or parallel application specific digital signal processors (DSPs).

2. GBIT/S HOME CONNECTIVITY: GETTING THERE WITH THE WIGWAM PROJECT

It was pointed out in the previous chapter that current wireless standards cannot meet the user requirements for high quality wireless A/V connectivity. Currently there are several initiatives to develop high-data rate short-range communications systems. Standardization in IEEE 802.11n is focused on extending the IEEE 802.11a/g standards with data rates up to 250 Mbit/s. Within the IEEE 802.15.3 group, short-range communications systems are being developed using ultrawide band (UWB) technology or mmWave frequency band technology. There are research activities looking beyond, and among them is the WIGWAM project. A consortium of 10 main contractors (Alcatel, DaimlerChrysler, IHP, Infineon, MEDAV, Nokia, Telefunken Racoms, Philips, Siemens, and TU Dresden) was formed, with the support of 17 German research institutes and universities with funding by the German Ministry of Science and Education (BMBF). The aim of the WIGWAM project is to design a wireless system that is able to support a maximum aggregate data rate of 1 Gbit/s, using resources in the 5 GHz band, with extensions to 17, 24, 38, and 60 GHz. This project includes the development of technologies that allow the adaptation of the user data rate to the specific mobility pattern. Due to the extremely high data rates and carrier frequencies, the currently prevailing technological frontiers will be touched or even have to be extended.

2.1. User Scenarios

In first phase of the project, user scenarios with the accordingly required parameters were defined without considering any technological assumptions. They form a basis for subsequent system concept development. Four scenarios were identified:

- **Home:** Due to its characteristic as a mass market, fulfilling the user needs arising in the home environment constitutes a major design criterion. The massive use of high quality multimedia applications (streaming audio and video) with data rates well in excess of

100 Mbit/s for numerous users is one major reason for the need to have 1 Gbit/s overall throughput. Other key requirements in this scenario are self–configuration, zero maintenance features, and low transmit powers, with a pupose to minimize electromagnetic- radiation exposure.

- **Office**: WLAN solutions available to date have already enabled office staff to work away from their desks to some degree. However, with 100 MBit/s and gigabit ethernet backbones being state of the art, and business applications, such as voice over IP (VoIP) and video conferencing demanding high quality of service, several key challenges are yet to be tackled in order to provide office users with the desired service quality. To ensure the confidentiality of information transmitted over the air, powerful encryption is a core issue.
- **Public access scenario:** In a future heterogeneous wireless network, large scale coverage will be provided by next generation cellular networks whereas high data rate access in urban and hot spot environments will be provided by short range wireless systems. Expected high variations in user data rates and differing service requirements call for a highly flexible MAC. In order to enable user mobility, horizontal and vertical handovers must be supported, i.e., the system will have to interoperate with other B3G standards.
- **High velocity**: The aim is to establish wireless links with vehicles moving at very high speeds (up to 600 km/h) as a backbone access method required to provide coverage within the vehicle. Classical examples for this scenario are bullet trains and cars on highways. The wireless link will be set up to *strings of access points* forming a row, e.g., on the highway's center divide, with line-of-sight connection to cars. The communication is supposed to use a directed beam for the link. Due to having line of sight in typical train scenarios, directional antennas (typical 5°, or 20°) may provide this type of beam. This approach calls for fast and accurate frequency offset correction, as well as technologies that enable reliable high speed (soft) handovers.

2.2. Cross-Layer Design of System Concept

To meet user requirements and to face technology challenges, the system concept has to embrace all layers involved in wireless data transmission. Therefore, five working groups with specialized know-how have been setup to address specific research topics arising in the different layers:

- **System concept:** The task within this working group is to coordinate all research efforts and to use the defined scenarios as well as data from standardization bodies to develop parameter sets that are then used as

guidelines for other working groups. Furthermore, the availability of spectrum resources is evaluated, and the system concept is provided to relevant standardization bodies (e.g., IEEE, ETSI) and frequency regulatory agencies.

- **Hardware platform**: This working group develops technologies that support transmission, reception and processing of data at 1 Gbit/s, thereby calling for novel approaches in antenna technology, analog and digital signal processing, analog-to-digital conversion (ADC), and digital-to-analog conversion (DAC). The DAC's can be seen as a bottleneck at these high transmission speeds. However, within the envisioned time frame, the development of a 12-bit DAC's, operating at 100 MHz is possible. Given the stated bandwidth requirements, the effect of *Dirty RF* significantly affects the system performance. Appropriate analog predistortion as well as digital compensation techniques have to be developed, too. Besides, cost-efficient implementation and a low power budget clearly favor system-on-chip designs solutions for the multiband multiple input–multiple output (MIMO) RF front end.
- **Physical layer**: Transmission of 1 Gbit/s at 100 MHz bandwidth requires a spectral efficiency of 10 bit/s/Hz which can only be achieved by using MIMO technology. After developing channel models for the propagation conditions prevailing in the different scenarios, adaptive modulation and coding techniques will be a key focus of investigation. This will be in order to facilitate the efficient adaptation of the user data rate to the prevailing channel capacities.
- **Link layer**: Wireless communication at such high rates generally requires new MAC techniques that need to be as resource-efficient as possible in order to leave the maximum possible bandwidth for the payload, where high throughput and low latency are key requirements. Very short transmission ranges at high frequencies (60 GHz) require the inclusion of multihop strategies into the developed solutions. Connection with wire-based (IP-based) networks and interoperability with other wireless network standards calls for appropriate hand over mechanisms and convergence layers in such heterogeneous environments.
- **Network layer**: Providing 1 Gbit/s to a mobile user who is moving fast through a system of short range cells, is a challenging task for the mobility management. Cross-layer optimization of network layer and radio resource management will help to develop handover techniques between neighboring cells and different wireless standards that meet the quality of service requirements.

2.3. System Concept Development

Despite ongoing research, the shape of the system concept was visible during midterm of the project in spring 2005.

2.3.1. Frequency ranges

- **5 GHz:** This is the main band considered for the WIGWAM project. The World Radio Telecommunication Conference (WRC-03) agreed in July 2003 to allocate a total of 455 MHz in the bands 5.150–5.350 GHz and 5.470–5.725 GHz for wireless access systems including RLANs. To protect radar and satellite operations, dynamic frequency selection (DFS) and transmit power control (TPC) are mandatory.
- **17 GHz:** In Europe, there is a band of 200 MHz bandwidth at 17 GHz dedicated to RLANs to be used on a non-protected and noninterference basis. This is an extension band for the WIGWAM system concept.
- **24 GHz:** There is an unlicensed band at 24.0–24.25 GHz with 250 MHz bandwidth available in Europe and the USA.
- **38 GHz:** In Europe, there is a band at 38 GHz dedicated to directional communication links. For example, the special high velocity train system TRANSRAPID is operating at 38 GHz. The WIGWAM high-speed scenario will operate at this frequency.
- **60 GHz:** This mm-wave band has attracted a lot of attention recently, e.g., in the IEEE 802.15.3c Millimetre Wave Interest Group. There is large bandwidth available in the range of 59–64 GHz. The limited propagation range, and therefore high interference isolation, makes this band especially attractive for home and office environments. The design of low-cost integrated transceivers at 60 GHz is extremely challenging. Only recent technological progress has made 60 GHz realistic for mass-market applications.

2.3.2. PHY

Orthogonal frequency division multiplexing (OFDM) has been chosen as modulation format. To support IEEE 802.11g devices in the same frequency range, bandwidth has been kept flexible in $n \times 20$ MHz steps, up to 100 MHz. Multiple antennas at the transmitter and the receiver can be used for spatial multiplexing or link quality enhancement. The concept adapts the transmission scheme flexibly to the amount of channel state information (CSI) at the transmitter. Thus, scenarios with CSI, with long-term statistical, or no CSI at the transmitter are supported. With four transmit and

four receive antennas, four parallel MIMO (multiple input–multiple output) data streams can be transmitted. However, a spatial equalizer is necessary at the receiver. The challenge is to provide sufficient performance in terms of packet error rate with a reasonable computational complexity. Linear processing, successive interference cancellation and sphere decoding (SD) are among the suitable algorithms [2].

Advanced coding concepts like low density parity check (LDPC)-codes, or multilevel coding are considered (see [1] and references therein). They provide a performance gain compared to conventional convolutional codes, and have the potential to have a lower complexity. However, since convolutional codes are used for a long time in many standards, highly optimized hardware solutions exist now. Within this project, hardware solutions for LDPC codes are developed.

Another design parameter of OFDM systems is the length of the guard interval. Channel measurements for the WIGWAM bandwidth and carrier frequency were conducted by MEDAV and TU Ilmenau. Delay window, delay spread, coherency time, angular spread, and spatial correlation are.determined in various environments and with several antenna configurations [3].

The subcarrier spacing, FFT size and symbol length are other design parameters. It is advantageous to fix the subcarrier spacing to a multiple or fraction of the 802.11a/g values for easier reuse of hardware components.

The resulting PHY data rate is:

$$\text{Data Rate} = \text{spatial streams}\ \underbrace{\frac{\#\text{data subcarriers}}{\text{FFT size}}}_{\text{guard band overhead}}\ \underbrace{\frac{\text{coded bits}}{\text{subcarrier}}}_{\text{constellation size}}\ \underbrace{\frac{\text{info bits}}{\text{coded bits}}}_{\text{code rate}}\ \frac{1}{\text{total symbol length}}$$

Degrees of freedom to increase the data rate of an OFDM system are:

- Bandwidth, and hence the ratio of FFT size and symbol duration, scales the data rate significantly. At 100 MHz, the data rate compared to 20 MHz increases at least by a factor of five, and the relative guard band overhead gets smaller.
- The number of spatial streams has a significant impact. With 2×2 MIMO, the data rate can in principle be doubled, and with 4×4 MIMO, quadrupled. However, the necessary SNR for the same constellation size depends strongly on the used MIMO detection algorithm.

- The constellation size is also important. Going from 64–QAM to 256-QAM, a factor of 8/6 can be achieved. However, phase noise and other *Dirty RF* effects may become dominant.
- The ratio of data subcarriers to the number of overhead subcarriers (pilots, zero subcarriers) also determines the bandwidth efficiency.

There are additional degrees of freedom to increase the system efficiency, but their impact is not as visible in the first place:

- Pilots are necessary for synchronization and channel estimation. In MIMO configurations, orthogonal pilots have to be transmitted from all antennas for channel estimation purposes. There are two main concepts for pilots: preamble-based pilots and superimposed pilots [25]. The first method has no interference problem, but spectral efficiency is wasted for short packets. With interference cancellation, superimposed pilots may have capacity advantages.
- Signaling information and other applications have very short data bursts (like acknowledgements, web-browsing mouse-clicks). If frame aggregation cannot be applied, the utilization of large OFDM symbols will be poor. Tailored OFDMA packets or superimposed multi-carrier spread-spectrum (MC-SS) signaling are two methods for increasing the efficiency.
- Scheduling, link adaptation, bit loading and power allocation for multiuser MIMO systems are extremely important. They have to be considered right at the beginning of a system concept development [26].

2.3.3. MAC and mobility support

The MAC schemes are developed to suit the user scenarios. Thus, both centralized and ad-hoc options are considered. As multiple access schemes, both OFDMA and MC-CDMA are suitable. Mobility, i.e., seamless handover to 2G, 3G, and WLAN networks is also provided by the WIGWAM system concept.

2.4. RF Impairment Characterization and Correction

When considering high data rate communications systems, the impact of imperfect analog hardware components, known as RF impairments, cannot be neglected. Concepts for coping with these impairments include the avoidance of critical frequencies (e.g., zero subcarriers low frequencies, in order to avoid impairments due to DC offset and flicker noise) and the digital compensation of the analog impairments.

In this chapter, three *Dirty RF* effects are characterized exemplary, and compensation methods are discussed. Again the available technology

at the market-entry point should be already considered at the system concept definition stage.

2.4.1. I/Q imbalance characterization and compensation

2.4.1.1. Impact of I/Q imbalance

One of the most critical RF impairments is the so-called I/Q imbalance, which leads to a limited rejection of the image signal. The performance of the receiver is characterized by the image rejection ratio (IRR). For multi-carrier systems, the I/Q imbalance in direct-conversion receivers translates to a mutual interference between pairs of symmetric subcarriers with respect to the DC subcarrier. As a consequence, symbol errors due to the I/Q imbalance arise.

Understanding the quantitative link between hardware parameters (such as the IRR) on the one hand and system level parameters (such as the symbol error probability) on the other is essential for design and dimensioning of a communication system. A digital compensation of I/Q imbalance is possible, resulting in an increased IRR. Therefore, the numbers presented in this section fulfill two purposes — first, to determine whether or not the system level requirements can be met without digital compensation; and second, how much total IRR (both analog and digital) is required, if a digital I/Q imbalance compensation is applied.

In our analysis we considered an independent fast fading in each subcarrier, which is the worst case from I/Q imbalance point of view. In this case, the impact of I/Q imbalance can be interpreted as a degradation from the real signal-to-noise-ratio (SNR) at the receive antenna to an effective SNR, which is:

$$\frac{1}{SNR_{\text{eff}}} = \frac{1}{IRR \times SIR} + \frac{1}{SNR}$$

The term SIR denotes the signal-to-interferer-power-ratio before downconversion, where the *interferer* is the subcarrier at the image frequency. Most practical systems (e.g., the IEEE 802.11a WLAN) are designed such that all data subcarriers have the same power, i.e., $SIR = 1$. However, $SIR \neq 1$ might

Table 2.2-1. Required effective SNR.

Target symbol error probability	Modulation order	
	64-QAM	256-QAM
10^{-1}	24.9 dB	31.2 dB
10^{-2}	34.9 dB	41.2 dB
10^{-3}	44.9 dB	51.2 dB

be relevant, e.g., in OFDMA systems. Table 2.2-1 shows the required minimum SNR_{eff} at the highest WIGWAM modulation orders (64QAM, or 256QAM) for reaching three exemplary target symbol error probabilities.

The presence of I/Q imbalance leads to a reduction of SNR_{eff}. In other words, a higher channel SNR is required in order to reach the same SNR_{eff} after degradation by I/Q imbalance. The design of the receiver (analog front-end with digital compensation) should be such that the SNR degradation due to I/Q imbalance is small at the targeted symbol error probability. With this constraint the required minimum IRR can be derived. Table 2.2–2 shows the resulting values for an allowed SNR degradation of 1 dB and 0.1 dB, respectively. For the purpose of a clear presentation, $SIR = 1$ was assumed here. The requirements to the image rejection rise, if the case $SIR < 1$ becomes relevant within the WIGWAM system concept.

2.4.1.2. I/Q imbalance compensation

The most intuitive approach for an estimation of I/Q imbalance parameters is to feed the I/Q mixer with dedicated calibration or training signals. However, such techniques are limited to a certain class of communications standards with the presumed pilots. Furthermore, in a practical scenario the pilots are likely to be affected not only by the receiver I/Q imbalance, but also by other impairments, such as the transmission channel and various RF impairments at the transmitter and the receiver. Because all these effects have to be considered, an accurate pilot based I/Q imbalance compensation is likely to be very complex. The dependence on known pilots is avoided by applying *blind* signal processing techniques. The compensation of I/Q imbalance in the time domain can be done by blind signal separation (BSS) [4]. In [5], an approach for blind estimation of the unknown I/Q imbalance is proposed. It does not require any pilot signals, which makes it independent from the targeted communications standard.

Figure 2.2-5 shows the achievable symbol error rate (SER) for an IEEE 802.11a/g – like OFDM system transmitted over a frequency-selective channel (ETSI H/2 A). The blind I/Q imbalance parameter estimation is

Table 2.2-2. Required image rejection ratio (IRR).

Target Symbol Error Probability	1 dB SNR Degradation		0.1 dB SNR Degradation	
	64-QAM	256-QAM	64-QAM	256-QAM
10^{-1}	31.8 dB	38.1 dB	41.3 dB	47.6 dB
10^{-2}	41.8 dB	48.1 dB	51.3 dB	57.6 dB
10^{-3}	51.8 dB	58.1 dB	61.3 dB	67.6 dB

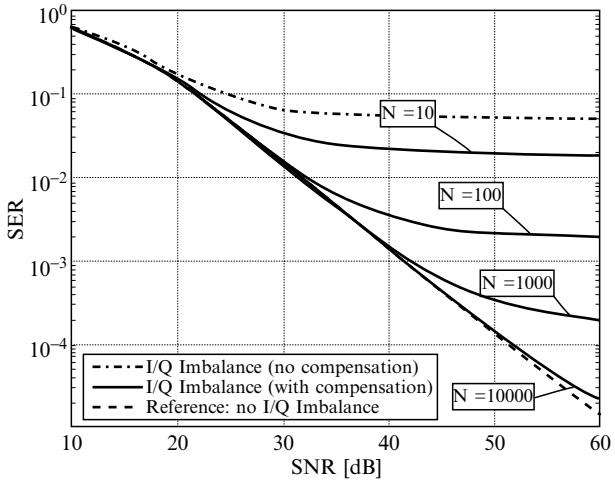


Figure 2.2-5. Symbol error rate versus SNR, 5% gain imbalance, 5° phase imbalance.

done base on unknown data symbols only (48 out of 64 subcarriers total), i.e., no pilots are used. The undesirable error floor due to the I/Q imbalance decreases to arbitrary low levels as the number of received OFDM symbols N increases. Therefore, the proposed estimation technique allows for a flexible trade-off between accuracy, computational effort, and measurement time. In order to fulfill the requirements of a chosen communications standard, the effects of I/Q imbalance can be compensated digitally, while scaling down the demands to the analog part of the receiver.

2.4.2. Phase noise characterization and compensation

The true test of the practicality of the system design is measured by the performance under impairments caused by the RF subsystems. As robust as OFDM systems are, e.g., to frequency selective fading, they are equally sensitive to some nonlinear distortions. In this section we will focus on the investigation of distortions caused by the phase noise in oscillators. The system becomes sensitive to phase noise with the use of bandwidth efficient higher order modulations, higher frequency bands, e.g., 60 GHz, and with decreasing the subcarrier spacing. Since these are some of the directions for increasing the data rates within the WIGWAM project, characterization and suppression of the effects caused by phase noise are important.

Let us consider a coded OFDM transmission model. We assume a system with N subcarriers and symbol duration $N_{\text{tot}} = N + G$ samples, where N and G correspond to the duration of a useful part and a cyclic prefix (CP), respectively. Without loss of generality, we consider OFDM

symbols in an interval $[-G, \ldots, N-1]$, with the useful signal part in the interval $[0, \ldots, N-1]$.

OFDM transmission is block-based transmission. Time domain samples of the useful symbol part are obtained by applying inverse discrete Fourier transform (IDFT) on the data vector $\mathbf{X}$ of length N in the frequency domain. A CP is added to the useful signal part $\mathbf{x}$. The resulting signal $\mathbf{s}$ is upconverted to RF and transmitted through the channel, which is described by an impulse response $\mathbf{h} = [h(0), h(1), \ldots, h(N_c - 1)]^T$.

For simplicity, we adopt a direct conversion receiver, which means that both upconversion and downconversion are done in one step. Oscillators at the transmitter and receiver are ideally harmonic functions of the form $x_c(t) = e^{j2\pi f_c t}$, where f_c stands for the carrier frequency. However, phase noise is inherently present in oscillators and its effect is equivalent to a random phase modulation of the carrier, thus the imperfect carrier has the form $x_c(t) = e^{j[2\pi f_c t + \phi(t)]}$.

2.4.2.1. Phase noise model

It is found that the phase noise of free-running oscillators is well modeled as a Wiener process. In modern transceivers however, local oscillators are frequency synthesizers, which are realized using phase locked loop (PLL). In our work we have adopted the PLL phase noise model developed by A. Mehrotra [7]. Other authors use much more simplified models. They model phase noise as a colored Gaussian noise with the specified power spectral density (spectral mask). Detailed discussion on the phase noise models can be found in [8] and references therein. Further discussion in this section does not depend on the phase noise model that is used., However, all algorithms can be applied for a general phase noise model. Simulation results provided in this document are obtained by using a Wiener process phase noise model. The quality of an oscillator is described with the relative phase noise bandwidth δ_{PN}. This parameter is defined as the ratio between an oscillator 3 dB bandwidth and the subcarrier spacing of an OFDM system.

2.4.2.2. Effect of the phase noise on an OFDM transmission

If the phase noise is present at the transmitter and receiver, after down conversion, the received signal vector with the cyclic prefix $\mathbf{r}_{\mathrm{CP}}$ is given as

$$\mathbf{r}_{\mathrm{CP}} = e^{\mathbf{\Phi}_{R_x}} C e^{\mathbf{\Phi}_{T_x}} \mathbf{s} + \mathbf{s}$$

where C is the time domain channel Toeplitz convolution matrix and $\mathbf{\Phi}_{T_x} = \mathrm{diag}(\boldsymbol{\varphi}_{T_x})$ and $\mathbf{\Phi}_{R_x} = \mathrm{diag}(\boldsymbol{\varphi}_{R_x})$ are diagonal matrices, which model

influence of the phase noise at the transmitter and the receiver, respectively. Terms $\boldsymbol{\varphi}_{T_x}$ and $\boldsymbol{\varphi}_{R_x}$ represent vectors of the sampled noise process during the whole OFDM symbol including the CP. Vector $\boldsymbol{\varsigma}$ represents additive white Gaussian noise.

This model can be linearized if one assumes that phase noise does not change much during one OFDM symbol. It follows that $\boldsymbol{\Phi}_{T_x} = e^{j\phi_{av},\,R_x}\mathbf{I}_{N_{\text{tot}}} + j\Delta\boldsymbol{\varphi}_{R_x}$ and $\boldsymbol{\Phi}_{T_x} = e^{j\phi_{av},\,T_x}\mathbf{I}_{N_{\text{tot}}} + j\Delta\boldsymbol{\varphi}_{T_x}$, where $\phi_{av,\,T_x}$ and $\phi_{av,\,R_x}$ represent mean values of the phase noise at the transmitter and the receiver respectively within one OFDM symbol.

At the receiver, after removing the cyclic prefix part and performing an FFT on a linearized model, the resulting signal vector of all subcarriers in the frequency domain is given as:

$$\mathbf{R} = e^{j\phi_{av},\,R_x} e^{j\phi_{av},\,T_x} C_D x + \boldsymbol{\xi}_{\text{ICI}} + \boldsymbol{\eta}$$

where C_D represents a diagonal matrix of channel frequency domain samples, and $\boldsymbol{\xi}_{\text{ICI}}$ represents inter-carrier interference (ICI) vector.

In the last equation, the multiplicative distortion term $e^{j\phi_{av},\,R_x} e^{j\phi_{av},\,T_x}$, common to all subcarriers of one OFDM symbol, corresponds to the constellation rotation by the mean of the phase noise during one OFDM symbol. This term is referred as common phase error (CPE). The common phase error must be corrected to obtain acceptable performance. This is done by estimating CPE term, and then derotating the received constellation. Pilots are used for this purpose, since all the pilots (as other subcarriers also) are rotated by the same angle. Simple averaging of the pilot rotation from the reference will give the rotation angle. Details can be found in [9], [10]. Note that in these references it is assumed, that the phase noise is present only at the receiver. The same principle however can be used if the phase noise is present at the transmitter also, as presented here. If the phase noise is present at both the transmitter and receiver, then to our knowledge, it is only possible to correct for common phase error as it is described above. There are no approaches in the literature that consider suppressing ICI in this case. This is due to the complex structure of the ICI. If the phase noise is present only at the receiver the system model is the same as already presented, except that the phase noise term at the transmitter is excluded. Additionally, the structure of the inter-carrier interference (ICI) is such, that it is possible to obtain some information about the phase noise, which can be used to suppress ICI.

For systems where phase noise is dominant (or present only) at the receiver, the inter-carrier interference can also be suppressed. In sequel we will present our approach for suppressing ICI [11]. Alternative approach can be found in [14].

2.4.2.3. Phase noise at the receiver only

In this case the demodulated carrier amplitudes $R(s)$ at subcarrier $s = 0, 1, \ldots, N-1$ of one OFDM symbol are given as:

$$R(s) = X(s)H(s)\underbrace{J(0)}_{\text{CPE}} + \underbrace{\sum_{\substack{\nu=0 \\ \nu \neq s}}^{N-1} X(\nu)H(\nu)J(s-\nu)}_{\text{ICI}} + \eta(s)$$

Here $X(s)$, $H(s)$ and $\eta(s)$ represent transmitted symbols on the subcarriers, the sampled channel transfer function at subcarrier frequencies and transformed white noise which remains additive white Gaussian noise (AWGN). The terms $J(i)$, $i = -N/2, \ldots, N/2-1$ correspond to the DFT of the realization of $e^{j\varphi_{R_x}(n)}$ during useful OFDM symbol, and are calculated as

$$J(i) = \frac{1}{N}\sum_{n=0}^{N-1} e^{-j2\pi\frac{ni}{N}} e^{j\varphi_{R_x}(n)}.$$

The properties of the ICI have been investigated by many authors [11, 13, 14]. It has been noticed that ICI should not be treated as a Gaussian random variable [11, 14]. In [11], an approach for an exact ICI power calculation is presented.

2.4.2.4. ICI correction idea

A phase noise compensation beyond the simple CPE correction will be possible only if one knows the instantaneous realization of the phase noise process. The already introduced factors $J(i)$, $i = -N/2, \ldots, N/2-1$ represent the DFT coefficients (spectral components) of one realization of the random process $e^{j\varphi_{R_x}(n)}$. The more spectral components $J(i)$ of the signal are known, the more is known of the signal waveform $e^{j\varphi_{R_x}(n)}$, and thus $\varphi_{R_x}(n)$. The signal $e^{j\varphi_{R_x}(n)}$ has the characteristics of a low-pass signal [6] with power spectral density of the form $1/(1+f^2)$, where f denotes the frequency. Additionally, phase noise has a very small bandwidth compared to the subcarrier spacing. Due to the shape of the spectrum of $e^{j\varphi_{R_x}(n)}$, very few low pass spectral components will suffice to give a *good* approximation of the phase noise waveform. This is illustrated by the example in Figure 2.2-6, where it can be seen that already second order approximation gives a much better phase noise approximation than only the DC value. Therefore, knowledge of the coefficients $J(i)$ gives the

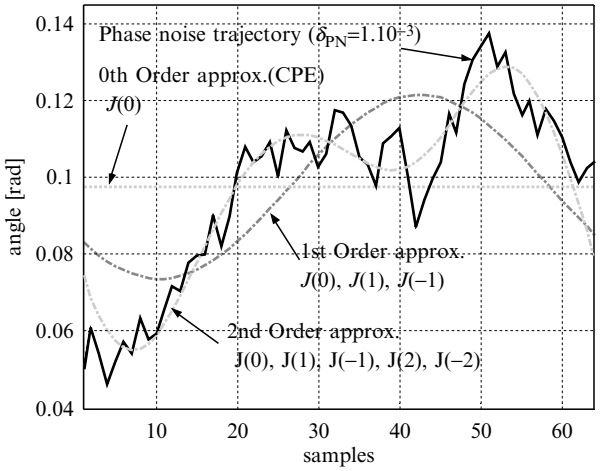


Figure 2.2-6. Phase noise approximation using Fourier series.

possibility to approximate the phase noise waveform to a higher order, and allows a better compensation of it than with CPE correction only.

2.4.2.5. ICI correction algorithm

The details of an ICI suppression algorithm can be found in [11, 15]. The proposed ICI suppression algorithm estimates as many spectral components (i), $i = -N/2, \ldots, N/2 - 1$ as possible using minimum mean square estimation (MMSE). The information about these spectral components is hidden in the ICI part of the signal at the output of the DFT demodulator $R(s)$, $s = 0, 1, \ldots, N - 1$. The estimation algorithm is a decision feedback algorithm, since it requires transmitted symbols. As transmitted symbols estimates, the symbols after necessary CPE correction are adopted. Once the DFT coefficients of the phase noise are known, one possesses enough information on the phase noise waveform, in order to suppress it. The correction can be done in the time domain by multiplying the received signal with $e^{-j\varphi_{R_x}(n)}$, or in the frequency domain as proposed in [11].

2.4.2.6. Iterative phase noise suppression

Described algorithm for ICI suppression is the decision feedback algorithm. It is to expect that falsely detected symbols after initial CPE correction are fed to the MMSE estimator and influence the estimation process. The reduction of the symbol error of the symbols, which are fed back, will improve the quality of the phase noise estimation and thus

the quality of the phase noise suppression. This can be achieved if the algorithm described in the previous section is applied iteratively. The details of the proposed algorithm can be found in [16].

2.4.2.7. Numerical results

System parameters correspond to the IEEE-802.11a standard. We use 64QAM modulation, standard convolutional code with rate $r = 1/2$ and random interleaving. One transmission block consists of 10 OFDM symbols, which comprise one code word. Number of 10,000 packets is transmitted, to assure valid statistics. Hard decision Viterbi decoder is used at the receiver. Within simulations six scenarios are compared: (1) without phase noise (no PN); (2) with phase noise and genie CPE correction (ICPE); (3) with phase noise and CPE correction using least squares (LS) algorithm [13]; (4) with phase noise and genie ICI correction of certain order u; (5) with ICI correction of uth order, and (6) with phase noise and an iterative phase noise suppression (number of iterations denoted by numbers).

A set of simulation results in terms of PER is plotted in Figure 2.2-7 for the ETSI A channel. The adopted relative phase noise bandwidth $\delta_{\mathrm{PN}} = 5 \times 10^{-3}$ is quite large. The ICI correction order adopted is $u = 3$.

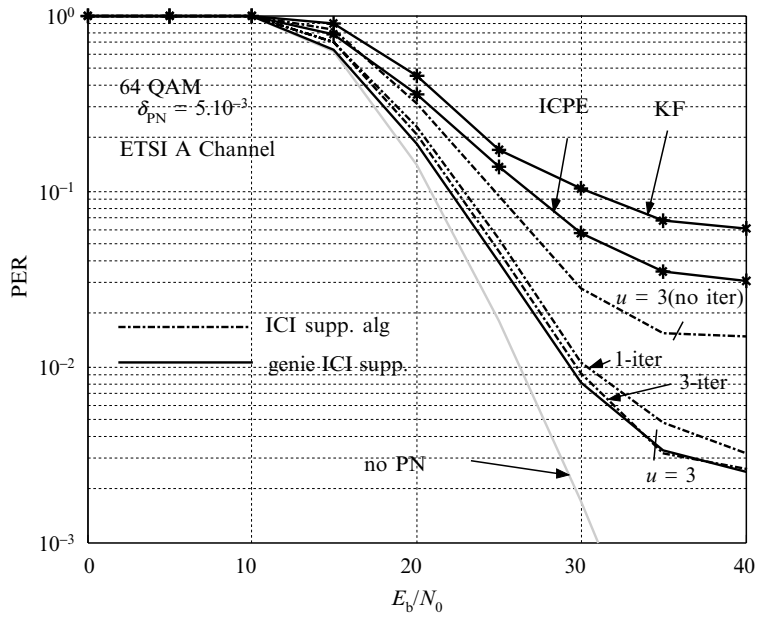


Figure 2.2-7. Performance of phase noise suppression algorithms.

The ICI correction algorithm shows better performance than the pure CPE correction, however, the results are much worse than the achievable genie correction of the specified order. This performance limitation is due to the decision feedback nature of the algorithm. Falsely detected symbols, from a standard algorithm, which are used for estimation of the phase noise DFT coefficients, will influence the estimation process. This problem is more pronounced if the phase noise bandwidth is large, because then ICI is large, which influences also the estimation of the CPE.

To improve the performance of this algorithm, or in other words to reduce the error propagation problem, the iterative approach for phase noise suppression should be considered. This algorithm provides results that are very close to genie phase noise suppression of the corresponding order. However the complexity of the algorithm is quite large. Therefore, the quality of the phase noise suppression is the trade off between complexity and performance.

The performance presented here is in terms of the packet error rate (PER). It is interesting to note that the bit error rate performance can even worsen with increasing number of iterations, while the PER decreases. For OFDM symbols, for which, after the initial CPE correction, many subcarriers are erroneously detected, the ICI estimation can produce additional errors. In the iterative algorithm this causes error propagation. However, for OFDM symbols with only few falsely detected subcarriers, the ICI algorithm is capable of correcting these errors. Packets with few errors will be recovered by the algorithm, while packets with many errors after the initial CPE correction will probably have even more errors.

2.4.3. Nonlinear power amplifier issues

2.4.3.1. Capacity of nonlinear channel

The capacity of multicarrier systems with complex signals impaired by nonlinear distortion and AWGN has been evaluated in [17]. It was shown that severe clipping results in only a moderate reduction of the system capacity. These capacity bounds can be used as an upper bound for evaluation of receive algorithms for nonlinearly distorted multicarrier signals. Furthermore, deliberate clipping at the transmitter, possibly in the digital baseband domain, can be a reasonable approach to developing power-efficient transmitters. However, this comes at the price of more complex receive algorithms.

2.4.3.2. Data detection considering nonlinear PA

The time-domain signal in an OFDM data transmission system is the superposition of many carriers by means of an inverse discrete Fourier

transform (IDFT). This results in an approximately Gaussian distribution of the I-components and Q-components of the complex baseband signal because of the central limit theorem [18]. Consequently, OFDM systems require transmit and receive signal-processing blocks with a high dynamic range, which leads to costly RF components. The high peak-to-average power ratio (PAPR) of the OFDM transmit signal is especially problematic for the power amplifier (PA), as PAs with a large linear dynamic range are less efficient than PAs with a smaller linear dynamic range for a given supply voltage level [19]. There has been active research in recent years in the area of preprocessing of OFDM signals for PAPR reduction (data predistortion) [20, 21] and signal predistortion [19]. The work in [24] follows a third approach: If some nonlinear distortion at the transmitter is allowed, the requirements on the RF front end can be relaxed. Information theory shows that signal clipping in an OFDM transmitter only results in a marginal reduction of channel capacity [17], which motivates the search for powerful receive algorithms. Based on maximum-likelihood (ML) detection, a suitable algorithm is derived in Ref. [24].

Figure 2.2-8 shows the uncoded bit error rate (BER) performance of a clipped OFDM signal (64 subcarriers) with severe clipping (input power back off, IBO = 0 dB) and AWGN. As reference, an unclipped signal affected by AWGN is given. For quadrature phase shift keying (QPSK), clipping introduces severe performance degradation, but the proposed sequential mean-square error reduction algorithm has a similar performance like unclipped QPSK. It even surpasses linear AWGN performance in some SNR regions. The reason is that clipping reduces the average power of the transmit signal, which results in an SNR gain for the clipped system. Clippping and the corresponding detection algorithm can be viewed as an encoder/decodet pair. For 16-QAM, a second iteration of the algorithm results in a further performance gain.

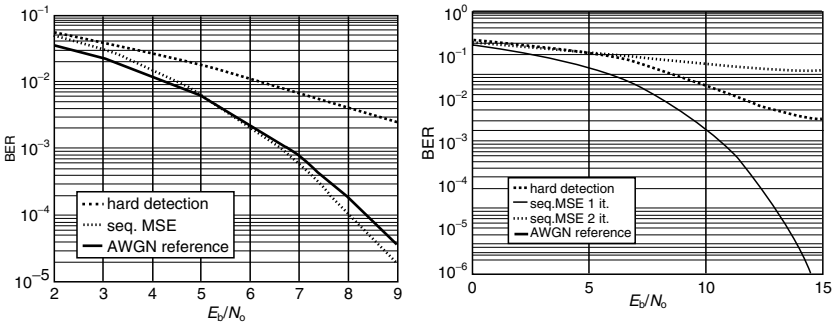


Figure 2.2-8. Performance of sequential MSE reduction for QPSK (left) and 16 QAM (right), 64 subcarrier OFDM, soft limiter back off 0 dB.

3. CONCLUSIONS

High quality wireless connectivity for A/V connectivity cannot be provided with current technology, since the necessary data rate is not available. The WIGWAM system concept shows a path towards technology enabling 1 Gbit/s A/V connectivity, which is required from 2007 and beyond. However, ideas from the concept need to be adopted by international standardization to be successful.

REFERENCES

[1] Zimmermann, E., Pattisapu, P. and Fettweis, G., 2005, Bit-flipping post-processing for forced convergence decoding of LDPC codes, *Proc. 13th European Signal Processing Conference*, Antalya, Turkey, September 2005.

[2] Marsch, P., Zimmermann, E. and Fettweis, G., 2005, Smart candidate adding: A new low complexity approach towards near-capacity MIMO detection, *Proc. 13th European Processing Conference*, Antalya, Turkey, September 2005.

[3] Trautwein, U., Landmann, M., Sommerkorn, G. and Thomä, R., 2005, System-oriented measurement and analysis of MIMO channels, *COST 273 TD(05)063*, Bologna, Italy, January 2005.

[4] Rykaczewski, P., Blaschke, V. and Jondral, F., 2003, I/Q imbalance compensation for software defined radio OFDM based direct conversion receivers, *Proc. 7th Intl. OFDM-Workshop*, Hamburg, Germany.

[5] Windisch, M. and Fettweis, G., 2004, Standard-independent I/Q imbalance compensation in OFDM direct-conversion receivers, *Proc. 9th International OFDM Workshop* (InOWo), Dresden, Germany, 15–16 September 2004.

[6] Demir, A., Mehrotra, A. and Roychowdhury, J., 2000, Phase noise in oscillators: A unifying theory and numerical methods for characterisation, *IEEE Trans. Circuits Syst. I*, **47**(5), 655–674.

[7] Mehrotra, A., 2002, Noise analysis of phase-locked loops, *IEEE Trans. Circuits Syst. I*, **49**(9), 1309–1316.

[8] Petrovic, D., Rave, W. and Fettweis, G., 2005, Properties of intercarrier interference due to phase noise in OFDM, *Proc. International Conference on Communications* (*ICC'2005*). Seoul, Korea, 16–20 May 2005, 2605-2610.

[9] Wu, S. and Bar-Ness, Y., 1998, A phase noise suppression algorithm for OFDM-based WLANs, *IEEE Communications Lett.*, **6**(12), Dec 2002, 535–537.

[10] Robertson, P. and Kaiser, S., 1995, Analysis of the effects of phase noise in OFDM systems, *Proc. ICC*, **3**, 18–22 June 1995, 1652–1657.

[11] Petrovic, D., Rave, W. and Fettweis, G., 2004, Intercarrier interference due to phase noise in OFDM – estimation and suppression, *Proc. IEEE Vehicular Technology Converence* (*VTC Fall*), Los Angeles, September 2004.

[12] Casas, R. A., Biracree, S. and Youtz, A., 2002, Time domain phase noise correction for OFDM signals, *IEEE Trans. on Broadcasting*, **48**(3) September 2002, 230–236
[13] Wu, S.and Bar-ness, Y., 2002, Performance analysis of the effect of phase noise in OFDM systems, *IEEE 7th ISSSTA*, **1**, 133–138.
[14] Piazzo, L. and Mandarini, P., 2002, Analysis of phase noise effects in OFDM modems, *IEEE Trans. Commun*. **50**(10) October 2002, 1696–1705.
[15] Petrovic, D., Rave, W. and Fettweis, G., 2003, Phase noise suppression in OFDM including intercarrier interference, *Proc. International OFDM Workshop* (*InOWo*), Hamburg, Germany, 24–25 September 2003.
[16] Petrovic, D., Rave, W. and Fettweis, G., 2005, Limits of phase noise suppression in OFDM., *Proc. 11th European Wireless Conference* (*EW'2005*). Nicosia, Cyprus, 10–13 April, **1**, 68–73.
[17] Zillmann, P. and Fettweis, G., 2005, On the Capacity of Multicarrier Transmission over Nonlinear Channels, *Proc. IEEE Vehicular Technology Converence (VTC Spring) 2005*, Stockholm, Sweden, May 30–June 1.
[18] Bahai, A. R. S. and Salzberg, B. R., 1999, *Multi-Carrier Digital Communications – Theory and Applications of OFDM*., Kluwer Academic/Plenum Publishers, New York.
[19] Cripps, S. C., 1999, *RF Power Amplifiers for Wireless Communications*, Artech House, Norwood, MA, USA.
[20] Mestdagh, D. and Spruyt, P., 1996, A method to reduce the probability of clipping in DMT-based transceivers, *IEEE Trans. Communications*, **44**, 1234–1238.
[21] Shepard, S., Orriss, J. and Barton, S., 1998, Asymptotic limits in peak envelope reduction by redundancy coding in OFDM modulation, *IEEE Trans. Communications*, **46**, 5–10 January 1998.
[22] Banelli, P., Leus, G. and Giannakis, G., 2002, Bayesian estimation of clipped processes with application to OFDM, *Proc. EUSIPCO 2002*, **1**, 181–184.
[23] Zillmann, P., Nuszkowski, H. and Fettweis, G., 2003, A novel receive algorithm for clipped OFDM signals, *Proc. WPMC 2003*, **3**, 380–384.
[24] Fettweis, G., Löhning, M., Petrovic, D., Windisch, M. and Zillmann, P., 2005, Dirty RF: A new paradigm, *Proc. IEEE International Symposium on Personal Indoor and Mobile Radio Communications* (*PIMRC*), Berlin, September 2005.
[25] Liang, T. -J., Irmer, R. and Fettweis, G., 2004, On the spectral bit rate loss of superimposed pilot and preamble synchronization, in *9th International OFDM Workshop*, Dresden, Germany, 15–16 September 2004.
[26] Jorswieck, E. and Boche, H., 2003, Transmission strategies for the MIMO MAC with MMSE receiver: Average MSE optimization and achievable individual MSE region, *IEEE Trans. Signal Processing*, **51**, 2872–2881.

Chapter 2.3

BODY AREA NETWORKS
The Ascent of Autonomous Wireless Microsystems

Bert Gyselinckx, Chris Van Hoof, and Stephane Donnay
IMEC, Leuven
bert.gyselinckx@imec-nl.nl

Abstract This chapter gives an overview of results of the IMEC's Human++ research program [1]. This research aims to achieve highly miniaturized and nearly autonomous sensor systems that assist our health and comfort. It combines expertise in wireless ultralow power communications, packaging and 3D integration technologies, MEMS energy scavenging techniques and low-power design techniques.

Keywords Autonomous; BAN; integration; microsystem; wireless

1. INTRODUCTION

It is anticipated that microsystem technology will increase the functionality of therapeutic and diagnostic devices to gradually match the needs of a society, which is ageing and spending more money on healthcare. It is expected that by the year 2010, technology will enable people to carry their personal body area network (BAN) [2] that provides medical, sports or entertainment functions for the user (Figure 2.3–1). This network comprises a series of miniature sensor/actuator nodes each of which has its own energy supply, consisting of storage and energy scavenging devices. Each node has enough intelligence to carry out its task. Each node is able to communicate with other sensor nodes or with a central node worn on the body. The central node communicates with the outside world using a standard telecommunication infrastructure, such as a wireless local area or cellular phone network. The network can deliver services to the person

S. Mukherjee et al. (eds.), AmIware, 73–83.

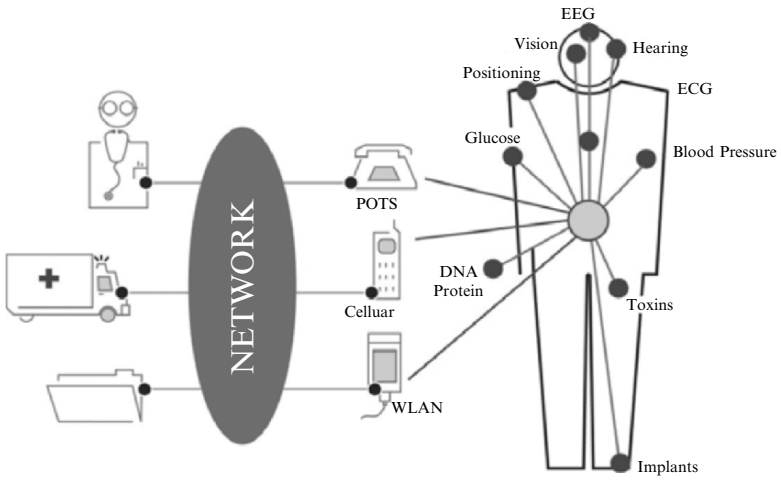


Figure 2.3–1. The technology vision for the year 2010: People will be carrying their personal body area network and be connected with service providers regarding medical, sports, and entertainment functions.

using the BAN. These services can include the management of chronic disease, medical diagnostic, home monitoring, biometrics, and sports and fitness tracking.

The successful realization of this vision requires innovative solutions to remove the critical technological obstacles. First, the overall size should be compatible with the required formfactor. This requires new integration and packaging technologies (Figure 2.3–1). Second, the energy autonomy of current battery-powered devices is limited and must be extended. Further, interaction between sensors and actuators should be enlarged to enable new applications such as multiparameter biometrics or closed loop disease management systems. Intelligence should be added to the device so that it can store, process, and transfer data. The energy consumption of all building blocks needs to be drastically reduced to allow energy autonomy.

2. AMBULATORY EEG AS TESTCASE

Electroencephalogram (EEG) is a monitoring tool used by neurologists to measure the electrical activity of the brain and trace neurological disorders, such as epilepsy. In hospitals, it is typically used during several days and involves hospitalization of the patient. Ambulatory monitoring of the brain activity would improve the patient's quality of life to a great

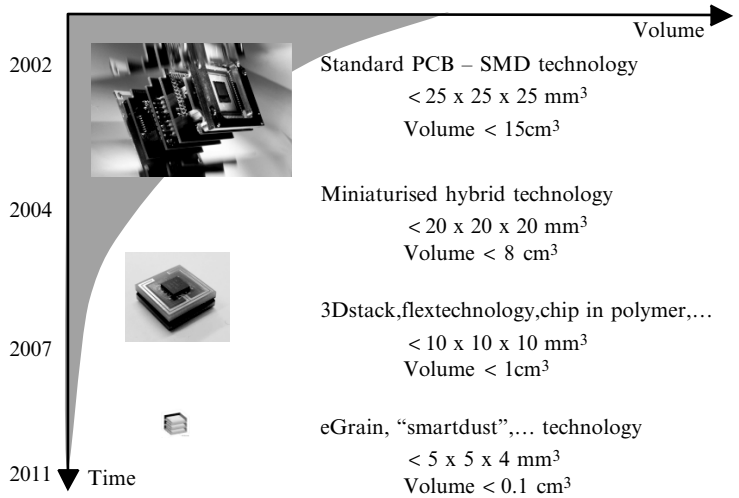


Figure 2.3–2. System integration roadmap for the coming decade — the development of 3D stacking technology, flex technology, and full wafer-scale 3D integration will lead to 2 orders of magnitude reduction in volume and enable smart unobtrusive autonomous sensor systems.

extent and therefore, wireless EEG was selected as a test case. The development started with a prototype using off-the-shelf components. This system consists of a portable, battery-powered transmitter with 24 EEG electrode inputs and a personal *health assistant* receiver that is within the reach of the patient. This assistant stores all activity, and if required, streams EEG data to other monitoring equipment. The transmitter, worn on the patient's body, digitizes 24 channels of EEG data at a sample rate of 256 Hz with 12-bit resolution. This complies with the industry standard. Digitized data is transmitted over a wireless link at 868 MHz. The output power is -10 dBm and the data rate over the air is 75 kbps. The system has an average power consumption of 145 mW. Running on 4AA batteries, an operational lifetime of 3 days is achieved. The system occupies a volume of over 500 cm^3 and is shown in Figure 2.3–3.

In order to improve the convenience of the patient, we used our in-house 3D stack technology to reduce the volume of the system to 1 cm^3 and extend the operational lifetime to 1 month. Assuming that half of the volume is reserved for a Li-battery with a typical energy density of 200 Wh/l, the stored energy is 100 mWh. In order to run 30 days on this energy, the average power consumption has to be lower than 140 μW.

Figure 2.3–3. Child carrying the first-generation ambulatory EEG transmitter.

In the remainder of this chapter we will show how advances in wireless communication, energy scavenging, and system integration can enable such systems in the near future.

3. WIRELESS COMMUNICATION

The radio in the sensor will have to operate at an average power of 50 μW. However, low power radios, such as Bluetooth and Zigbee [3] cannot meet the stringent wireless BAN power requirements. If one takes the body environment and the RF properties of the body into consideration, the power consumption in the sensor node can be brought down by at least one order of magnitude. Networking and MAC protocols can be optimized for the BAN context, taking into account the simple network topologies, the relatively small number of nodes and the (e.g., latency) requirements of body monitoring applications. In Ref. [4] we showed how an electromagnetic (EM)-field tends to stick to the skin like a wave creeping from one side to the other. Figure 2.3–4 shows how a TE-field propagates around the human body.

Thanks to the low data rate of typical sensors the radio can be operated in burst mode with a minimal duty cycle (e.g., with a burst data rate of a few hundreds of kbps and an average data rate of around 1 kbps, leading to a duty-cycle in the range of 0.1–1%).

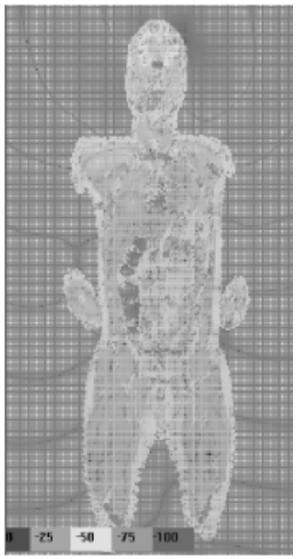


Figure 2.3–4 TE field propagating along the human body.

The power budget in the sensor node and in the master device is very different. The sensor has an extremely tight power budget, whereas the master has a slightly more relaxed power budget. In the air interface definition this asymmetry is exploited by shifting as much complexity as possible to the master device.

For these reasons we have chosen to make use of ultra-wideband (UWB) modulation. This will allow us to use an ultra-low-power, lowest-complexity transmitter and shift as much as possible the complexity to the receiver in the master. We have developed the first silicon of a pulser, which is key building block for a UWB transmitter. The system generates position-modulated pulses that comply with the FCC regulation mask. It operates between 3 GHz and 5 GHz and the signal bandwidth can be tuned from 500 MHz up to 2 GHz as shown in Figure 2.3–5.

The system can deliver a pulse rate up to 40 MHz. The pulses are modulated in position and the position modulation can be tuned from 4 ns to 15 ns.

Figure 2.3-6 shows the chip micrograph.

4. MICROPOWER GENERATION

In order to relax the stringent power consumption requirements of the sensor node, energy scavenging may be an option. Although a rechargeable

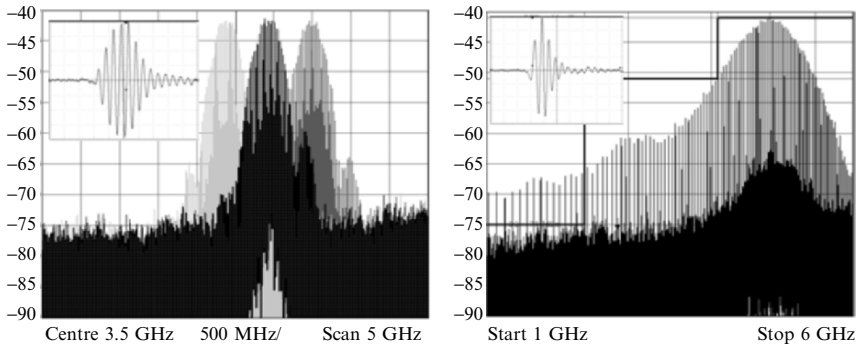


Figure 2.3–5. Measured pulse spectra and time waveform (inset) of pulser ASIC. Left shows 528 MHz wide pulses centered at 3.432 GHz, 3.960 GHz, 4.488 GHz. Right shows 2 GHz wide pulse.

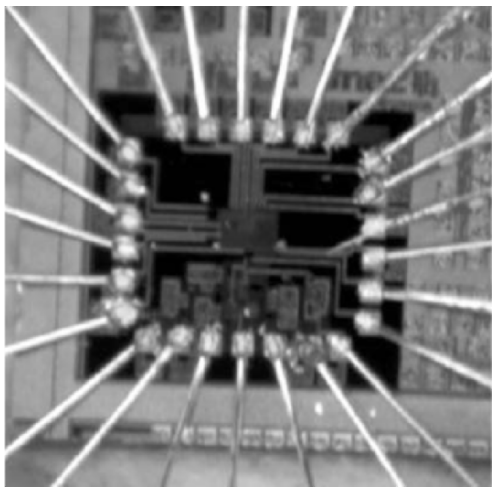


Figure 2.3–6. Pulser die micrograph.

battery will still be needed in conjunction with the scavenging source, this battery will be smaller than in the case of a primary battery. For body applications, mechanical and especially thermal scavengers are well suited as alternatives or complements to solar energy. The human body produces approximately $10\,\mathrm{mW/cm^2}$ of waste thermal energy. Thermopiles can be used to convert this thermal energy into electrical energy. Since individual thermopiles produce only a very limited voltage and power, one needs many of them interconnected electrically in series. Thermally the thermo-

piles have to be interconnected in parallel, so each thermopile can benefit from a maximum heat flux and corresponding temperature difference.

We have realized working mechanical and thermal micropower generation prototypes. The thermal scavenger shown in Figure 2.3–7 generated an average power of 100 μW. The main target for thermal scavenging is miniaturization using MEMS technology for improved SiGe or BiTe thermopiles as shown in Figure 2.3–8.

5. INTEGRATION TECHNOLOGY

One form factor suitable for many applications is a small cubic sensor node. To this end, a prototype wireless sensor node has been integrated in

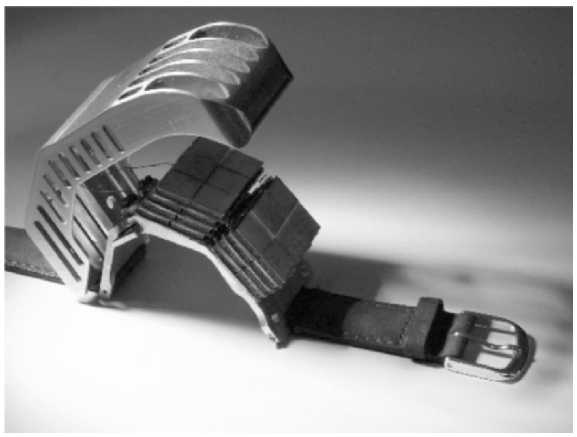


Figure 2.3–7. Thermal micropower generator prototype.

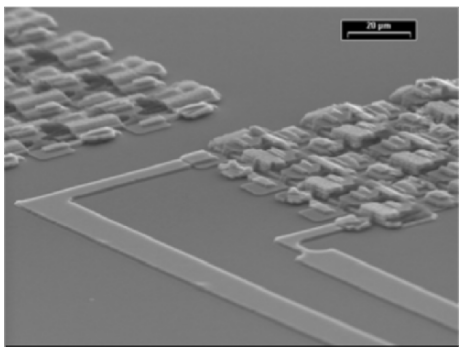


Figure 2.3–8. Thermal micropower generator – detail of partly micromachined module.

a cubic centimeter (see Figure 2.3–9). In this so-called three-dimensional system-in-a-package approach (3D SIP) [5], the different functional components are designed on separate boards and afterwards stacked on top of each other through a dual row of fine pitch solder balls. This system has the following advantages: (1) modules can be tested separately, (2) functional layers can be added, or exchanged depending on the application, (3) each layer can be developed in the most appropriate technology. The first generation 3D stack offers a complete system-in-a-package (SiP) solution for low power intelligent wireless communication. The integrated stack includes a commercial low power 8 MIPS microcontroller [6] and 2.4 GHz transceiver [7], crystals and all necessary passives, as well as a matched dipole antenna custom-designed on the top layer laminate substrate. The bottom layer has a BGA footprint, allowing standard techniques for module mounting. This sensor module has been integrated with the thermal scavenger presented above and this is the basis for sensor networks [8, 9, 10], which, unlike most of their predecessors, are fully energy autonomous.

Recently, parallel research was started to implement the same technology on a 2D carrier. The ultimate target is to create a small and smart band-aid containing all the necessary technology for sensing and communication with a base station. It will provide a generic platform for various types of applications (wound healing, UV-radiation, EEG, ECG, EMG, etc.).

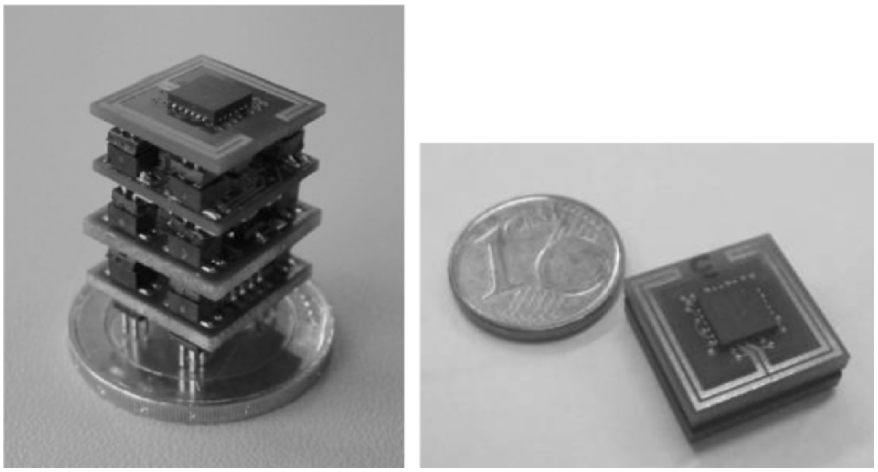


Figure 2.3–9. Wireless sensor module as miniaturized but conventionally-connectorized module (left), or as integrated 1 cm^3 volume 3D stack (right).

The first prototype (Figure 2.3–10) is10 times smaller than a credit card (12×35 mm) and thin as a compact disc (1–2 mm). The flexible 25 μm polyimide carrier contains a microprocessor and a wireless communication module (2.4 GHz radio). It enables IMEC to optimize the antenna for its activity on human skin. Current focus lies on adding the necessary sensors and energy equipment (rechargeable battery, energy scavenger and advanced electronics to keep energy consumption as low as possible). IMEC targets an ultimate device thickness of approximately 100 μm.

The biggest challenges in developing this kind of modules are the extreme miniaturization and its effects on the functionality of the used components. Some of the many problems to tackle are the use of naked chips, chip scaling, assembly processes like wire bonding and flip-chip on a flexible substrate, application of thin-film batteries and solar cells and integration of the entire technology in a biocompatible package.

6. BODY AREA SENSOR NETWORK

A network of the above sensors was demonstrated where the sensors share a single communication medium (a radio-channel in the 2.4 GHz band). The low-duty cycle, non time-critical measurements typical for a network of low-power sensor modules, allow for a time division multiple access (TDMA) method to share the medium [11]. A sensor module's power consumption profile during a TDMA cycle is shown in Figure 2.3–11. This TDMA cycle returns at a specified measurement interval, with the system returning to a 6 μW sleep mode in between.

Figure 2.3–10. Prototype sensor in a flexible band aid.

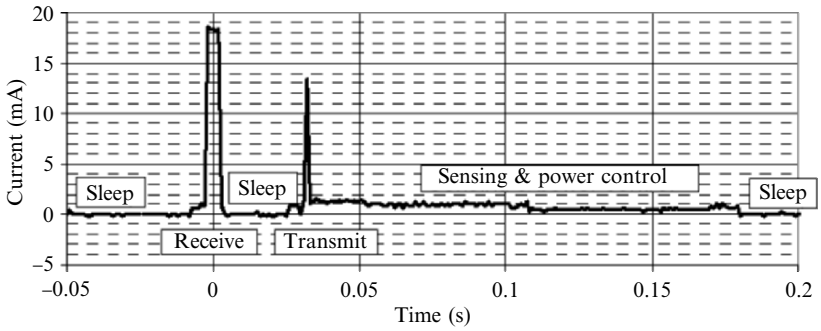


Figure 2.3–11. Power consumption profile of the wireless sensor module [12].

The resulting average power consumption for long measurement intervals and in practical operating conditions is 100 μW.

7. CONCLUSIONS

This chapter gave an overview of the Human++ research program at IMEC, which is targeted at developing key technologies and components for future wireless BAN for health monitoring applications. Several working prototypes have been discussed, such as micropower generation devices and a 1 cm^3 low-power wireless sensor node. This modular wireless three-dimension stack is now used as a platform for the integration of future developments (sensors and actuators, energy scavenging devices, ultralow-power local computing and transceiver) in order to realize fully integrated, autonomous ultra-low-power sensors nodes for body area networks.

REFERENCES

[1] http://www.imec.be/ovinter/static_research/human++.shtml.
[2] Schmidt, R., et al., 2002, Body area network ban, a key infrastructure element for patient-centered medical applications, *Biomed Tech* (*Berlin*), **47**(suppl. 1, part 1), 365–368.
[3] http://www.bluetooth.com, http://www.zigbee.org.
[4] Ryckaert, J., De Doncker, P., Meys, R., de Le Hoye, A. and Donnay, S., 2004, Channel model for wireless communication around the human body, *Electronics Lett*., **40**(9), 543–544.

[5] Stoukatch, S., Ho, M., Vaesen, K., Webers, T., Carchon, G., De Raedt, W., Beyne, E. and De Baets, J., 2003, Miniaturization using 3D stack structure for SIP application, *Proc. SMTA* (*Surface Mount Technology Association*) *International Conference*, September 21–29, 2003, Chicago.
[6] TI MSP430F149, www.ti.com.
[7] Nordic nRF2401, www.nvlsi.com.
[8] Tubasihat, M. and Madria, S., 2003, Sensor networks: An overview, *IEEE Potentials*, **22**(2), 20–23.
[9] Ahmed, A. A., 2003, A survey on network protocols for wireless sensor networks, *Information Technology: Research and Education*, *2003 Proceedings*, 301–305.
[10] Vieira, M. A. M., et al., 2003, Survey on wireless sensor network devices, *IEEE Emerging Technologies and Factory Automation 2003 Proceedings*, **1**, 537–544.
[11] Warneke, B. A., et al., 2002, An autonomous 16 mm3 solarpowered node for distributed wireless sensor networks, *IEEE Sensors 2002 Proceedings*, **2**, 1510–1515.
[12] Torfs, T., Sanders, S., Winters, C., Brebels, S. and Van Hoof, C., 2004, Wireless network of autonomous environmental sensors, *Proc. IEEE Sensors 2004*, October 24–27, 2004, Vienna.

Chapter 2.4

WIRELESS COMMUNICATION SYSTEMS

Neil C. Bird
Philips Research Eindhoven
neil.bird@philips.com

Abstract The vision of ambient intelligence (AmI) brings with it the need for a complex wireless communication infrastructure in the home. Today, some of the requirements can be met with existing wireless standards, but for others results from ongoing research into new wireless systems will be needed to complete the broad portfolio of wireless links that AmI demands. In addition to wireless communication, the home infrastructure will also need to provide location information so that the environment can adapt and respond in an intelligent manner. This chapter discusses these needs from the wireless systems perspective, and also considers the issues that arise during the installation of such a complex system.

Keywords ambient intelligence; 60 GHz; indoor positioning; wireless communication; WLAN; WPAN; ultralow power radio

1. INTRODUCTION

From the perspective of wireless communications, the vision that ambient intelligence (AmI) translates into a set of diverse and demanding requirements for the wireless infrastructure will serve the *smart home* of the future. The concept of AmI is already well-known and documented [1]. It describes future digital environments that are sensitive and responsive to the presence of people. Key characteristics include: *context awareness*—the environment has knowledge of what is happening; *personalization*—actions and responses are tailored to your needs; and *adaptivity*—the environment will change in response to events. Specifically, the vision describes an environment in which the technology is embedded, hidden,in the background, and one that demands the presence of a multitude of invisible distributed devices.

S. Mukherjee et al. (eds.), AmIware, 85–104.

Looking from the technology stand-point, and taking a somewhat more practical view of how such a vision is to be realized, it becomes immediately apparent that wireless data links will form a key part of the communication infrastructure that will be required to connect the large number of devices in the environment. In addition to the sheer number of links that will be required, the scale of the technical challenge is augmented by the range of data rates that are called for, the distances over which these will have to operate, and consequently the need for simultaneous operation and coexistence.

In addition to the (wireless) transfer of data around the environment, in many cases, it is essential to know the location of the source (and destination) of the data. For example, information from a particular sensor is much more useful if the system knows the location of the sensor. This, then, implies that the AmI environment will also be served by an indoor positioning system that is capable of locating people and objects. The use of the wireless communication infrastructure for this purpose is an attractive possibility.

Finally, a key aspect of AmI, and one that is perhaps insufficiently addressed, is the question of how such an environment can be constructed (installed) in the home. Issues, such as configuration of the wireless network and set-up of the indoor positioning system play an important role here. These need to be solved if the set-up of the AmI environment is to be as intuitive and straightforward as the use of the environment itself.

In the following sections of this chapter, we examine the demands that AmI places on the wireless communication infrastructure: first from the data transfer perspective, and then the implications of using the same infrastructure for wireless indoor positioning. The chapter concludes with a brief look at installation and configuration.

2. WIRELESS DATA LINKS

From what has been remarked above, and written in the many other publications on AmI, it is clear that wireless data communication will form the core of any such system. The complexity of the overall solution results from the combination of three diverse aspects:

- the wide range of required data rates (the variety of applications translates into hugely varying data rates);
- the distance over which data needs to be sent (ranging from a meter, to across the whole house, which will strongly influence architecture choices);
- simultaneous operation of multiple links.

The first two of these topics will be discussed in the following sections. The third, although being an essential component, is beyond the scope of this chapter. Suffice to say that in the definition of virtually all wireless standards, consideration of interference, simultaneous operation and co-existence has played an important role and to a greater or lesser extent has been addressed. Further, in the definition of new standards, these issues continue to play a dominant role. For the remainder of this chapter, we take the robustness of a particular wireless standard as a given, and concentrate on how to combine the various standards into a suitable wireless infrastructure.

2.1. Data Rates

Concerning the wide range of data rates, a brief look at the types of applications in, and features of, AmI will illustrate the point. At one end of the scale, there is the need for extremely high data rate links – of the order of 1 Gbit/s. These will be required for the streaming of vast amounts of video data to the high definition television (HDTV) displays that will populate the environment. For example, an uncompressed HDTV display with a resolution of 1920×1080 operating at 60 frames/s with a color depth of 16 bits, results in a data stream of 2 Gbit/s. Other applications that need Gbit/s data rates are those where bulk data transfer is concerned. As an example, consider a MP3 player. Today, these devices commonly have a capacity of 128 MB – approximately 1 Gbit – and within a few years several tens of Gbits will be mandatory as the function of the MP3 player changes from audio to video. In an AmI scenario, the environment would be aware that you are rushing to catch a plane, and most likely will want to take your favorite films with you. Time is of the essence, so as you pass out of the front door, 10 Gbits of data are downloaded into your portable multimedia player. Given that you are within the range of the wireless transmitter for only a few seconds, a data rate of several Gbit/s is therefore required.

At the other end of the data rate range are the wireless links that support the ubiquitous sensors that will be found in the environment. Here, the requirements are very different. Typically, a given sensor will only need to transmit data occasionally—a few times per day—and each message will be of the order of a few hundred bits of information (a 64-bit address, a 128-bit data packet, and a protocol/error correction overhead). In this case, channel capacity is not the issue, but rather the need to minimize the energy per message because the sensor will be powered from a battery or rely on energy scavenging techniques. Low power radio topics are addressed in more detail in Section 2.3.2, which draws the conclusion that the optimum data rate is in the order of tens of kbit/s.

Figure 2.4–1 shows the range of data rates for AmI, and how the various applications are mapped onto these. The immediate point of interest is that the range covers 6 orders of magnitude! This clearly tells us that we are not in a situation where *one standard fits all*, and we will have to deploy and interconnect multiple wireless systems.

The central part of the Figure 2.4–1, from 100 kbit/s to 54 Mbit/s is covered by existing wireless standards, such as ZigBee [2], Bluetooth [3] and the 802.11 variants [4]. Table 2.4–1 compares the key features of these standards.

Commercial solutions for data rates above 54 Mbit/s are not yet generally available. Two technologies are being developed to address this: 802.11n and ultra-wideband (UWB). The first of these, 802.11n, which can employ multi-input-multiple-output (MIMO) techniques, promises data rates of up to 200 Mbit/s, while the current driver for UWB development is the realization of data rates over short ranges of 480 Mbit/s to support the wireless USB standard. In the case of UWB, extending this data rate to 1 Gbit/s seems to be feasible. This could be achieved by employing higher order modulation schemes and trading distance against data rate, or by using allocated spectrum in the higher bands: the FCC allows UWB to operate in between the 3.1 and 10.6 GHz with a maximum transmitted power level of −41.3 dBm/MHz (current implementations for the 480 Mbit/s system use only spectrum between 3.1 and 4.6 GHz). While the use of the additional spectrum to increase the data rate is certainly possible, the return will not be linear because of the larger path losses at the higher frequencies. Consequently, moving into the multi-Gbit/s region will require the definition of a new system. This is discussed in Section 2.3.1.

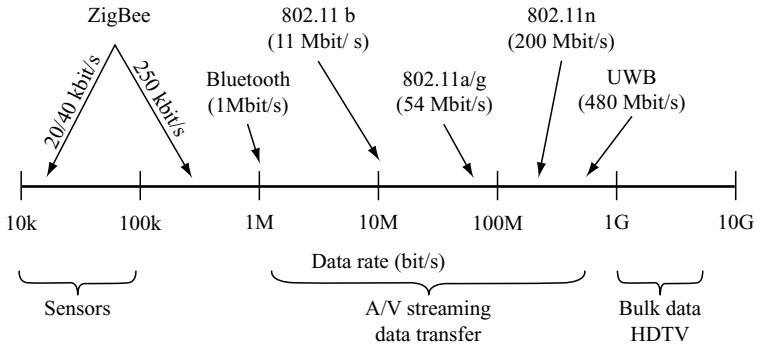


Figure 2.4-1. Data rates for Ambient Intelligence.

Table 2.4-1. Comparison of wireless standards.

Standard	ZigBee	Bluetooth (1.1)	802.11b	802.11a/g
Frequency	2.4 GHz ISM	2.4 GHz ISM	2.4 GHz ISM	5 GHz ISM 2.4 GHz ISM
Max Data Rate	250 kbit/s	1 Mbit/s	11 Mbit/s	54 Mbit/s
Channel	5 MHz	1 MHz	20 MHz	20 MHz
Range	10 m	10 m	$<$ 50 m indoors	$<$ 50 m indoors

At the lower end of the data rate range (e.g., for sensor applications) the most suitable standard is ZigBee (802.15.4). As shown in Table 2.4–1, ZigBee provides over-air data rates of 250 kb/s in the 2.4 GHz band, and 20–40 kb/s in the 868 MHz (Europe) and 915 MHz (USA) bands, respectively. From the data rate perspective, ZigBee is well-suited (and indeed was designed with such applications in mind). However, for many applications in the AmI sphere, further developments are required to meet future size and power consumption requirements. This is considered in Section 2.3.2.

2.1.1. Wireless LAN and PAN

Data rate is only one of the metrics that should be considered when contrasting the various wireless standards. As will be seen in the next section, another important aspect is the distance over which the data can be transmitted, and whether the wireless standard provides for sufficient transmitted power to go through walls.

A convenient categorization of wireless standards that takes into account the transmission distance is the division into wireless local area networks (WLAN) and wireless personal area networks (WPAN). Broadly speaking, WLANs such as 802.11 and its variants are intended for communication over distances of several tens of meters or even 100 m, and are capable of transmitting through walls. However, the attenuation of the signals by walls compromises the maximum achievable data rate to levels below those shown in Table 2.4–1. Conversely, WPAN systems, such as Bluetooth and UWB, are intended for communicating over short distances of typically less then 10 m, and typically aimed at in-room communication. Power levels are generally insufficient for data transmission through walls.

2.2. Infrastructure

Having looked at the wide range of wireless data rates that are required to support AmI, the next step is to consider the architecture of the overall

wireless infrastructure. As remarked earlier, the wireless infrastructure must be capable of supporting data rates that vary over 6 orders of magnitude, and at the same time, handle distances that vary from 1 m to coverage of the entire home.

The two key factors here are: (1) the physical structure of our homes, and (2) the way in which different types of data (and hence data rates) play a role in the environment.

For the first of these, the division of our homes into rooms is fundamental. In many circumstances data will have to be transmitted through one or more walls, which will attenuate the signals and cause reflections (multipath)—both of which will have a deleterious effect on the maximum achievable data rates. In addition, the fact that data can be transmitted through walls in our *own* homes also means that the signals from our network will be present in our neighbors' homes (especially in an apartment block scenario), and conversely *their* network signals will be present in our own home. Apart from issues of security, this will significantly increase the level of interference and further impact the maximum data rates between rooms.

The second major determinant in the design of the wireless infrastructure is the type of data transfers that are required, and whether the data needs to be transmitted within a room, or between rooms. Table 2.4–2 shows a categorization of the different data types.

In Table 2.4–2, a tacit assumption has been made, for the time being, that the majority of audio and video content is stored on a central server in the home, and that this central server is connected to sources of content coming into the home from outside. This explains why all of the streaming data, including broadcast, will need to cross room boundaries. However, such a set-up would place unfeasible requirements on the wireless system. Take, for example, the situation where HDTV is being streamed from one room to another. In order for this data to go through walls, the maximum

Table 2.4-2. Data types and characteristics.

Data Type	Data Rate (single link)	Comment
A/V streaming		
Broadcast TV	15 Mbit/s max, 3–8 Mbit/s	Dig. Broadcast (MPEG2)
HDTV (compressed)	19 Mbit/s (USA), 24 Mbit/s (Japan)	
HDTV (uncompressed)	2 Gbit/s	
DVD	10.8 Mbit/s max, 3–8 Mbit/s avg	MPEG2
Bulk Data Transfer	Several Gbit/s	
Internet	1–4 Mbit/s	
Control	10–100 kbit/s	e.g., remote control
Status/Environment	10–100 kbit/s	e.g., sensor networks

RF frequency is limited to around 6 GHz and apart from other considerations, sufficient bandwidth would have to be found at these frequencies. However, since the RF signals *do* go through the walls there will also be interference between our HDTV streaming application and potentially those of our neighbors. Returning to the scenario of an apartment, this could mean that up to 9 wireless HDTV streams could be coexisting (you, your neighbors on either side, and the neighbors above and below). This results in an aggregate payload data rate of 18 Gbit/s and assuming a 50% overhead for protocol and error correction, results in an over-air data rate of 27 Gbit/s. Even if as much as 1 GHz of bandwidth was to be allocated for such applications, the spectral efficiency would need to be close to 30 bits/s/Hz. This is far beyond what can be achieved with current standards, and when this requirement is combined with the power levels that would be required to penetrate walls, only the most optimistic designer would say that this approach will ever be feasible.

However, the need for wireless streaming remains a fundamental part of AmI, so an alternative structure is required. The solution is to construct a wireless infrastructure that uses a WLAN system for inter-room communication and WPAN systems for intra-room links. However, as discussed above, this will only give limited data rates between rooms. In addition, therefore, the infrastructure needs to incorporate large amounts (GBs) of local storage in each room, and use this as the data source for streaming applications. The local storage then becomes a buffer to mitigate the difference between the bandwidths of WLAN (inter-room) and WPAN (intra-room). In this way, the wireless streaming of HDTV and bulk data becomes an in-room application. Table 2.4–3, shows how the characteristics of the different data types change as a result.

Table 2.4-3. Data types and characteristics (local storage).

Data Type	Distance	Comment
A/V streaming		
Broadcast TV	Room-to-room	
HDTV (compressed)	In room	Also room-to-room when broadcast
HDTV (uncompressed)	In room	
DVD	In room	
Data transfer		
Fast bulk	In room	Several Gbit/s
Transfer to local storage	Room-to-room	e.g., overnight data transfer (tens of Mbit/s)
Internet	Room-to-room	
Control	In room	Room-to-room via the WLAN backbone
Status/Environment	In room	Room-to-room via the WLAN backbone

Now only the broadcast stream and internet access are required to operate between rooms. In addition, a new bulk data transfer category has been added, for data, which is required to cross room boundaries. This is to allow non-real-time download of data from broadcast sources (satellite, cable, etc.) to the local storage (e.g., overnight).

Figure 2.4–2 shows the details of a hierarchical wireless infrastructure employing both WLAN and WPAN systems.

Here, each room is served by one or more WPANs, possibly Bluetooth or UWB depending on the required data rates. In either case, the important feature is that the RF signals should not leave the room. This can be achieved by ensuring that the amount of RF power transmitted is kept to a minimum, or by using RF frequencies above 10 GHz. In this way, inference between WPANs in different rooms is prevented, and maximum spatial reuse of spectrum can occur. Returning to the HDTV streaming example, if an in-room WPAN is used to stream the data from local storage to the display, then adjacent rooms (or neighbors) can each accommodate such an application without problems.

For inter-room communication, a WLAN system is employed, and this supports the lower data rates that are needed for broadcast streaming and data transfer between rooms. Taking current developments into account (e.g., 802.11n and MIMO techniques), this part of the wireless infrastructure will, in the future, be able to support over-air data rates of hundreds of Mbit/s and point-to-point links of 10–20 Mbit/s. In this way, by combining both WLAN and WPAN in one infrastructure, it seems possible to satisfy the various demands for wireless data rates by taking into account the need to transfer data both within rooms and between rooms.

An alternative for room-to-room communication is to use a wired backbone, based on optical fiber or low-cost coax. Such a solution solves the issue of high data rate transfers between rooms, and is an attractive

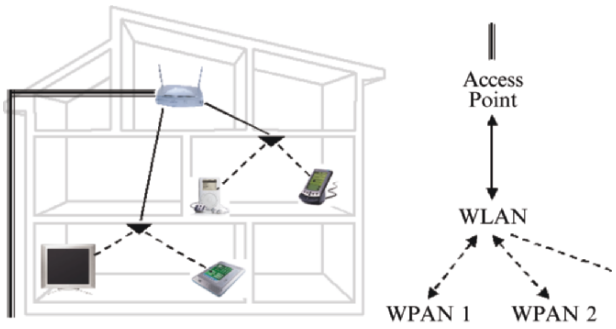


Figure 2.4-2. WLAN/WPAN infrastructure.

proposition for new homes. However, in existing buildings the cost and disruption of installing a wired infrastructure is likely to be too expensive, which favors a wireless solution. Consequently, wired infrastructures will not be considered further in this chapter.

2.3. Future Technologies

In terms of new technologies that are required to support the whole range of wireless data rates for AmI, two gaps in the portfolio can be identified. The first is a technology for providing multi-Gbit/s data links in an in-room WPAN system, and the second is a solution for ultra low power wireless links for various sensor applications.

2.3.1. Multi-Gbit/s wireless links

From the discussions earlier in the chapter, a target data rate of several (3–5) Gbit/s has been identified for applications, such as HDTV streaming and high-speed bulk data transfer. This is far beyond what current systems (and those under development) are capable of, so it is clear that a new solution is required. Efforts to address this fall into two categories. First, many systems are employing techniques to make more efficient use of the spectrum (i.e., increasing bits/s/Hz). These include the use of more complex modulation schemes, such as OFDM combined with a high order QAM. However, per today, there is a practical upper limit to the bandwidth efficiency of a digital wireless system, which is around 2.5 bit/s/Hz as illustrated in Figure 2.4–3.

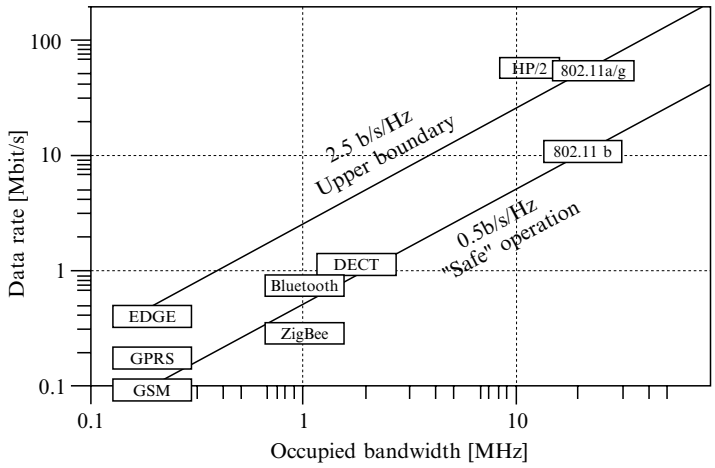


Figure 2.4-3. Spectral efficiencies.

The exceptions are MIMO systems, such as those in 802.11n, which exploit multiple channels in a rich multipath environment using space-time coding techniques. With such systems, spectral efficiencies as high as 10 bits/s/Hz are anticipated, but this comes at the cost of considerable added complexity.

This leads us to the second alternative, which is to use more bandwidth, c.f. Shannon's Law in Equation (2.4–1).

$$C = B\log_2(1 + S/N) \tag{2.4-1}$$

Exploiting more bandwidth has the attractive property that the channel capacity rises in direct proportion with the bandwidth employed. UWB does exactly this. However, with increasing levels of congestion at lower RF frequencies (especially < 6 GHz), extra bandwidth has to be found, and this is only structurally possible at higher RF frequencies. One consequence of using higher RF frequencies above 6 GHz is that attenuation by walls increases, which means that such frequencies cannot be used for inter-room communication. However, from the Ambient Intelligence perspective, this is an advantage, as we want to have room-sized WPAN cells, in order that spatial reuse of spectrum can be achieved, with the consequent reduction in coexistence and interference issues.

Table 2.4–4 shows a comparison between the various candidates for a multi-Gbit/s wireless WPAN system. For completeness UWB has been included in this table. In Europe and the USA, spectrum has been allocated at 17 GHz and 24 GHz for unlicensed applications. Both of these suffer from the lack of allocated bandwidth, with a maximum of 250 MHz being allocated. In order to reach a data rate of 3 Gbit/s, the system would

Table 2.4-4. Comparison of candidate systems for multi-Gbit/s links.

Standard	17 GHz ISM	24 GHz ISM	UWB (MBOA)	60 GHz
Regulatory	Europe	USA	USA (RoW in progress)	USA (5 GHz) Japan (7 GHz) Europe (3 GHz)
Bandwidth	200 MHz	250 MHz	1584 MHz (Group A)	3 GHz
Spectral efficiency for 3 Gbit/s link	15	12	2	1
Path loss (10 m) compared to 2.4 GHz	17 dB	20 dB	4 dB	28 dB
Antenna directionality	0–12 dB	0–12 dB	0 dB	15–20 dB

have to support spectral efficiencies of at least 12 bits/s/Hz. The key issue with UWB is the limited amount of power that can be transmitted (−41.3 dBm/MHz) because of concerns over possible interference issues. With this power level, current developments are typically targeting 480 Mbit/s, and even by exploiting the whole of the 7 GHz bandwidth, expectations are that the limit of UWB will be in the 1–2 Gbit/s range.

This, then, brings us to the spectrum close to 60 GHz that has some major advantages [5]. The first, and most significant, is the amount of allocated bandwidth. In the USA, the FCC has allocated 59–64 GHz for unlicensed applications; in Japan 7 GHz of bandwidth has been allocated between 59 and 66 GHz; and in Europe it is expected that 3 GHz will be allocated between 59 and 62 GHz for WPAN-like applications. This means that there could be 3 GHz of bandwidth *globally* available. Consequently, a 3 Gbit/s data stream can be achieved with a spectral efficiency of 1 bit/s/Hz—a figure that is easily achieved by today's wireless standards. Other advantageous features of the 60 GHz band include the high level of absorption by oxygen (10–15 dB/km) and water, and the limited ability to pass through walls—both of which are important for interference-free and secure systems. Another advantage of 60 GHz is the short wavelength—5 mm. This means that a $\lambda/4$ patch antenna measures only 1.25 mm across, and the prospect of using directional antenna arrays to increase the link budget becomes feasible. The small size of an antenna also means that a 60 GHz radio can ultimately be smaller than comparable lower RF frequency systems and consequently lend themselves much more to being hidden in the environment—one of the key aspects of AmI.

Turning to the technical challenges imposed by 60 GHz, a key issue is that of achieving low cost. In practice, this means that 60 GHz transceivers will have to be implemented in a low-cost Si or SiGe based technology. Systems are available today using GaAs MMICs, but these are, and are likely to remain, too expensive for consumer applications. In order to assess the suitability of technologies for given RF frequencies, two figures of merit are often used: f_T, the transition frequency; and f_{MAX} the maximum frequency of oscillation. As a general rule of thumb, f_{MAX} should be approximately $2f_T$, and f_T should be at least 2–3 times the RF frequency. This means that silicon IC processes with f_T and f_{MAX} above 120 GHz and 240 GHz respectively are candidate technologies for 60 GHz applications. Such processes are being developed as shown in Figure 2.4–4. Processes in the top-right segment of the figure are potentially suitable for 60 GHz systems.

Apart from the IC technology itself, other topics need to be addressed, for example, high-quality passive components, signal handling at 60 GHz and, of course, circuit and system level aspects. However, many groups are

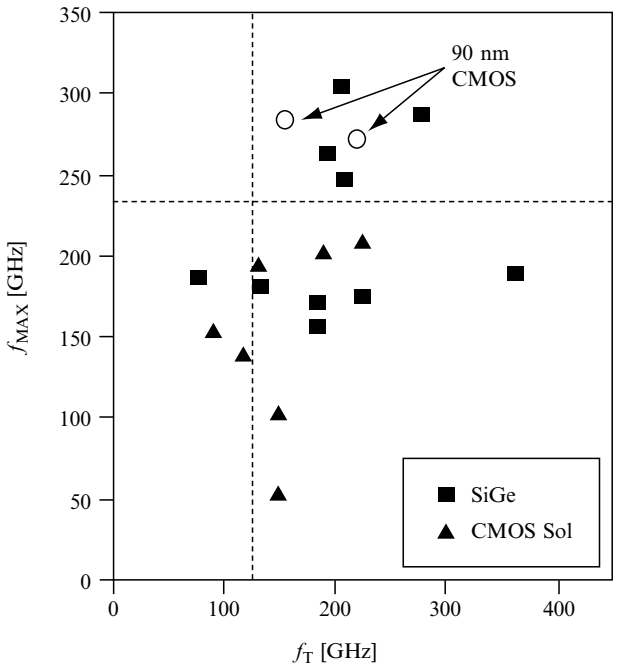


Figure 2.4-4. SiGe and CMOS IC technologies.

already active in this field, and it is to be expected that a low-cost 60 GHz WPAN system will become available in the coming years and therefore be available to form part of the wireless infrastructure for AmI.

2.3.2. Ultralow power radio

Apart from high-speed communication, the other key requirement is the need for low data rate links that will interconnect the vast number of sensors in the environment. The fact that AmI calls for such a large number of sensors (e.g., for temperature sensing, object identification, light level detection and so on) essentially defines the core attributes of these wireless devices: they must have an extremely small form factor, consume very small amounts of power and, of course, be low cost. This broad set of requirements suggests that conventional passive RFID tags could be suitable for many of the applications. This is indeed the case, but for other applications—especially those where the tag must *initiate* the conversion—a new approach based on ultra low power radio is required. An example of such an application is one where temperature sensing is involved. If this is implemented by a tag being periodically activated by a

reader, then the tag can only report that the temperature had *previously* exceeded a given threshold: with a radio device that can initiate the conversation, it can report a situation where the temperature is *about* to exceed the threshold, so appropriate action can be taken.

Table 2.4–5 compares some features of conventional RFID tags and existing low power radio systems, such as ZigBee.

Many attempts are being made to bridge the gap between the functionality of RFID tags and conventional radios. These include active RFID tags and NFC. Active RFID tags are aimed primarily at logistics applications where a larger distance between the tag and the reader needs to be supported together with the need for robust operation (e.g., reading the tags in hostile environments). A solution is to augment the passive RFID tag with a battery power source. As with conventional RFID systems, the tag is activated by energy transmitted from the reader, and the battery is then used to increase the return range of the tag. However, the problem still remains that the tag cannot initiate the conversion. Near field communication (NFC) [6] on the other hand, is a symmetrical system that does allow the link to be activated from both ends. However, this is based on 13.56 MHz magnetic coupling, where the magnitude of the field drops off at $1/d^3$, which means that the transmit power needs to be unfeasibly high (of the order of Watts) for 10 m transmit distances.

Consequently, for sensor networking, where the communication can be initiated from either end of a link, and where distances of up to 10 m need to be covered, a conventional radio solution (i.e., one that employs a transceiver at both ends of the link) is required. However, such a solution must still satisfy the requirements of small size, ultra low power, and low cost. Attributes, such as physical size and cost can be addressed by using appropriate (combinations of) technology for the integration of the transceiver. The power source is a critical component in the system, as it is likely to determine the final size and form factor of the radio device. New generation thin and flexible batteries, of the type produced by power paper and others are extremely attractive and could enable a *stick-on* radio (a radio as easy to install as sticking on a *Post-It* note). In addition,

Table 2.4-5. Comparison of low power radio and RFID technologies.

Standard	Bluetooth	ZigBee	RFID	NFC (active mode)
Frequency	2.4 GHz ISM	2.4 GHz ISM	13.56 MHz + various	13.56 MHz
Data rate	1 Mbit/s	250 kbit/s	e.g., 26 kbit/s	106, 212 kbit/s
Link initiation	Either end	Either end	Reader only	Either end
Power	100 mW	40 mW	$<$ 1 mW (tag)	$<$ 100 mW

the use of energy scavenging techniques holds the promise of the battery-less radio.

In order to reduce power consumption—or more accurately, energy per message—the radio system needs to be optimized at both the circuit and the system level. At the circuit level, a complete radio transceiver needs to be designed to consume of the order of 1 mW when active, and much is being done in this area.

At the system level, much has been done already. Systems, such as ZigBee have protocols that are optimized for applications where the sensor node needs only occasionally to transmit data back to the master—for most of its life the ZigBee radio is in a power down mode. However, unlike RFID, the protocol is designed so that the sensor node can initiate a conversion at anytime. The other system level consideration is the optimum data rate. From the energy-per-bit perspective, a high data rate is more energy efficient, so in principle a system such as UWB would be optimum. However, we also need to consider the energy consumed as the transceiver powers up from the sleep state, as this contributes to the overall energy per message.

Figure 2.4–5 shows the energy per bit and turn-on energy for various systems and also for a target ultralow-power solution. The message length is taken as being 288 bits, and the turn-on (from sleep to being ready to send a message) is 200 μs. If the ratio of data rate to power consumption is too low (*Slow*), the energy used to send the message dominates. On the

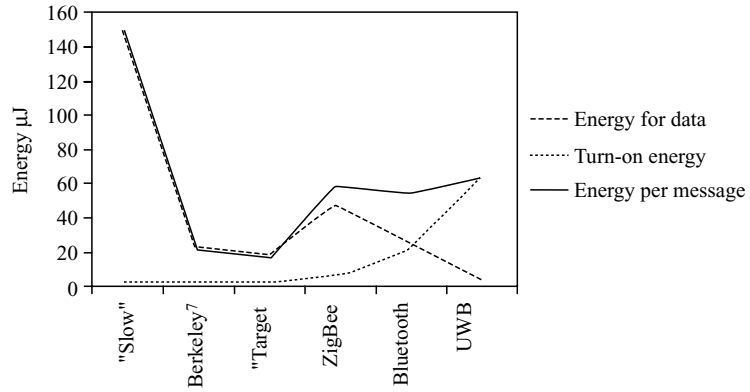


Figure 2.4-5. Energy per message.

right hand side, the turn-on energy dominates. From this simplified calculation, the optimum is a transceiver capable of data rates of around 50 kbit/s at an active power consumption of less than 3 mW (*Target*). Recently published results [7] show that this is feasible.

3. WIRELESS POSITIONING

So far, we have concentrated on the wireless transmission of data around the environment, be it audio/video content or information from sensors. However, AmI has concepts such as context-sensitivity at its core, and consequently will rely on the environment knowing the locations of objects and people. In order to do this, some form of indoor positioning system is required. The accuracy to which the position of an object needs to be determined is highly application dependent and can range from knowing whether two objects are in the same room, to accuracies of the order of 1 cm (e.g., for when an application needs to determine whether a particular object is on a desk or in the drawer just beneath).

The most well-known indoor positioning systems are based on the measurement of the time-of-flight (ToF) of ultrasound pulses from one location to another. Typically, such a system comprises a number of beacons at known fixed points in the environment, and by measuring the distance from the object to be located to each beacon, a three-dimensional position can be determined. In order to provide timing synchronization between the beacons and the object, an RF signal is transmitted at the same time as the ultrasound pulse, and the ToF of the ultrasound pulse is simply determined by subtracting the time-of-arrival (ToA) of the ultrasound pulse from the ToA of the RF signal. The RF signals are also used to identify the beacons.

Ultrasound systems in general work well, and can achieve accuracies of better than 5 cm. However, given that ultrasound does not go through walls, each room will then need to have its own installation of beacons. This helps to provide easy identification of the room in which an object is located, but inevitably complicates the overall positioning infrastructure.

An attractive alternative is to use the wireless communication infrastructure for the indoor positioning function. The principles of operation are essentially similar to those of the ultrasound system, but the implementation is more challenging because RF signals travel approximately 10^6 times faster than ultrasound signals, so the accuracy of the ToF measurement has to be a factor 10^6 higher if the same positional accuracies are to be achieved.

Figure 2.4-6 shows the different techniques that can be employed in complete indoor positioning system architecture. These will be discussed in the context of an RF implementation, but the structures could equally be employed in an ultrasound system.

The first of these is the use of *simple beacons*. If the object to be located can *hear* the beacon, it must then be within a certain distance of the beacon. Such an approach can be implemented using, for example, Bluetooth, and is useful for determining the logical position of the object. The drawback is that the required number of beacons is a function of both the area to be covered and the positional accuracy required. The result is that a large number of beacons are needed. *Signal strength* measurements are an extension to the straightforward beacon concept and, in principle, can be used to improve the positional accuracy of the beacon approach. However, in practice the improvement in accuracy is severely limited as the received signal strength will be strongly affected by reflections and other objects in the vicinity. Accuracy can be further improved by the use of *fingerprinting*. With the fingerprinting technique, an array of beacons is installed around the area to be covered, and the signal strength at every point in that area is measured and stored in a look-up table. In this way, the variations in signal strength due to other objects etc., can be calibrated out and accuracies of the order of 1 m or less can be achieved (again, dependent on the presence of mobile objects such as people). The serious drawback, however, is the high cost of installation. Additionally, in order to maintain the accuracy of the system, the fingerprinting exercise will

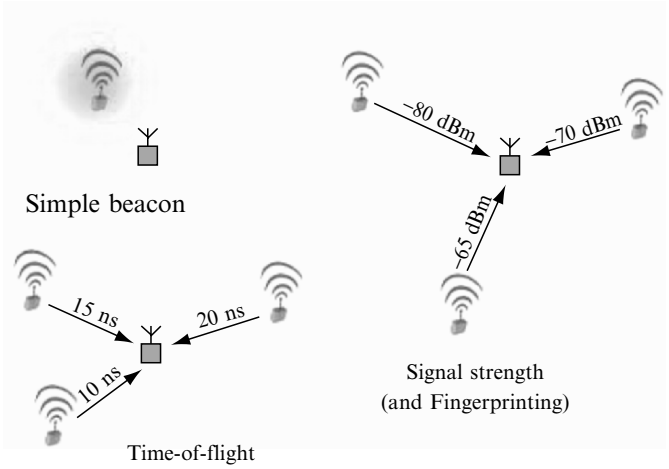


Figure 2.4-6. Indoor positioning techniques.

have to be redone whenever there are changes in the environment, such as the addition of new furniture and other large items.

The last approach is based on the ToF principle as used in ultrasound positioning systems. Using this technique with conventional standards such as 802.11b, accuracies of around 1 m should be possible, and if the high bandwidths available in systems such as UWB can be exploited, accuracies of a few centimeters will be achievable.

Figure 2.4-7 shows how these different positioning techniques compare as a function of the required accuracy and the coverage area. Simple beacons lie on the diagonal line on the right hand side of the figure, with RF tags providing this functionality at the 1 m level and Bluetooth offering 10 m accuracy. Below this line are systems whose performance is worse than that of simple beacons. Above the line in the top right-hand part of the diagram is where accuracies are not sufficient to identify a particular room. So, if the aim is to cover the whole house using a limited number of beacons, the resulting system should not be operating in this part of the diagram. To the left is the region in which it might be possible to identify a given room, but the accuracy of a few meters is not good enough to allow accurate associations to be made (for example, to verify the object X is next to object Y). In the bottom left part of the figure is the area in which in-room systems should operate (i.e., a coverage range of 10 m), but with accuracies of significantly better than 1 m. Superimposed on these various regions are the currently available RF-based indoor positioning systems.

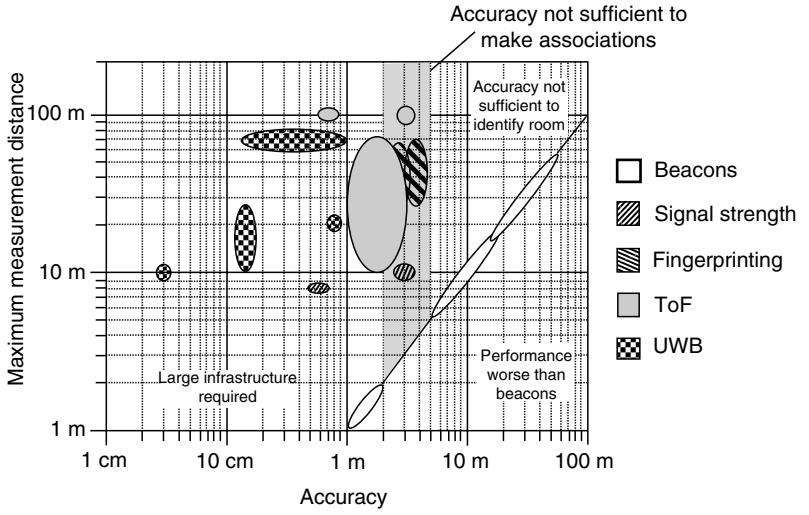


Figure 2.4-7. Comparison of wireless positioning systems.

Going back to the requirements from the AmI perspective, it is clear that the overall positioning infrastructure will have to be capable of determining positions of people and objects to an accuracy of better than 1 m, while at the same time having a coverage area equal to the whole house. Figure 2.4-7 clearly indicates that a homogenous RF-based positioning system is not going to be able to provide this functionality without a considerable additional infrastructure. Some hierarchy is therefore needed, and the obvious solution is to exploit the hierarchical structure of the wireless communication system itself. In other words, within each room the WPAN system is used to provide the centimeter-level accuracy positioning capability (exploiting the high RF bandwidth of the WPAN). This can be used for applications in which associative information is needed. On the wider scale (e.g., for locating an object in the home), the (room-bound) WPAN can locate the exact whereabouts in the given room and then communicate this to the backbone WLAN system. This, in turn, will know which WPAN has located the object, and therefore the room in which the object is located. In this way, the position is determined and reported in a natural way (i.e., *in the dining room*, *next to the PC*).

4. SYSTEM INSTALLATION

Having addressed the topics of wireless data communication and indoor positioning based on RF signals, the next subject is somewhat orthogonal—the set-up and configuration of the underlying wireless infrastructure in a manner that is easy and understandable.

The issue is one of complexity, as is evident from the foregoing sections. We are talking about a hierarchical, multistandard wireless network, with the possible added complication of beacons for the indoor positioning system. Comparing this with the simple wireless networks that are present in some homes today, the scale of the installation problem comes into sharp focus. It is not simply a question of providing easier to understand instruction manuals (of course, this is an essential part of the process) because, apart from other considerations, the number of possibilities and variations in a specific system are so vast, the manual would likely become unmanageable.

At least part of the solution is to use technology within the network to guide and assist the installation procedure. Take, for example, the videocassette recorder (VCR). From its inception, the process of installing the VCR was lengthy, complicated, and often misunderstood by the user. Typically, the process involved connecting a number of cables between the VCR and TV; carrying out various tuning operations to find the TV

channels, and then making sure that they are in the same order as on the TV. Then, of course, the clock had to be set—easy the first time, when the manual is in front of you, but a common source of frustration when the clocks move forwards or backwards twice per year, and the manual can no longer be found. In recent years, the process has improved considerably with the introduction of concepts, such as Philips' *Easy Link*. The complete set-up procedure now consists of connecting the cables and pressing a single button—the VCR and TV together do the rest. Of course, such simplicity does not appear overnight, the design of such a simple system is a huge challenge in itself and, importantly, often requires elements that need to be standardized.

Returning to the wireless system for AmI, many similar set-up examples can be identified. Lighting is a key element, and in many cases individual lamps will need to be controlled via a wireless link. On the face of it—a relatively straightforward problem to solve. However, on further inspection it becomes more complex. The lamp will need some form of ID number (preassigned during manufacture or assigned during installation), which needs to be programmed into the light control system together with the information about the location of the lamp, and so on. In essence, enough information has to be put into the system so that the user, when seated in a chair, can metaphorically 'point' to a given lamp to turn it on or off. Given the number of lamps in an average home, the set up of even this part of the environment is already looking like a daunting prospect.

The challenge is clear. If the set-up of an AmI environment is to be as simple and intuitive as its use, then we need to start considering the set-up process at the earliest stage in the design of the underlying wireless infrastructure. But why is this so important for *wireless*? Of course, the issue of set-up extends to all systems in the environment, but in the case of the wireless infrastructure there is one important complication compared to the VCR case given earlier: if a link doesn't work you can't look around the back to see if the cable has been plugged in!

5. CONCLUSIONS

In this chapter, we have discussed the attributes and requirements of a wireless infrastructure that is capable of supporting the features and applications that will exist in an AmI environment. The two key requirements that result from the overall concept are wireless data transfer around the home, and the ability to locate both people and objects in the environment. For the first of these, the requirements are diverse, from multi-Gbit/s communication for bulk data transfer to ultralow power

wireless links that will connect the huge number of sensors that will be hidden in the environment. Add to this the need for data to be transferred from room to room, and the complexity of the required infrastructure becomes clear. One solution is to adopt a structure comprising in-room WPAN communication, coupled to a WLAN backbone, together with local (room-based) storage to buffer differences in data rates. For the indoor positioning function, the main issue is accuracy (many applications will require location accuracies of the order of a centimeter) combined with coverage of the entire house. Inevitably, this will require a considerable amount of infrastructure. Although indoor location systems based on ultrasound are available today, a much more attractive possibility is the use of the wireless communication infrastructure for the indoor positioning system. Achieving the required level of performance is most certainly a challenge, but the benefits of reusing the infrastructure are clear.

Finally, there is the question of how to install and set-up such a complicated wireless infrastructure. This is likely to be the first exposure to AmI that most users have, and consequently the experience must be as simple and intuitive as the interaction with environment itself.

ACKNOWLEDGEMENTS

As usual, many people have contributed to the content of this chapter. In particular I would like to acknowledge the specific contributions of: Alan Davie, Jean-Paul Linnartz, Adrian Payne, and Martin Wilcox. In addition, discussions with members of the Integrated Transceivers Group at Philips Research Eindhoven have brought a continuous stream of new ideas.

REFERENCES

[1] Aarts, E. and Marzano, S., *The New Everyday*, 010 Publishers, ISBN 90-6450-502-0.
[2] http://www.zigbee.org.
[3] http://www.bluetooth.org.
[4] http://grouper.ieee.org/groups/802/.
[5] Bird, N. and Sanduleanu, M., 2004, Towards integrated transceivers in mm-wave applications, *GAAS 2004*, European Microwave Week, October 2004.
[6] http://www.nfc-forum.org.
[7] Molnar, A., Lu, B., Lan Zisera, S., Cook, B. W., Pister, K. S. J., An ultra-low power 900 MHz RF transceiver for wireless sensor networks, Custom integrated circuits conference, 2004. Proceedings of the IEEE 2004 3–6 Oct. 2004, pages 401–404.

Section 3

Smart Sensors

Chapter 3.1

INTERCONNECT AND PACKAGING TECHNOLOGIES FOR REALIZING MINIATURIZED SMART DEVICES

Eric Beyne
IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
eric.beyne@imec.be

Abstract In order to become truly ubiquitous, the systems envisioned to create the ambient intelligence (AmI) landscape must be highly miniaturized and manufacturable at low cost. Together with *system-on-a-chip* (SoC) technology, interconnect and packaging technologies are key enabling technologies for realizing smart *system-in-a-package* (SiP) solutions. The realization of such devices requires new developments in the field of packaging and interconnection technologies.

The continuing scaling trend in microelectronic circuit technology has a significant impact on the different integrated circuit (IC) interconnection and packaging technologies. These latter technologies have not kept pace with the IC scaling trends, resulting in a so-called *interconnect technology gap*. Multilayer thin film technology is proposed as a *bridge*-technology between the very high density IC technology and the coarse standard PCB technology. It is also a key enabling technology for the realization of true SiP solutions, combining multiple SoC IC's with other components and also integrating passive components in its layers.

A further step is to use this technology to realize new functionalities on top of active wafers. These additional *above-IC* processed layers may be used for low loss, high speed on-chip interconnects, clock distribution circuits, efficient power/ground distribution, and to realize high Q inductors on chip.

1. INTRODUCTION

According to the international technology roadmap for semiconductors (ITRS) [1], the scaling of the smallest feature size on chips is continued unrelentingly. Figure 3.1-1 illustrates this scaling trend with respect to the

S. Mukherjee et al. (eds.), AmIware, 107–123.

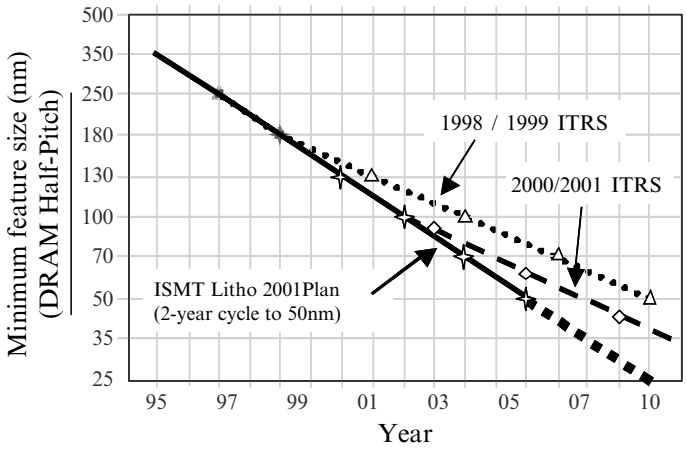


Figure 3.1-1. Scaling trend in time of the minimum lithography dimensions, after ITRS [1].

on-chip lithography trends. The observed trend confirms the so-called Moore's scaling law, which predicts that the number of transistors on a chip would double every 12–18 months. With each new generation of smaller devices at the wafer level, faster operating speeds can be achieved. This has a significant impact on how packaging and interconnection technologies are conceived and realized. Sizes have to be minimized, and dimensions have to be chosen specifically to match required electrical characteristics.

Another consequence of Moore's law from the IC-packaging and interconnection engineers' perspective is that, for an unchanged circuit design, the IC will become smaller and the input and output (I/O) contact pads have to be fitted on a smaller die. This results in an increased I/O-*density*. In the field of circuit design, the so-called Shannon's law, states that the circuit architecture complexity (number of gates or memory bits on a chip) grows even faster in time than Moore's scaling law. Another empirical observation, the so-called *Rent's rule*, states that the number of circuit I/Os increases exponentially with increasing circuit complexity. In combination, all these trends result in a significant increase in the I/O pad density on a chip. This by itself has a big impact on the interconnect and packaging technologies for the future, as these technologies can be seen as *geometry translators*—from the submicron?scale on the chip to the millimeter scale at system level.

The newest IC-technologies allow for incredibly complex circuit integration on a single die. This is often referred to as system-on-chip (SoC), architectures. This approach is highly successful in integrating the major

part of a system, in particular digital signal processing, on a single deep submicron CMOS technology. However, in most cases the SoC technology is not able to integrate the entire system on a single piece of silicon. The SoC chip still needs packaging and interconnection to other system elements. A system (may) also consist(s) of many non-silicon parts, such as passive components, displays, sensors, antennas, connectors, etc. Furthermore, within the IC-technology there is a growing divergence between different types of technologies. Besides the *mainstream* high-density logic CMOS technology, there are many different specific technologies for memory, analog, high voltage, RF, MEMS, and electro-optical circuits. It is highly unlikely that all these functions can be integrated in a cost effective fashion in a single SoC Si-technology platform.

Although SoC cannot really offer a true single chip system, it can significantly reduce the size as well as the cost of a system or subsystem. In combination with high-density interconnection and packaging techniques it enables the realization of a so-called system-in-a-package (SIP). Multilayer thin-film technology is a key enabling technology for the realization of a SIP.

2. THE INTERCONNECTION GAP

The main off-chip interconnection technology used in industry today is the so-called printed circuit board (PCB) or printed wiring board (PWB) technology. This technology is based on the lamination of a stack of multiple interconnect circuit layers. Each layer typically consists of a double-sided metallized sheet of epoxy resin-impregnated glass-fiber cloth. These *double* sided *inner* layers are individually patterned using wet etching of the copper metallization. Both sides of the layer are connected to each other by mechanically drilled and chemically metallized holes (buried via's). Several of these inner layers, interlaced with additional isolating layers of resin impregnated glass cloths, are finally laminated together in a hot isostatic press to form the final multilayer PCB boards. The different layers in the stack are interconnected through mechanically drilled holes that are chemically metallized (through via's). Wet etching is typically used to pattern the top metal layer.

The technology described above has been the major interconnect technology for the past 30–40 years. Evolution towards smaller feature sizes has been rather slow compared to the scaling frenzy in IC-technology. The smallest dimensions in this technology are now in the order of 100 μm line width and spacing, drilled via holes of 200 – 300 μm diameter with via-contact lands of 300 – 400 μm diameter.

In the past ten years, a major advance in this technology has been brought about by the use of so-called *sequential build-up* (SBU) technology. In this case additional resin layers and metallic layers are applied to the PCB board, after the *stack-up* lamination process, turning the process from a parallel to a more sequential process flow. For this technology, via holes are mainly realized using laser drilling, although photosensitive dielectrics and plasma etching of holes are used sometimes.

An overview of the currently most advance SBU technologies is shown in Figure 3.1-2 [2]. For commercial products, the smallest line-widths available today are 75 μm. Via diameters are typically larger than 50 μm, their further shrinkage is limited by the materials used, the relatively thick SBU dielectric and the metallization techniques used, which all limit the aspect ratio of the via's in SBU to about 1. The most significant limitation is however the size of the via metallization pad. This size is minimally 100 μm larger than the via hole size. This difference is caused by the large tolerances on PCB substrates, caused by their relatively poor dimensional stability during processing.

From Figure 3.1-2, it is clear that even when using the most advanced PCB SBU interconnection technologies under development, a large *interconnect gap* still exist with the on-chip interconnect technologies. Multilayer thin-film lithography based interconnection technology is required to bridge this *gap*. Line widths and spaces can scale down to 5 μm

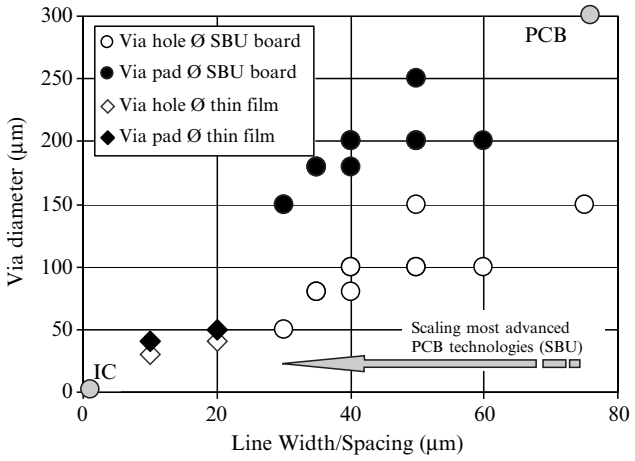


Figure 3.1-2. Illustration of the *interconnect gap* between the on-chip interconnect density and the most advanced PCB technologies, showing the scaling of minimal feature sizes of the most advanced SBU PCB technologies under development [2]. Multilayer thin-film technology is a *bridge* technology that may close this gap.

dimensions, and via hole diameters could be smaller than 40 μm or even 20 μm without major problems. A via metallization pad, 10 μm larger than the via hole diameter is sufficient to account for any alignment and process tolerances. This makes thin-film lithography based interconnections an ideal bridge technology between the IC and the PCB dimensions.

The thin-film lithography based technologies are not intended for the realization of large area interconnection boards, such as PCB's. This would result in low yield and high cost substrates. It is, however, the ideal technology for realizing SIP devices, not larger in size than typically 3×3 cm [3].

3. MULTILAYER THIN-FILM TECHNOLOGY

The key features of a multilayer thin-film technology are the use of wafer-like process steps and the use of thin-film planar 1-X lithography (Mask feature size is equal to the printed feature size, in contrast to reduction steppers where the features on the mask are reduced to smaller dimensions on the wafers). The basic elements of such an interconnect technology are a thin-film, high density metallization technology and a thin-film, dielectric deposition technique, capable of realizing very small via holes in the isolation layers to allow for high density interconnects between the different layers in the structure. An example of a cross-section through such a multilayer build-up is shown in Figure 3.1-3.

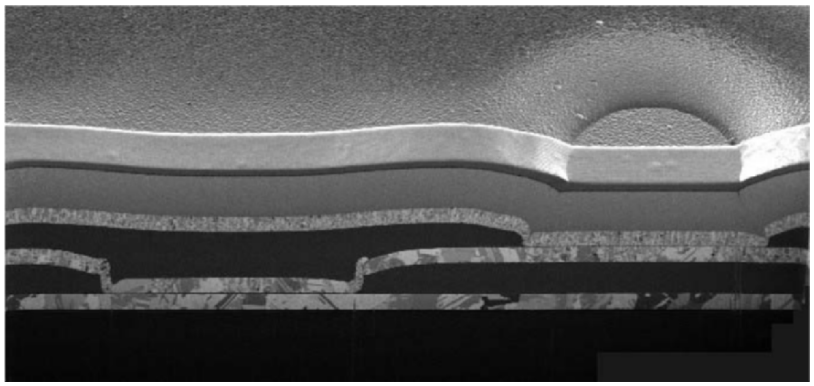


Figure 3.1-3. Cross-section through a multilayer thin film structure, showing 2 μm thick copper conductor layers and 5 μm thick BCB dielectric layers (IMEC).

3.1. Metal Interconnect Lines

Different types of metals and different deposition and patterning techniques can be used. The main interconnect materials used today are Al and Cu. Deposition of the materials is typically performed using physical vapor deposition (PVD) or, in the case of Cu, by electroplating. The PVD layers are limited to thin layers of $1 - 2\,\mu m$—electroplating is more effective for thicker layers. Patterning is generally performed by wet chemical etching or, in the case of Cu, by semi-additive plating techniques. Wet etching has a limited resolution due to the unavoidable under-etching. Semi-additive plating has an excellent conductor width control through the use of a thick photo resist *mould*. The Al metallizations can also be patterned with great accuracy by using dry etching techniques. Another metallization technique is the use of evaporation and lift-off technology. This technology is typically used for materials that are difficult to etch on the wafer. It is also restricted to rather thin layers. For realizing low resistive, high density interconnect lines, the semi-additive Cu plating technique is the preferred process. It can also be applied to other metals, such as Ni, Au, Ag, and Co [4].

3.2. Dielectric Layers

In order to obtain a proper electrical performance of the interconnection layers, the isolation layers between the different metal interconnect layers have to be relatively thick layers, typically $5 - 10\,\mu m$. For good high frequency performance they also need to have a low k value. Typical dielectrics used are polymer spin-on dielectrics, such as polyimides and benzo-cyclobutene (BCB) [5]. These materials are preferably photosensitive, allowing for a high yield, fast, and high-density patterning process. Most materials available today are negative type photosensitive materials. Via openings down to $20\,\mu m$ can be typically realized without specific problems.

3.3. Integration of Passive Components

While the on-chip integration increases, the relative area occupied by chips on boards (at the system level) decreases rapidly. The bottleneck for further size reduction is often determined by the size occupied on the board by passive components, such as resistors, capacitors, and inductors. As a result, the integration of passive components in the interconnect and package technologies is receiving strongly increasing attention. Thin-film technology is well suited for the integration and miniaturization of passive components. Complex materials can be deposited with high repeatability to form the highest quality resistor or capacitor layers. The thin-film

lithography assures a high dimensional accuracy, enabling small tolerances and increased miniaturization and, therefore, avoiding the need for *trimming* of resistor or capacitor values. The electroplated copper lines, described above, are ideally suited for realizing high quality inductors, particularly those required for high frequency applications.

4. CONNECTING HIGH DENSITY IC'S USING THIN-FILM TECHNOLOGY

The problem of interconnecting a high I/O density IC to a standard PCB circuit board can be solved using an intermediate *interposer* substrate, schematically shown in Figure 3.1-4. The interposer is a high-density interconnect substrate that can *translate* the on-chip fine geometries to the PCB-level coarse geometries. It also acts as a mechanical interposer, *absorbing* mechanical stresses that could occur between the die and the outside world.

4.1. Wire Bonding

The chip is typically connected to the interposer substrate by wire-bonding or flip-chip technology. As the contact pads are traditionally located around the perimeter of the die, the increasing I/O pad density on chips results in a need for a reduced I/O pad pitch. The most widely used interconnect technology used today is thermo-sonic Au wire bonding. Where a few years ago 100 μm pitch was seen as a lower limit for this technology, currently 60 μm pitch is in production and the estimates are that this technology will move down to 40 μm pitch by 2005. When using traditional PCB substrates, such small pitches cannot be maintained on the substrate side. This inevitably results in long interconnect wires, fanning out from the small on-chip I/O pitch to a coarse package-level pitch of typically 200 μm or more. Multiple rows of pads have to be used on the package level in order to avoid too large packages.

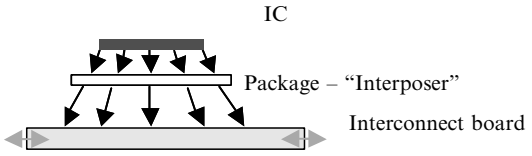


Figure 3.1-4. The package function as an "interposer" between the fine pitch chip I/O connections and the coarse pitch PCB-level contact pads.

4.2. Flip Chip Technology

One way to overcome the wire-bond problems is the use of flip chip bumping technology. When direct bumping of the IC to the interposer is foreseen, a flip chip bump pitch on the substrate, down to 40 μm, has to be envisaged. This cannot be achieved by SBU PCB technologies, but is well within the capabilities of thin-film lithography. As an example, in Figure 3.1-5 a thin-film on laminate technology is shown [6–8].

The thin-film technology can also be applied on the silicon wafer for rearranging the bond pads in a more sparse area array format. This redistribution process is often referred to as a *wafer level packaging* (WLP) technology. The cross section of such a thin-film build-up is shown in Figure 3.1-6. An application example is shown in Figure 3.1-7. The typical redistributed flip chip bump pitches will be 100–250 μm.

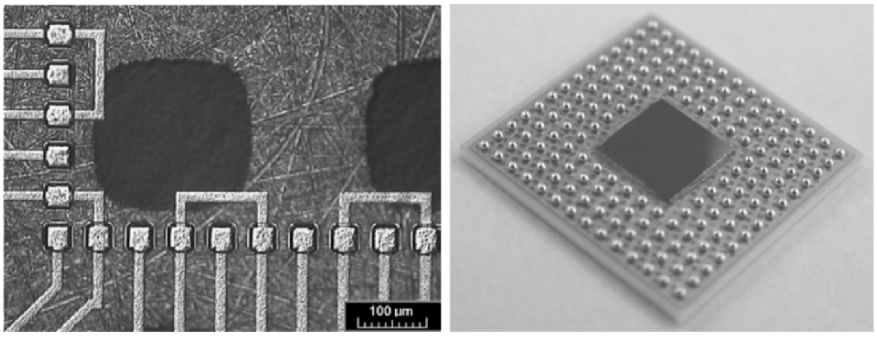


Figure 3.1-5. Example of a *thin-film on laminate* interposer substrate. Left: substrate detail showing 60 μm flip chip pads; Right: flip chip mounted die on laminate substrate (chip size = 5 × 5 mm) [8].

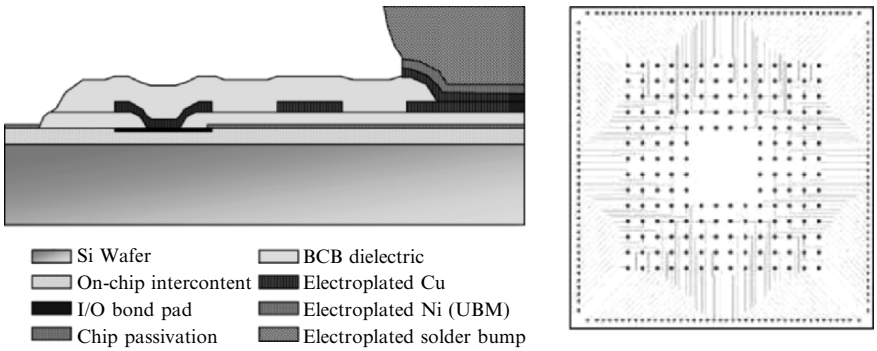


Figure 3.1-6. Thin-film interconnect pattern on chip, used for the *redistribution* of the peripheral I/O contact pads into an area array configuration. Left: schematic cross section; Right: typical redistribution layout.

For a die with a low I/O density, such as memory die, this can result in a significantly larger pitch of 500–800 μm. In that case, preformed solder balls can be applied directly on the wafer, resulting in a truly *chip sized package* (CSP) that can be mounted directly on standard PCB boards.

4.3. Multi-Chip Modules

When using a thin-film interposer substrate, multiple die can be attached to the interposer substrate, using the high-density interconnection capabilities of the multilayer thin-film technology. An example of such a high density, high speed interconnect technology, developed at IMEC, is shown in Figures 3.1-8 and 3.1-9.

This technology is also referred to as MCM-D technology (D from deposited dielectric). Very high interconnect densities can be achieved using this approach. The typical build up consists of 5 metal and 4 dielectric layers. Power and ground planes are used to provide a good reference to the signal interconnect lines on the X-interconnect and

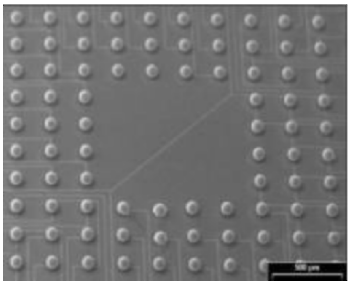


Figure 3.1-7. Photograph of a "redistributed" array of flip chip bumps on a Si-chip.

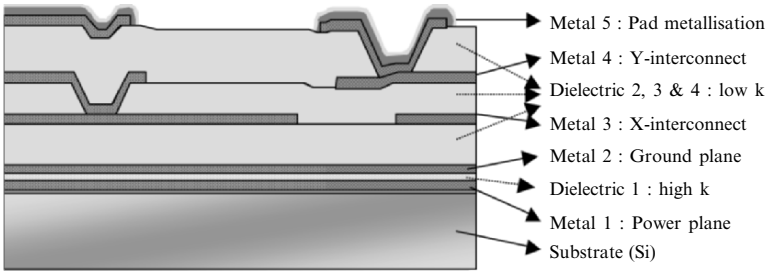


Figure 3.1-8. Schematic cross-section of a typical MCM-D interconnect module [9–11].

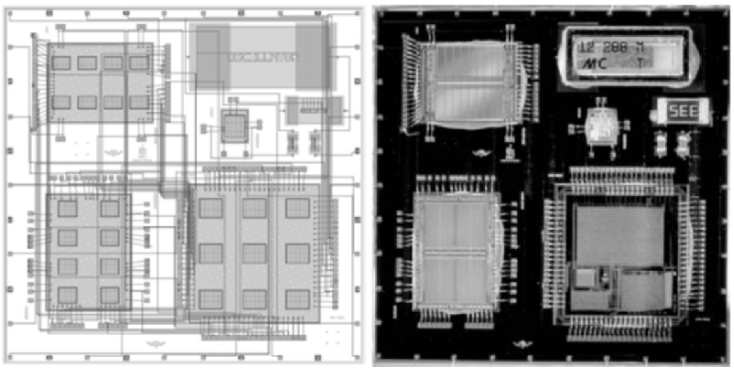


Figure 3.1-9. Example of an MCM-D module, measuring 2.5×2.5 cm. In this case, the chips are connected to the substrate using wire bonding. Left: interconnect layout substrate; Right: finished module [11].

Y-interconnect planes. The typical width and spacing of these interconnect lines is 10–20 μm. These layers are routed perpendicularly to each other, in order to minimize the electric coupling between the lines. Between the power and ground planes, a thin, high k dielectric is used to integrate a large value decoupling capacitor (up to $1\,\mathrm{nF/mm^2}$). The other dielectric layers use low k dielectrics with a typical thickness of 5–10 μm per layer. [3, 9–11]

One disadvantage of this technique is the requirement for additional packaging of the thin-film substrate. One approach that is under development at IMEC is the realization of the multilayer thin-film technology on top of a laminate interconnect substrate, similar to the thin-film on laminate interposer technology shown in Figure 3.1-5. [6]

4.4. Integrated Passives

Electronic systems also contain many non IC-components. Most prominent are the many resistor, capacitor, and inductor passive components seen in most systems. Also these components do not shrink as fast as the IC-technology, resulting in their relative increased use of substrate area. In particular for high frequency circuits, integration of passive components can lead to a significant size reduction as well as to an improvement in performance. An example of an RF-MCM-D technology with integrated passives, developed at IMEC, is shown in Figure 3.1-10. Some application examples are shown in Figure 3.1-11 [12–16].

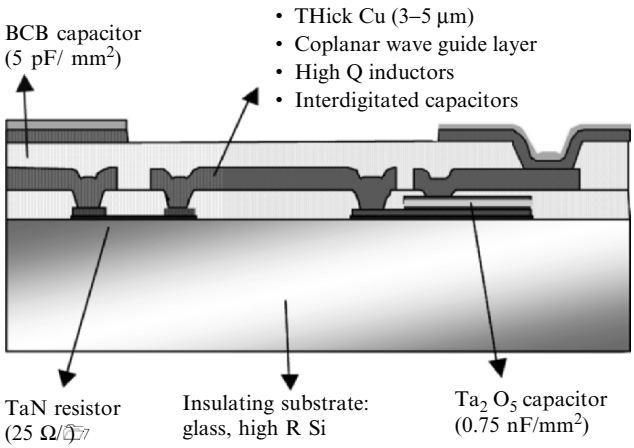


Figure 3.1-10. Schematic cross-section of an RF-MCM-D interconnect module with integrated passives as developed by IMEC [16].

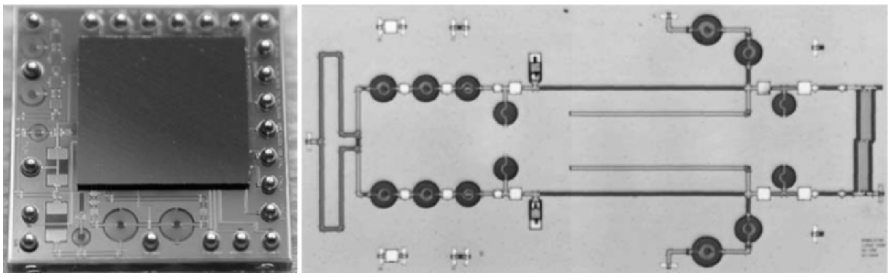


Figure 3.1-11. Examples of RF-MCM-D modules. Left: RF-section of a wireless front-end circuit showing a flip chip RF-chip on a glass RF-MCM-D substrate with 300 μm solder balls (CSP) for mounting on a PCB board (size = 7 × 7 mm); Right: sub harmonic QPSK modulator at 7/14 GHz (size = 8 × 17 mm) [16].

5. *ABOVE-IC* PROCESSING

The flip-chip technology discussed, in Section 4.2, is performed on top of active silicon waters. The flip-chip redistribution layers can also be considered an extension of the on-chip wiring layers. This *above-IC* technology allows for the realization of lines that are only 5 μm wide, 5 μm spaced, and 5 μm thick. Though small for traditional package and interconnect technologies, these lines are very *fat* compared to the back-end of line (BEOL) interconnect layers and therefore less resistive. The MCM and integrated passives technologies, described previously, can be applied

above-IC, rather than on an intermediate substrate. This opens potential applications of these layers for realizing high-Q on-chip inductors, low resistance power and ground supply lines, and fast on-chip interconnect lines.

5.1. Low Resistance *above-IC* Cu Wiring

The thick copper conductors, realized by electroplating, may be exploited in analog and power semiconductor technologies to lower the electrical resistance of the connections from high current drivers to the external package pins. Figure 3.1-12 shows some plated Cu-lines on top of the IC passivation with 10 μm thickness and 10 μm line and space width.

5.2. Fast on-Chip Electrical Wiring

The resistance of the electrical wires on the redistribution layer can be a factor 20 lower than the resistance of lines in the BEOL layers of the die. At the same time, the capacitance per unit length of these lines will not be significantly different, as they are basically *scaled-up* versions of the BEOL interconnect layers and also use low-*k* dielectric materials. The RC-delay of these *above-IC* lines will thus be more than an order of magnitude smaller than the on-chip interconnects.

Obviously, to be practical, two interconnect planes with perpendicularly routed interconnect lines will be required. As the wiring pitch on these layers will only go down to some 10 μm, the available wiring density will be relatively limited. This should however not be considered as a severe limitation for their applicability as they are only intended to be

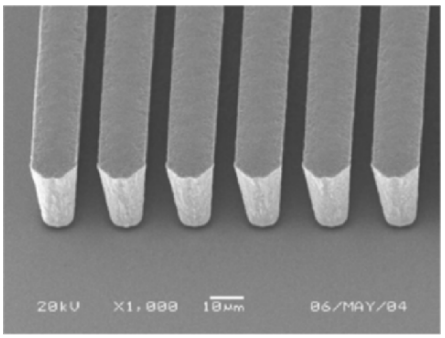


Figure 3.1-12. "Above-IC" electroplated Cu lines with 10 μm thickness, 10 μm nominal with spacing.

used for the longest global nets on the chip, such as interconnect lines between functional blocks on the die or lines distributing the clock circuit with high quality across the die.

The *above-IC* routing technology can be considered a *down-scaled* version of thin film multi-chip (MCM-D) technologies as they were proposed in the nineties and shown in Figures 3.1-8 and 3.1-9. Many of the techniques developed for these technologies can be used for on-chip electrical connections. One example is shown in Figure 3.1-13. In this approach, first a parallel power and ground plane are realized on top of the chip passivation. These planes form an integrated decoupling capacitor (up to $5\,\mathrm{nF/mm^2}$), a low resistive and inductive on-chip power and ground supply and provide a well-defined reference for the interconnect lines that are realized in two subsequent routing. This approach requires eight mask steps.

The interconnect lines in such a structure behave as lossy transmission lines [9, 18]. These lines have less dispersion compared to RC lines, therefore enabling a higher speed performance as signal rise times are better preserved. As lines on chip are still relatively short, line termination is not required to avoid signal ringing up to high signal speeds. If the line impedance is matched to the output impedance of the driving circuit, the signal delay can be minimized to the signal delay time of the rlc line.

The simulated attenuation and transmission delay of such lines is given in Figure 3.1-14. Figure 3.1-15 gives the time domain response of such lines. The 5 μm thick, 5 μm wide line corresponds to the *above-IC* line, the 1 μm thick, 1 μm wide line corresponds to a global on-chip interconnect line. A comparison between modeled and measured attenuation of such lines is given in Figure 3.1-16, showing excellent agreement [18]. The *above-IC* lines can be treated as transmission lines above a few 100 MHz.

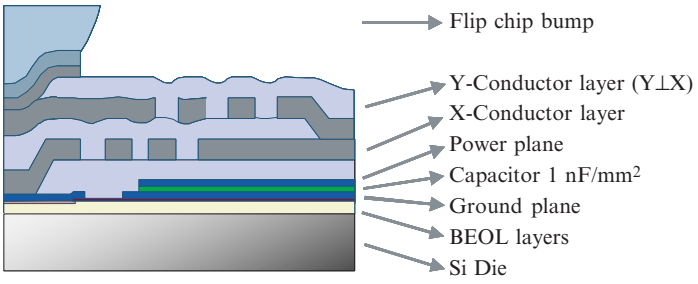


Figure 3.1-13. Schematic cross-section *above-IC* processed interconnect layers.

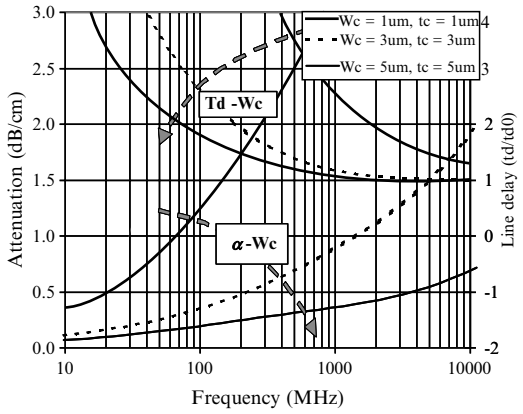


Figure 3.1-14. Attenuation (left axis) and normalized delay (right axis) of copper microstrip interconnect lines with cross sections from 1×1 to $5 \times 5\,\mu m^2$.

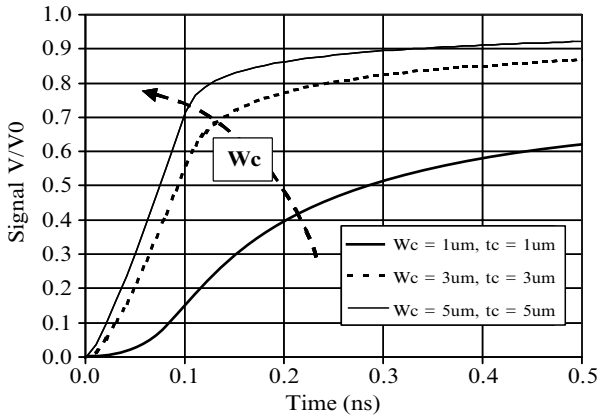


Figure 3.1-15. Time domain step-response of the copper microstrip interconnect lines shown in Figure 3.1-15.

5.3. High-Q on-Chip Inductors

For many high frequency RF-IC's, the poor quality factors of regular on-chip inductors is a limiting factor. This is due to the relatively high sheet resistance of the on-chip metallization and the losses in the semiconducting silicon substrate. By placing the spiral inductor in the redistribution layer, the distance between the spiral and the lossy substrate is greatly increased. By using a thicker, electroplated Cu conductor, a much lower track resistance is obtained. A FIB cross-section of such an inductor

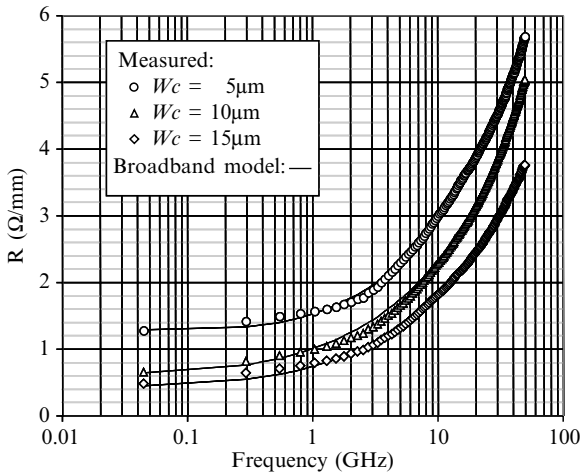


Figure 3.1-16. Measured and modeled frequency-dependent resistance of above-IC Cu interconnect lines (line thickness 3 μm) [18].

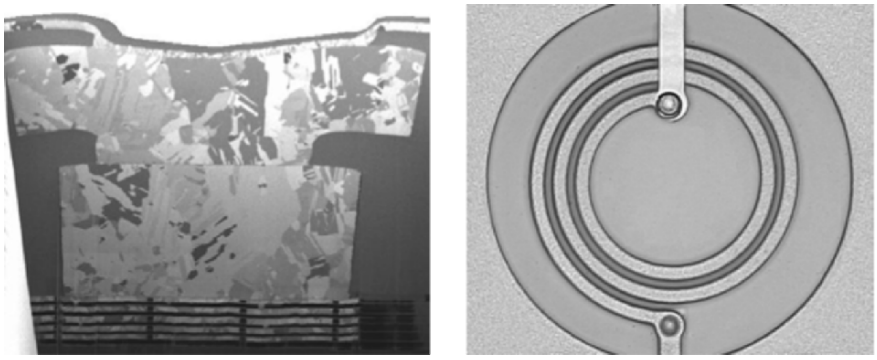


Figure 3.1-17. High-Q, 10 μm thick Cu inductor processed on top of a CMOS wafer. Left: cross-section contact inductor; Right: top view.

process is shown in Figure 3.1-17. In this case a 10 μm thick copper layer and a 18 μm thick dielectric is used. Inductors with Q-factors above 30 up to 5 GHz were obtained over 20 Ωcm Si CMOS wafers. [17] For differential inductors, even higher quality factors are obtained, as shown in Figure 3.1-18.

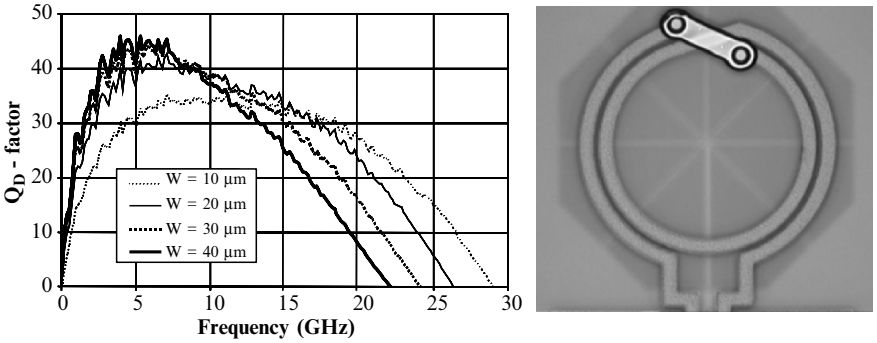


Figure 3.1-18. Photograph and quality factor of differential 1.6 nH *above-IC* inductors. Inductor parameters: 2 turns, line thickness of about 10 μm, line spacing of 10 μm, line widths ranging from 10 to 40 μm, and inner diameter of 250 μm.

6. CONCLUSIONS

Interconnect and packaging technology scales down much slower than IC-technology. This results in an increasingly bigger *interconnect gap* between the advanced IC-technologies and the advanced PCB technologies. Multilayer thin-film technology, using 1X photolithography, is a key enabling technology for bridging this gap and realizing the so-called SIP solutions.

The *above-IC* thin film technology used for flip chip redistribution can be considered as an integral part of the on-chip wiring hierarchy. It can also be used to realize very high-Q on-chip inductors; a low impedance power and ground supply; or very fast on-chip electrical interconnects.

REFERENCES

[1] http://public.itrs.net/.
[2] Kuniba, M., 2002, Semiconductor package substrates, technology trends and forecast, *Market Trends*, Gartner Inc.
[3] Beyne, E., 1996, Issues and opportunities in thin-film MCM, *2nd European Conference on Electronic Packaging Technology*, EuPac '96, January 31–February 2, 1996, Essen, Germany.
[4] Ruythooren, W., et al., 2000, Electrodeposition for the synthesis of microsystem, J. *Micromechanics and Microengineering*, **10**(2), 101–107.
[5] Beyne, E., Van Hoof, R. and Achen, A., 1995, The use of BCB and Photo-BCB dielectrics for high speed digital and microwave applications, *Proc. 1995 International Conference on Multichip Modules*, April 19–21, 1995, Denver, Colorado.pp.513–518.

[6] Beyne, E., Van Hoof, R., Webers, T., Brebels, S., Rossi, S., Lechleiter, F., Di Ianni, M. and Ostmann, A., 2001, A. high density interconnect substrates using multilayer thin-film technology on laminate substrates (MCM-SL/D), *Microelectronics International*, **18**(3), 36–42.
[7] Degryse, D., Labie, R., Vandevelde, B., Gonzalez M. and Beyne, E., 2002, Bond pad pitch reduction down to 40 μm: Influence on solder joint fatigue for flip chip structures, *Proc. EuroSIME 2002*, April 15–17, 2002, Paris, France, pp. 316–321.
[8] European Union FP5 project, IST–1999–10023 "CIRRUS"
[9] Peeters, J. and Beyne, E., 1994, Broad band modeling and transient analysis of mcm interconnections, *IEEE-CPMT Part B*, **17**(2), 153–160.
[10] Truzzi, C., Beyne, E. and Ringoot, E., 1997, Performance analysis of MCM systems, *IEEE* Transactions on Components, Packaging, and Manufacturing Technology, *Part B*, Advanced Packaging, **20**(3), 333–341.
[11] Truzzi, C. and Beyne, E., 1999, On the use of bare-die field programmable devices in miniaturized systems, *Proc. IMAPS HDP and Proc. SPIE*, April 7–9, 1999, Denver, CO, USA, **3830**, pp. 120–125.
[12] Becks, K. H., Beyne, E., Ehrmann, O., Gerlach, P., Gregor, I. M., Pieters, P., Topper, M., Truzzi, C. and Wolf, J., 1999, A MCM-D-type module for the ATLAS pixel detector, *IEEE Trans. Nuclear Science*, **46**(6, part 2), 1861–1864.
[13] Beyne, E., 2001, Technologies for very high bandwidth electrical interconnects between next generation VLSI circuits, *IEEE-IEDM 2001 Technical Digest*, December 2–5, 2001, Washington, D.C., pp. S23–P3.
[14] Pieters, P., Brebels, S. and Beyne, E., 1996, Integrated microwave filters in MCM-D, *Proc. IEEE-MultiChip Module Conference MCMC-96*, February 6–7, 1996, Santa Cruz, California.
[15] Pieters, P., Vaesen, K., Brebels, S., Mahmoud, S., De Raedt, W., Beyne, E. and Mertens, R., 2001, Accurate modeling of high Q-inductors in thin-film multilayer technology for wireless telecommunication applications, *IEEE Trans. MMT-S*, **49**(4), 489–599.
[16] Carchon, G., et al., 2001, Multilayer thin-film MCM-D for the integration of high- performance RF and microwave circuits, *IEEE Trans. Components and Packaging Technologies*, **24**(3), 510–519.
[17] Carchon, G., et al., 2003, High-Q RF inductors on standard silicon realized using wafer-level packaging techniques, *MMTS Conference*, June 9–12, 2003, Philadelphia.
[18] Balachandran, J., et al., 2004, Compact broadband resistance model for microstrip transmission lines, *IEEE CPMT and MTT 14*th *topical meeting on Electrical Performance of Electronic Packaging*, *EPEP 2004*, October 24–27, 2004, Portland, OR.

Chapter 3.2

CMOS IMAGE SENSORS FOR AMBIENT INTELLIGENCE

Albert J.P. Theuwissen, Martijn F. Snoeij, X. Wang, Padmakumar R. Rao, and Erik Bodegom
Department of Microelectronics/DIMES, Delft University of Technology
{a.j.p.theuwissen, m.f.snoeij, x.wang, p.rao}@tudelft.nl; bodegom@pdx.edu

Abstract Ambient Intelligence (AmI) hinges on three key technologies: ubiquitous computing, ubiquitous communication, and intelligent user-friendly interfaces. For intelligence and communication to be of relevance to people, sensor inputs are essential. One of the most versatile and, in our estimation, crucial of all possible inputs is the image sensor. To fulfill the dream of AmI and to assure widespread adoption of the image sensor, one must address the following criteria: cost, power, versatility, and adaptability. We will present an overview of the current state of the image sensor and indicate some of the ways in which hardware designers might meet the challenge of addressing these criteria.

Keywords image sensors

1. INTRODUCTION

Ambient intelligence (AmI) relies on inputs from the external world to enhance, improve, facilitate, and to secure the life of humanity. Applications range from automotive (collision avoidance, smart highways) [1, 2], security (biometrics, fingerprint recognition for payments in grocery stores, facial recognition for border control, and Nigerian scam baiters)[3, 4], medical (assisted living, diagnostics) [5, 6], hazardous operations (mine sweeping, smart buildings) [7], social (video conferencing, museums) [8], cultural (sports) [9], and gaming (virtual reality) [10] to the esoteric, such as satellites (star trackers) [11]. Inputting the external stimuli will very often involve visual cues and hence the primacy of the image sensor.

S. Mukherjee et al. (eds.), AmIware, 125–150.

It was in the mid-eighties that solid-state image sensors replaced the classical imaging tubes in video applications. Although the first papers on solid-state imaging were published in the sixties, it took almost two decades before solid-state imagers could make any inroads into the consumer market. The main reason for this long design-in period was the lack of a mature technology to fabricate the imaging devices. Photographic equipment has recently gone through the same process as the imaging tubes did in the past. The cost-quality ratio of solid-state image sensors is becoming so attractive that they pop up in many new, emerging markets that are the forerunners of AmI, like mobile imaging and automobiles. The AmI era will be the ultimate mass market for solid-state imagers.

Today's image sensors are based on two technologies: charge-coupled devices (CCDs) [12, 13] that are made in a dedicated semiconductor process and complementary metal-oxide-semiconductor (CMOS) imagers () that are produced in a standard semiconductor process [14–16]. For AmI applications, cost and power consumption are the two most crucial characteristics. Therefore, it should not be surprising that CMOS with their cost advantages, when produced on a large scale and their inherent power advantages over CCDs is believed to be the technology of choice for AmI.

The goal of this chapter is to provide an overview of the state of the art in CMOS sensors, and provide a roadmap for future developments. The layout of the chapter is as follows. In the first section, the overall underpinnings of image sensors are presented. The next section involves global power considerations and on-chip processing, followed by a section on the analog chain. The last section highlights the myriad possibilities that can be addressed by novel designs to solve unique issues. Finally this chapter concludes with a summary and an outlook for imager design and potential.

2. IMAGE SENSOR ARCHITECTURE

Imagers can be built up one-dimensionally (e.g., facsimile), but most of them are constructed in a two-dimensional configuration (e.g., video, automotive, and mobile).

2.1. Basics of CMOS Imagers

A CMOS XY-addressable imager is a matrix of photodiodes, each of which is provided with a MOS transistor acting as a switch. This setup is shown schematically in Figure 3.2–1. The basic structure of a unit cell

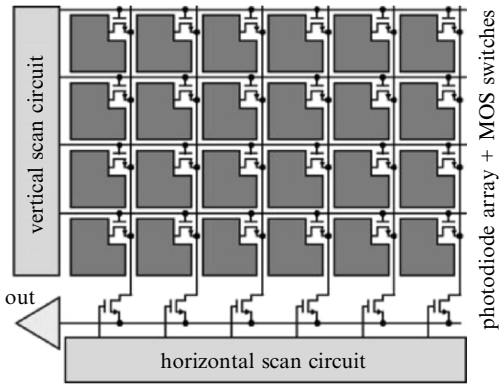


Figure 3.2-1. The basic architecture of a CMOS XY addressable imager with an array of passive pixels.

consists of a photodiode connected to the output by means of a sensing line. To convert the two-dimensional spatial information into the serial stream of electrical signals, electronic scan circuits are added to the device to address all pixels in a sequential mode and to read out their information. This addressing feature forms the basic operation of the device. At the beginning of a new field, the vertical scan circuit is activated. Suppose that the first row of pixels is selected. This is done by setting a high DC voltage on all gates of the MOS switches of this first row. Next, the horizontal scan circuit selects the pixels on one particular column by scanning its own outputs using a single high DC output, while all others are at a low level. This combination of one output of the vertical scanner and one output of the horizontal scanner at a high level and all the others at low level, selects one single pixel from the two-dimensional matrix. This pixel can be emptied and its information dumped into the output stage. Immediately after this action, the pixel can be reset and restart an integration. The neighboring pixel will then be addressed and read out.

The XY-addressed imager with passive pixels (the simple photodiode) is limited in its performance due to the relatively high levels of noise and fixed-pattern noise (FPN). FPN is the variation in output between various pixels given the same input. In a perfect imager, each pixel produces the same output given the same input, but in actual image sensors, the output of each pixel is different. These differences are partly caused by the small capacitance of the photodiode, which is connected to the large readout capacitance of the sensing line. A small amplifier can be added to every pixel to overcome these noise sources. A source-follower stage (amplification in the charge domain) is constructed within every single pixel, and

with appropriate correlated-double sampling circuitry (see Section 4) on the column buses, the noise and the fixed-pattern noise can be limited.

The architecture of the active pixel sensor (APS) is very similar to the passive pixel sensor (PPS). Figure 3.2–2 shows this architecture. In both cases, after selection of the appropriate pixel by the scan circuitry, the information is read from the pixel—the charge is transported from the photosensitive area towards a floating diffusion amplifier by means of the transfer gate. The driver of the source follower stage (the amplifier in the pixel) is implemented in the pixel itself, and the load of the source follower is common for all pixels on a column.

Compared to the PPS, the overall noise and fixed-pattern noise performance of the APS sensors is improved by at least one order of magnitude.

2.2. Pixel Configurations

The *passive pixel* is very simple in construction (see Figure 3.2–3) with 1 photodiode, 1 transistor, and 2 interconnects. The pixel is characterized by a large fill factor, but unfortunately also by a large noise level. After addressing the pixel by opening the row-select (RS) transistor, the pixel is reset along the column bus and through RS.

The *active pixel* has an amplifier within every pixel (see Figure 3.2–4). The amplifier is configured as a source follower where the driver of the source follower is located in the pixel itself, and the load of the source follower is placed on the column bus, but physically out of the focal plane. The two other transistors in the pixel are used for addressing (RS) and

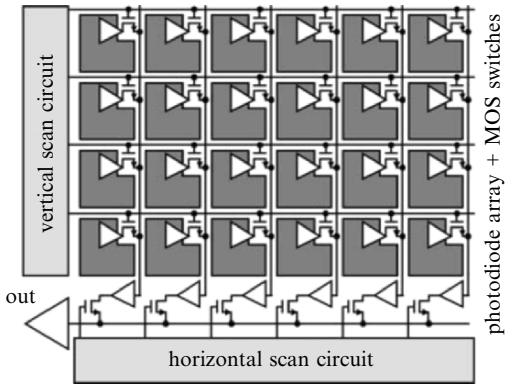


Figure 3.2-2. The basic architecture of a CMOS XY addressable imager with an array of active pixels.

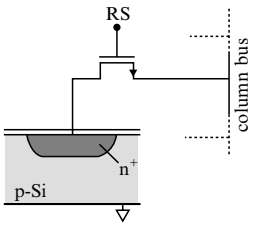


Figure 3.2-3. Basic configuration of a passive pixel.

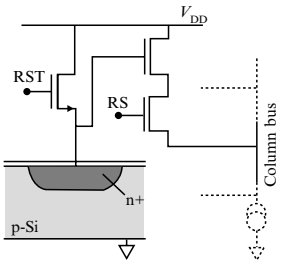


Figure 3.2-4. Basic configuration of an active 3T pixel.

resetting the pixel (RST). Every pixel has 1 photodiode, 3 transistors, and 4 interconnects.

After addressing a pixel, its actual video level is sensed by the source follower and fed to the column bus then the pixel is reset, and a new integration cycle can start. This APS pixel has become very popular in CMOS image sensors, because it solves or mitigates many of the noise problems. Unfortunately, in the configuration shown, it still suffers from a large reset noise component. This drawback is solved in the pixel configuration known as the *pinned-photodiode* (PPD) pixel (Figure 3.2–5). The operating principle of this PPD pixel is based on four major steps in the readout cycle: after addressing the pixel by RS, first the reset of the floating diffusion takes place, next the readout of the voltage across the floating diffusion is done, then the charges are transferred towards the floating diffusion and finally the readout of the voltage across the floating diffusion after this transfer. Subtracting the signals from the two readout cycles allows one to obtain a video signal that is almost completely noise-free. Because of this interesting characteristic, the PPD concept is becoming very popular as measured by the increase in the number of articles at the ISSC Conference. Some manufacturers are starting to deliver products based on this concept. The downside of this design is

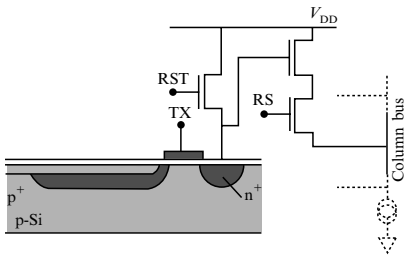


Figure 3.2-5. Basic configuration of a pinned-photodiode pixel.

that the pixel is quite large, because it is based on 4 transistors and 5 interconnects within every pixel.

A solution to the latter problem can be found in the *shared pixel* concept: four neighboring pixels share the same source-follower (see Figure 3.2–6). By means of the RS switch, a group of four pixels is selected, but the transfer of video information from a dedicated pixel towards the shared output stage is driven by its individual TX pulse. Therefore, overall selection is done by RS; individual pixel selection (within a group of four) is done by one out of the four TX pulses.

The shared pixel concept allows to combine within 1 pixel low-noise, high-light sensitivity with a minimum number of transistors (7 transistors/4 pixels = 1.75 transistors/pixel). Because of its compactness, this pixel is very well suited for AmI applications.

3. POWER

Focusing on AmI applications of the CMOS image sensor, low power is, besides cost, certainly the main issue. Considering the power distribution

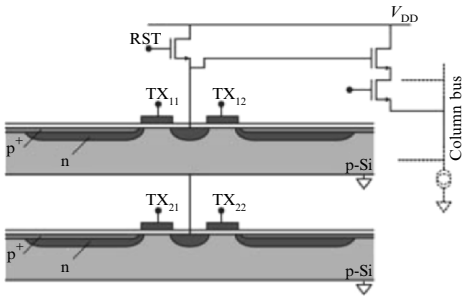


Figure 3.2-6. Shared transistor concept: four pixels share the source follower.

inside a camera, it makes little sense to consider only the pixel-level power consumption. Figure 3.2–7 depicts the power dissipation for a QCIF (176×144) and VGA (640×480) CMOS image sensor, both operating at 60 Hz frame rate with analog to digital converters (ADCs). It clearly shows that the majority of the power is actually not consumed in the pixel array. I/O, color signal processing, and ADC totally consume more than half of the power. Therefore, it is important to investigate the power issue in these components.

3.1. Power Consumption in the Signal Processing Unit

The signal processing can always be divided into two parts: analog signal processing and digital signal processing. In a CMOS image sensor, analog signal processing is essential to achieve high signal-to-noise ratios (SNR), and especially to minimize pixel and column FPN. The digital part normally takes care of the remaining image processing procedures (e.g., color interpolation and correction, white balancing, and aperture correction).

After analog processing (see Section 4), the video signal is converted to a digital signal by the ADC. Comparing with the analog part, the image digital signal processing certainly consumes much more power due to the added and multifaceted functionality. At the moment digital image processing units are mostly designed for photography applications [17, 18] (i.e., color imaging). The digital signal processing that is necessary to correct the various problems is so complex, that the corresponding algorithms can even vary from pixel to pixel within one sensor [19].

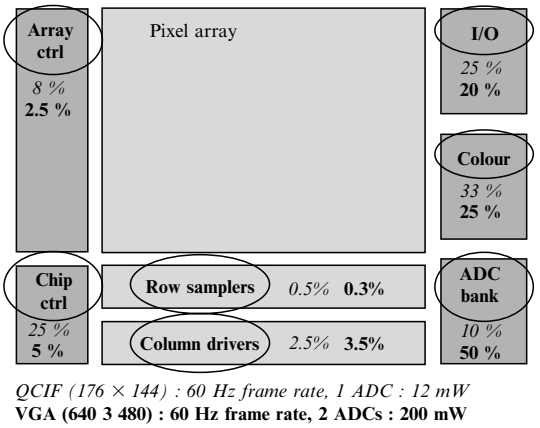


Figure 3.2-7. CMOS image dissipation for a QCIF and a VGA pixel.

In addition, due to this complex functionality, besides the power issue, the cost increases and these sensors are not very suitable for widespread AmI applications.

Fortunately, in many cases, color processing is not necessary for AmI. Therefore, it is critical to develop an application-specific signal processing (ASSP) design, like in, for instance, motion detection, edge detection, or 3D detection (of interest in, for example, virtual reality), to acquire a simplified structure and reach the ultralow power consumption needed. As a good example, researchers recently implemented a camera processor design using a CMOS image sensor especially for motion detection [20]. This is definitely an important application in AmI—to realize automatic inspection of environmental surroundings and detecting changes that are introduced. Figure 3.2–8 shows the block diagram of the proposed algorithm.

Traditionally, in order to detect motion, a previous image needs to be stored and compared with the current image. This leads to the need for a large memory and increased computing power during the required signal processing. In essence, the image data bit width used for comparison decides the power consumption. The novelty in this algorithm design is to choose only the most significant bit instead of all the data bits to make the comparison while not losing accuracy. The image signal processing unit as reported, includes an embedded microprocessor (ARM) and several hardwired modules using pseudo advanced microprocessor bus architecture (AMBA). Together with a power management block, which controls the system clocks by software, a reduction of 1/3 in overall power consumption can be achieved.

Even though the image processing functions are normally realized in the digital domain, the argument exists, that because digital processing require high resolution data, it may lead to a bigger chip size, and most of all, more power consumption. Therefore, there are designs [21] and

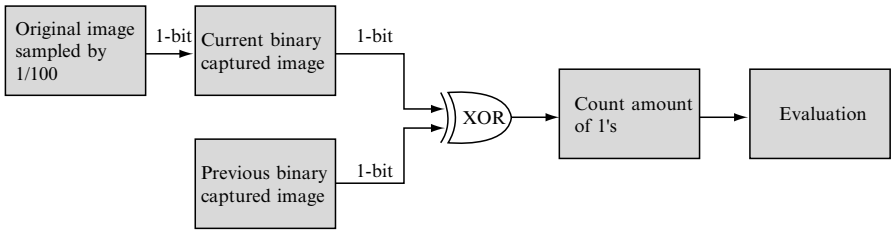


Figure 3.2-8. Block diagram of the algorithm for a low-power motion detection sensor.

products [22] to integrate most fundamental signal processing before the ADC. Figure 3.2–9 is the functional block diagram of one of such products, OV7640 color CMOS VGA CameraChip from Omnivision Technology. Without a DSP unit, the only digital domain components are output formatter and the I/O circuit. Comparing with Figure 3.2–7, the same VGA resolution (640 × 480) with a dual ADC configuration is reported to use only 40 mW active power (30 fps) including the I/O power. Even considering the lower operation speed, dramatically reduced power consumption is seen.

With an application-specified processing unit design, power consumption in signal processing unit can be optimized and reduced further. Moreover, there are still quite a few strategies for power reduction in the DSP unit (e.g., dynamic voltage and frequency scaling and resource hibernation). A lot of work is being carried out especially concerning power issues in processor design, both at the hardware and software level [23]. Also, due to the simpler functionality desired by AmI applications, further power reduction can be achieved by shifting the processing procedures from the digital to the analog domain. By doing this, the well-established low power analog circuit design techniques can also play a role.

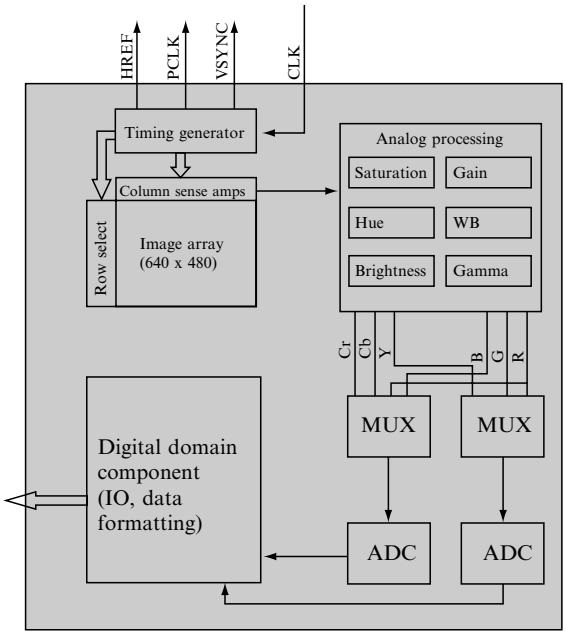


Figure 3.2-9 Functional block diagram for OV7640 VGA CameraChip.

3.2. Power Consumption in the I/O Unit

The I/O circuitry in a CMOS image sensor is another power-hungry component. It contains all the input and output control signals, such as clock, reset, and video output signals (analog/digital). For standard photography, a CMOS image sensor (e.g., a typical VGA resolution sensor) has more than 30 I/O pins (8-bit RGB output formats) [22].

In order to adjust output current according to different loading needs, driving circuits such as buffers and amplifiers are needed for video output ports. Unfortunately, the downscaling of the technology leads to a reduced on-chip capacitance, while the off-chip capacitance remains the same. Therefore, the driving ability of the I/O unit becomes extremely important; and as the consequence, higher power consumption is expected.

To achieve a low power design, the most straightforward method is to limit the amount of I/O. By doing this, the number of bits on the system-bus can also be limited, which results in a significant reduction of power dissipation due to its high switching activities and large capacitive load [24].

For example, a number of pixel parallel ADC circuits have been reported based on a pulse width modulation (PWM) scheme [25]. A similar design with on-pixel ADC based on PWM is reported [26]. Figure 3.2–10 is the pixel structure of this design. The main scheme of PWM remains the same. The difference of this design is that the on-pixel 8-bit memory can be reconfigured as a 4-bit counter. Therefore, the global

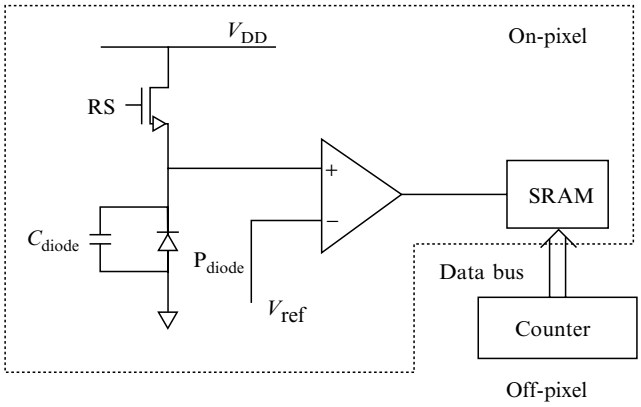


Figure 3.2-10. Pixel Structure of the on-pixel PWM ADC design.

data bus lines are also reduced by half. During integration, the in-pixel 4-bit counter generates the 4 MSB with respect to the video signal; an array-based off-pixel counter generates the 4 LSB based on the same video signal. Immediately after integration, the on-pixel 4-bit counter is configured back as an 8-bit memory. By transferring the 4 LSB video signals from an off-pixel counter, the on-pixel memory holds the complete 8-bit digital video signal. Then, a special time multiplexing strategy is used to readout 8-bit data through the 4-bit bus. Therefore, the whole integration and readout phase can be handled with a 4-bit system bus. Logically, the power consumption due to system bus is reduced.

However, unless much less data is required at the output, for the same amount of data, the technique of limiting I/O may not produce the results wanted. Sometimes, the problems are just shifted downstream. Even though the amount of I/O is reduced by half, the switching speed may need to be doubled and extra circuits may need to be added in the pixel or in the ADC, which certainly leads to extra power consumption. Also other effects (e.g., fill-factor, and ADC speed) may be hurt because of it. Obviously, a careful trade-off is necessary.

In other words, the most efficient method for power saving is still an application-specified design, both for the signal processing unit and the I/O unit. Particularly for the AmI applications, since an ordinary digital video signal (e.g., 8-bit RGB) data is not compulsory, a smaller number of I/O ports become decidedly more attractive. If a detection functionality is desired, for example, then a single output *Yes* or *No* may just be enough. Theoretically, the power consumption from the I/O unit is then optimized.

4. ANALOG SIGNAL PROCESSING IN CMOS IMAGERS

4.1. Introduction

The analog signal processing chain is an important component of a CMOS image sensor. First of all, the quality of this processing chain determines, together with the properties of the pixel array, the quality of the sensor. Second, a significant part of the total power of the image sensor is consumed in the analog signal processing, as is depicted in Figure 3.2–7.

The analog signal processing chain of a conventional CMOS image sensor can generally be divided into three parts, as is depicted in Figure 3.2–11 [27]. The front-end of the chain is formed by the in-pixel readout

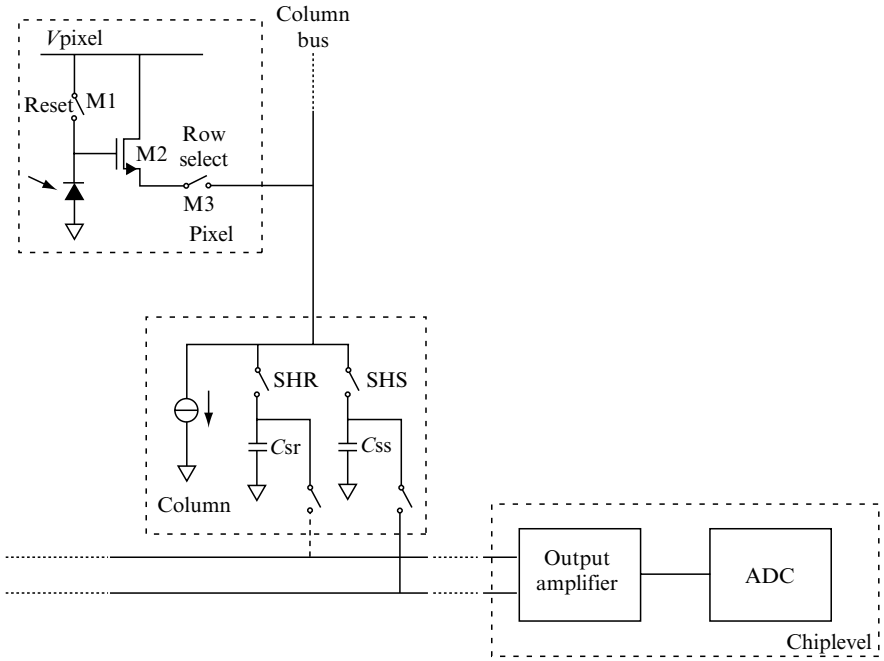


Figure 3.2-11. Overview of the analog signal processing chain of a CMOS imager.

transistor, which consists of only one source follower, as discussed in Section 2.2 of this chapter. To minimize the in-pixel area overhead, the source follower is usually of minimum size. As a result, it has a very high $1/f$ noise, particularly in modern deep-submicron CMOS processes. This $1/f$ noise cannot be decreased without fundamental changes to the processes, and usually therefore limits the overall quality of the processing chain.

The second stage of the chain is the column-level circuitry. This is a row of circuits situated below the imaging array, which is connected to one row of the imaging array at a time. The column circuit biases the in-pixel source-follower and samples both the pixel output signal and its offset. These voltages are stored on capacitors. Although many of these column circuits operate in parallel on a row of pixels, the combined power consumption of the first two stages is quite low. The reason for this is that in a conventional design, the column circuits are switched on only for a small fraction of the (line) time.

The last stage of the analog signal processing chain consists of a central output amplifier and an A/D converter. The output amplifier reads out the column-level capacitors. By subtracting the capacitor voltages of each

column circuit, a correlated double-sampling (CDS) is performed, thus removing offset and noise. Since only a single output amplifier is used, it reads out the row of column circuits at high speed. Sometimes, an automatic gain control (AGC) is added to the stage to decrease the necessary dynamic range of the A/D converter connected to the output of the amplifier. This A/D converter has usually a speed of 10–40 MSPS and a resolution of 8–12 bits. A pipeline or two-step architecture is commonly used for the ADC. The combination of output amplifier and ADC consumes most of the power in the analog signal processing chain, since it has to run at high speed, thus requiring high bandwidth circuits that consume a lot of power in order to have low noise.

In order to apply image sensors successfully in AmI, two aspects are of key importance. First and foremost, the power consumption of the imager has to be decreased as far as possible. Second, there should be a choice of output quality depending on the situation (i.e., it should be possible to adjust the necessary quality versus required power consumption dynamically). This has its impact on the analog signal processing chain, as will be shown in the next paragraphs.

4.2. Analog Signal Processing Topology for Minimum Power Consumption

Since most of the power of the analog signal processing chain is consumed in its ADC, efforts to minimize power should be focused on this part. Although there is a lot of research on optimizing the power consumption of general-purpose ADCs, imager read-out chains differ in one fundamental aspect: in an imager, several slower ADCs can be put in parallel without needing extra power in order to parallelize the signal path, as with a single-input general-purpose ADC. In recent years, research has been done on alternative signal processing chains, where the ADC is moved towards the front-end of the read-out. Several CMOS imagers were reported with a column-level ADCs [28–30] or even pixel-level ADCs [31, 32]. Some advantages of this approach are obvious, such as a possibility for high-speed read-out and/or very high-resolution imagers. However, how these different topologies compare on power efficiency is less obvious.

Although it is difficult to make an accurate comparison between the different topologies without designing them, some estimation can be made. Without regard to any specific ADC architecture, a fundamental aspect of analog circuitry is that voltage amplification is needed. The most basic circuit that delivers a gain is an inverting amplifier stage (Figure 3.2–12), loaded by an identical stage. If we regard only the transistor and consider its load resistor ideal, its bandwidth is determined by its transcon-

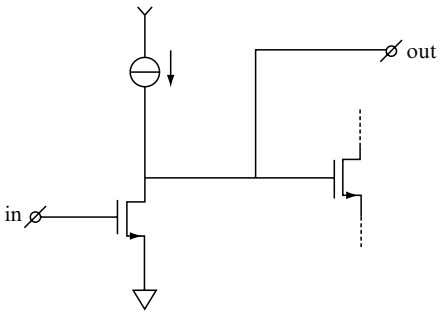


Figure 3.2-12. Basic amplifier configuration with one transistor.

ductance (g_m) and its parasitic capacitances, while its gain is determined by its transconductance and its output resistance. To maximize the power efficiency, it is therefore best to maximize the amount of transconductance the transistor can deliver for a given bias current. This means that the transistor has to be biased in weak inversion, where the ratio of transconductance to the bias current is maximum (typically around 25). This operation in weak inversion puts a lower limit on the *W/L* sizing of the transistor, which in turn determines the parasitic capacitance of the transistor. Therefore, if we choose to operate the transistor in weak inversion, there is a fixed relation between the transconductance and the parasitic load capacitance, and therefore the bandwidth does not depend on the bias current.

A simple computer simulation of the circuit (Figure 3.2–12) in a 0.18 μm CMOS process shows that the mentioned bandwidth is around 15 MHz. This figure is of course an upper limit to what should be the process' capability, in particular, in a real amplifier the capacitive load is probably higher. When a chip-level ADC is used, its speed should be at least 10MSPS for a VGA resolution sensor. As a result, the transistors in the ADC cannot be biased into weak inversion, since the amplifiers inside the ADC should have a bandwidth of at least 5 times the sampling speed (assuming switch-capacitor circuits). Based on this transistor-level analysis, multiple ADCs in parallel can be more power efficient if a suitable ADC architecture can be found that is at least as effective as a typical chip-level architecture (e.g., a pipeline ADC).

When we compare column-level and pixel-level ADCs reported in literature, it is immediately obvious that pixel-level ADCs are impractical for low-cost sensors due to their large pixel sizes. Moreover, the power consumption is probably not determined anymore by required transconductance, but by a minimum bias because of leakage currents, which will decrease the power efficiency. Therefore, column-level ADCs are the best compromise between necessary speed and amount of parallel circuitry. In the next paragraph, we will briefly describe some circuit aspects of such column-level ADCs.

4.3. Flexible Power/Quality Settings

Based on the assumption that a column-parallel ADC is the best topology to minimize power consumption, the next step in an ADC design for AmI is to design the ADC in such a way that the quality it delivers can be varied in order to vary the power consumption. This can be of particular interest in systems that do not have to deliver a high quality image most of the time, for instance a motion detection camera. When the camera image is only stationary, no high quality picture is needed, however, when motion is detected, a higher quality is desired.

This flexible quality and power can easily be implemented into a column-parallel ADC, if single-slope architecture is used [28, 29]. In Figure 3.2–13, a block diagram of such architecture, which is commonly used for column-level ADCs, is depicted. It consists of a single-slope ADC system, where a single, central ramp generator is driving a large number of comparators that are situated in each column. A central digital counter runs synchronized with the ramp generator and is connected to digital latches in each column. When a comparator detects that the ramp voltage exceeds the signal, the digital latch is triggered and a digital output is stored. The advantage of this architecture is that it minimizes the amount of circuitry needed in each column. Moreover, it is relatively easy to get a uniform response across all ADC channels, as only comparator offsets can

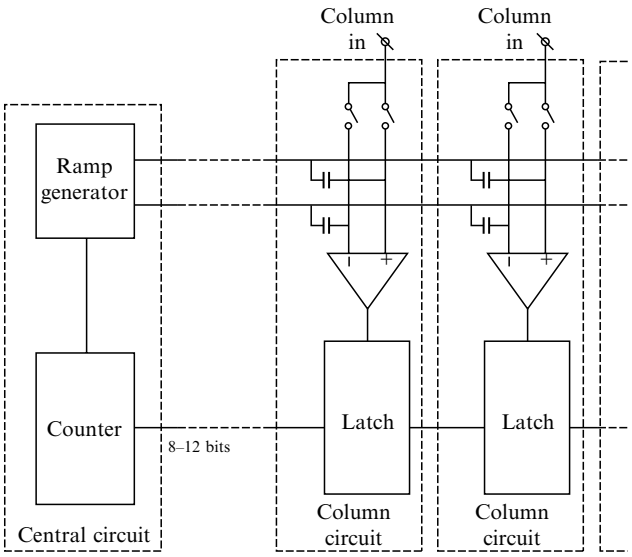


Figure 3.2-13. Column-parallel single-slope ADC architecture.

cause nonuniformities. These offsets can be corrected using the auto-zero technique.

The resolution of such an ADC depends on the steepness of the ramp and the clock speed at which the digital counter is running. Thus, the resolution of the system can be easily changed by changing the ramp generator and/or digital counter, which are both implemented on a central level. Therefore, although the analog signal path of the ADC is column-parallel, which reduces bandwidth and noise, the ADCs quality and thereby its power, can be easily changed at one central circuit.

Since it is easiest to design the column-level comparators for one speed, it is best to design them such that they are fast enough to convert an analog signal at the highest quality at the required speed. When a lower quality is needed, the steepness of the ramp can be increased, thereby reducing the resolution. As a result, the A/D conversion time decreases, and the comparators can be powered down for some of the time, thereby reducing power consumption.

Apart from changing the ramp steepness, we can also create a companding ADC by changing the shape of the ramp. When the ramp voltage is exponential, a dynamic compression of the input signal takes place, as is done in telephony systems. This can further reduce the power consumption. In particular, we can exploit the fact that some of the noise in an imaging signal increases at high signal values, due to photon shot noise in the pixel. Therefore, the resolution of the ADC can be decreased for high input signals without losing imaging quality.

5. SPECIAL DESIGNS

Technological improvements in the area of CMOS VLSI have fashioned devices to be more power efficient and cheaper, thus making it a suitable candidate for AmI applications. Moreover, a circuit designer has more flavors of architectures to choose from either to meet the stringent power requirements of AmI systems or that use specialized designs to achieve unique characteristics [33]. Logarithmic pixel and various current mode architectures can lead to very low power consumption. In the next level of hierarchy, a part of the computation usually processed in DSPs can be shifted to the pixel themselves by various spatial arrangement of the sensors that mimic the biological systems. In addition, employing massively parallel analog processing units for low-level image processing in the pixels further reduces the computational power budget of DSPs and application specific integrated circuits (ASICs) that follow the chain. Proper selection of the fabrication techniques to realize these sensors can

also help optimize the energy budget of a system. For example, silicon on insulator (SOI) techniques can be implemented for very low voltage and power applications. A brief description of these various techniques that can be used to improve the power budget for AmI systems or address special needs will be presented.

5.1. Various Electrical Structures

5.1.1. Logarithmic structure

The logarithmic structure, which has a logarithmic response to input radiation, is a good example of a *continuous response* pixel (see Figure 3.2–14). In this pixel architecture, the reset transistor (T1) operates in weak inversion mode since only a very small current flows through the load transistor when light is incident on the photodiode.

The current that flows through T1 can be written as

$$I = I_0 \exp\left(-\frac{V_g - V_s - V_{th(T1)}}{nV_t}\right),$$

where V_s and V_g are the source and gate voltages and $V_{th(T1)}$ is the threshold voltage of the load transistor. n and I_0 are process dependent parameters and V_t is the thermal voltage. The output voltage can be written as

$$V_A = V_{DD} - V_t \ln\left(\frac{I_{ph}}{I_0}\right).$$

The power consumption for an array of 525 × 525 pixels, employing the above structure in addition to on-chip calibration for FPN noise and

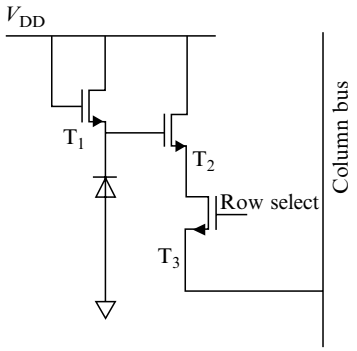


Figure 3.2-14. A logarithmic pixel structure.

ADC, can be typically some hundreds of milliwatts [34]. These pixels have the capability of capturing images with illumination ranges exceeding 120 dB. However, these pixel structures do suffer heavily from FPN, and image lag. Studies are being carried out to improve the noise factor and enhance the image quality.

AmI applications can also make use of the *Lin-Log* mode of pixel operation (see Figure 3.2-15) [35]. Here, by choosing the value of V_{bias}, the pixel can be chosen to operate either in the linear or in the logarithmic mode of operation. Thus, the image can be first captured using the *Log* mode of operation, and by using threshold and feedback blocks, the linear mode of operation can be carried out for specific pixels or region of interest by clever pixel select logic blocks. This effectively reduces the power consumption of the system.

5.1.2. Current mode pixel architectures

Current mode pixel structures greatly simplify integration of on-chip signal processing and lead to increased dynamic range while maintaining faster readouts and using a smaller silicon area. They are better suited for applications involving low voltages and are inherently buffered against voltage fluctuations. Current mode pixels tend to have large FPN, and efforts are on to rectify this problem [36].

Figure 3.2-16 presents such a configuration. During reset, transistors T1, T3, and T4 are on so that a known reference current I_{ref} given by

$$I_{\text{ref}} = \frac{1}{2}\beta(v_{\text{gsT2}} - v_{\text{thT2}})^2$$

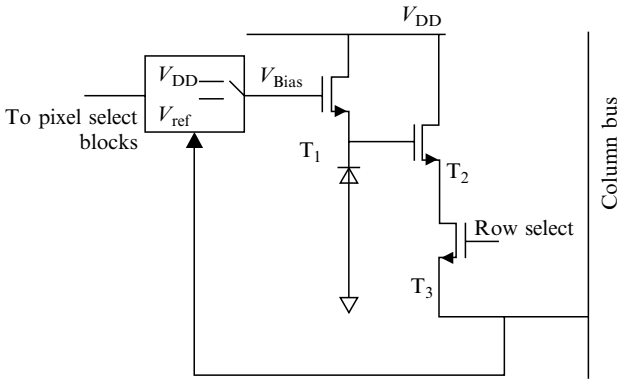


Figure 3.2-15. Lin-Log mode of pixel operation for smart systems.

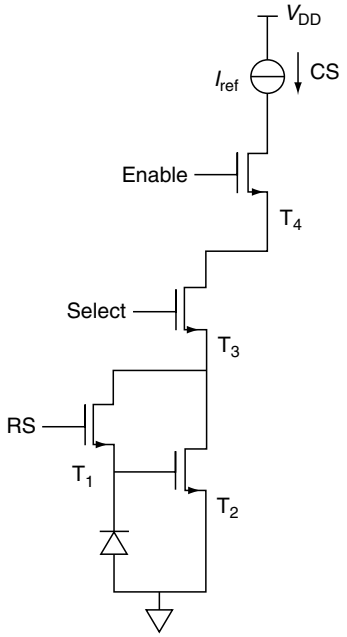


Figure 3.2-16. A current mode pixel employing a current mirror (T1 and T2).

flows through T2 and the gate voltage of T2 is *reset* to a voltage given by

$$V_{\text{ref}} = \sqrt{\frac{2I_{\text{ref}}}{\beta}} + v_{\text{thT2}}.$$

After an integration time, t_{int}, and a given photo current, I_{photo}, the voltage at the gate of T2 is reduced to $V_{\text{gsT2}} = V_{\text{ref}} - \Delta V$. This voltage can be read out as a current by closing T3 to connect the pixel to the column bus.

Furthermore, readout almost independent of threshold voltage of T2 is possible by keeping T4 enabled (automatic cancellation of FPN) [37], thus reducing the number of processing steps later in the chain.

5.1.3. The silicon retina chip for AmI

There have been improvements in smart system design based on biologically inspired structures. For most AmI applications, retina-like vision systems are well suited for object tracking and identification. The fovea and periphery approach introduces a higher concentration of sensors in

the fovea and an exponential decay of the density towards the periphery (log-polar imaging array). This architecture increases the data compression and lowers the power budget [38]. The power consumption of such a system can be made as low as 25 mW at video frame rates of 28 fps.

Image coordinates are mapped from the original image via a mapping template (a) to separate images for the fovea and the periphery. Data in the original image undergoes a one-to-one mapping to the fovea image. In the periphery, the receptive fields (RF) or the circles in Figure 3.2-17 cover multiple pixels. Hence, the data corresponding to a RF undergoes a many-to-one mapping in which all pixels within a receptive field are averaged to produce a single pixel value in the periphery image. RF's in the periphery are distributed along rays of angular displacement, $\Delta\theta$. All RF's on ring i have radial displacement given by $r_i = \alpha e^{i\beta}$ where α and β are determined by solving the equation for the given inner and outer edges of the periphery.

5.1.4. Pixel parallel analog processing

Analog processing in a parallel fashion in the pixels helps to reduce the computation budget that follows the chain. These computations can be considered as filters. Such filters can be realized by analog circuits, which

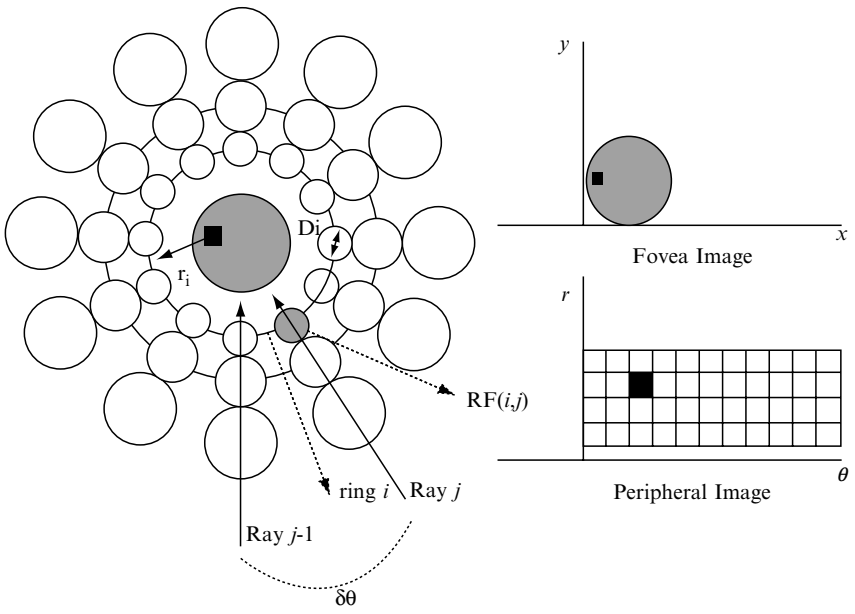


Figure 3.2-17. A foveated-mapping scheme.

have complex-valued impulse responses. For example, a filter has been implemented by Shi [39], which provides an orientation selective system. These filters can be electronically tuned by varying the bias voltages. One advantage of such filters is that they operate in continuous time. As an example, Figure 3.2-18 shows a section of a filter based on transistors operating in the weak inversion regime where the energy efficiency is maximized. The filter function can be thought of as the weighted sum of the pixel currents and a constant term. Transistor networks in weak inversion as shown in Figure 3.2-18 can realize such functions. The output signal (i.e., the processed image) provides a mechanism for differentiating object orientations.

In Figure 3.2-18, the weight of the coefficients of the filter function can be set by choosing the values of V_b and V_h appropriately. The value as a function of V_b and V_h is given by

$$\alpha_W \propto e^{\frac{k}{V_T}(V_b - V_h)}.$$

Current additions and subtractions can be performed at the nodes of the circuit. Current amplification, both positive and negative, can be performed by current amplifier circuits consisting of current mirrors. Thus, the currents indicated by $i(n-1)$, $i(n)$, and $i(n+1)$ are effectively

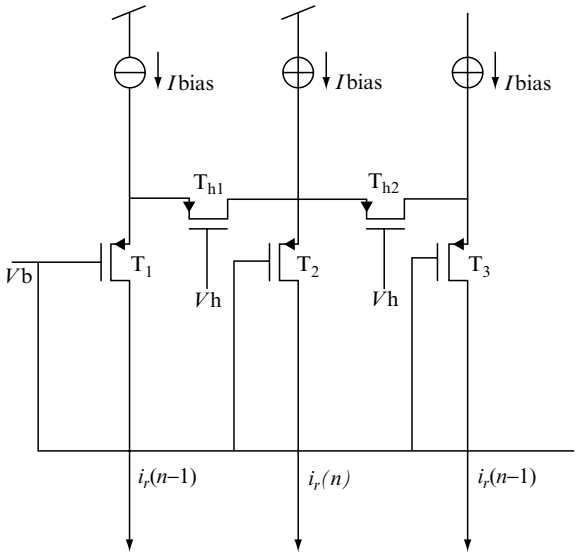


Figure 3.2-18. Weak inversion transistor network possessing a linear relationship.

the terms comprising the filter function. A network of such systems can be utilized to realize a filter completely, once the filter transfer function is known.

The power dissipation of such pixels is reported to be around 1.2 μW per pixel [39].

5.1.5. Pixel parallel analog to digital converters

As pixel readout is not limited by settling time as in column readout, pixel ADCs are an attractive solution for high-frame rate, low-power systems particularly for AmI applications. This also prevents the transfer of analog charges through long readout buses. A pixel parallel ADC employing the concept of a first-order synchronous Σ-Δ A/D converter is discussed below [25].

A free-running continuous oscillator when sampled at fixed intervals can be seen as a first order Σ-Δ ADC. Since it is a sampled free-running oscillator, it does not require clocked components and consists only of a single comparator.

The simple representation of a time mode pixel represented in Figure 3.2-19 compares continuously the node voltage of the diode and switches state when the voltage crosses a threshold value V_{low}. This information is obtained in the column bus as a PWM signal whose width corresponds to the slope of the integration curve. Pixel structures with power dissipation of less than 40 nW per pixel have been reported using this approach.

5.2. Structures at the Fabrication Level

Advancements in the area of fabrication technology have also helped improve the performance of sensors.

The SOI technology offers full dielectric isolation and various quasi-ideal properties like sharp subthreshold slope, low body effect, and high

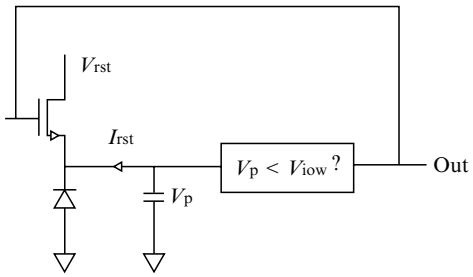


Figure 3.2-19. Time mode pixel architecture [25].

temperature operation as well as radiation hardness. These devices also promise low power operation, owing to the better coupling between the gate voltage and surface potential than in the bulk devices. The power dissipated in a circuit is proportional to $f \times C \times V^2$ where f is the frequency of operation, C is the sum of all capacitances in the circuit, and V is the supply voltage. Since all capacitances except the oxide capacitance are less in SOI devices, they dissipate less power as well. This makes a SOI device a good candidate for AmI applications. Figure 3.2-20 shows a hybrid bulk/SOI device structure [40]. In this device, the photodiode was fabricated in the usual SOI technique by an extra etching step to remove the buried oxide.

6. SUMMARY AND OUTLOOK

Some promising approaches that can reduce the power budget of image sensors fabricated in CMOS VLSI technology have been discussed. Options have been presented to improve and/or integrate the various structures for applications pertaining specifically to ambient intelligence. It is anticipated that new and/or improved designs will come along at an amazing rate, which, in turn, will lead to innovative uses that have not yet been contemplated. Use will be accelerated if interfacing standards are to be developed. A bright future awaits the CMOS image sensor, and chip designers will enjoy myriad challenges.

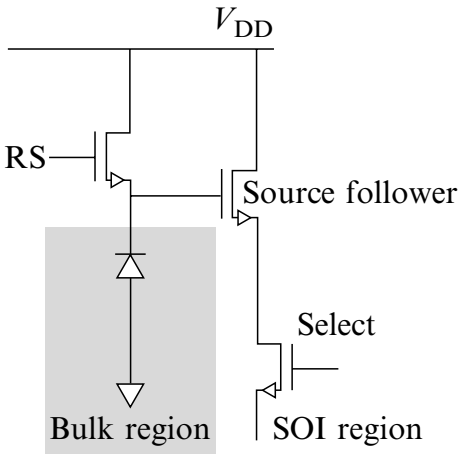


Figure 3.2-20. Layout of a hybrid bulk/SOI CMOS active pixel image sensor.

REFERENCES

[1] Klein, L. A., 2001, *Sensor Technologies and Data Requirements for ITS Applications*, Artech House Publishers, Norwood, MA, USA.

[2] Hosticka, B. J., Brockherde, W., Bussmann, A., Heimann, T., Jeremias, R., Kemna, A., Nitta, C. and Schrey, O., 2003, CMOS imaging for automotive applications, *IEEE Trans. Electron Dev*. **50**, 173.

[3] http://www.geocities. com/a_kerenx/prince.html.

[4] Lei, H. and Govindaraju, V., 2004, Direct image matching using dynamic warping, *First IEEE Workshop on Face Processing in Video*, Washington D.C.

[5] Bahadori, S., Cesta, A., Grisetti, G., Iocchi, L., Leone, R., Nardi, D., Oddi, A., Pecora, F. and Rasconi, R., 2003, RoboCare: An integrated robotic system for the domestic care of the elderly, *Proc. Workshop on Ambient Intelligence AI*IA-03*, Pisa, Italy.

[6] Daperno, M., Sostegni, R., Lavagna, A., Crocella, L., Ercole, E., Rigazio, C., Rocca, R. and Pera, A., 2004, *Eur. Rev. Med. Pharmacol. Sci*., **8**, 209.

[7] Micheloni, C., Foresti, G. L. and Snidaro, L., 2005, A network of cooperative cameras for visual-surveillance, *IEE Visual*, *Image & Signal Processing Special Issue on Intelligent Distributed Surveillance Systems* **152**, 205.

[8] Bertamini, F., Brunelli, R., Lanz, O., Roat, A., Santuari, A., Tobia, F. and Xu, Q., 2003, Olympus: An ambient intelligence architecture on the verge of reality, *Proc. 12th Int. Conf. Image Analysis and Processing*, 139.

[9] Bloomfield, J., Jonsson, G. K., Polman, R., Houlahan, K. and O'Donoghue, P., 2005, Temporal pattern analysis and its applicability in soccer, in L. Anolli, S. Duncan Jr., M.S. Magnusson and G. Riva (eds), *The Hidden Structure of Interaction: From Neurons to Culture Patterns*, IOS Press, Amsterdam.

[10] Manninen, T., 2003, 20 interaction manifestations in multi-player games, in G. Riva, F. Davide and W. A IJsselsteijn (eds), *Being There: Concepts*, *Effects and Measurement of User Presence in Synthetic Environments*, IOS Press, Amsterdam.

[11] de Boom, C. W., Leijtens, J. A. P., v.Duivenbode, L. M. H. and van der Heiden, N., 2004, Micro digital sun sensor: System in a package, *Proc. MEMS*, *NANO and Smart Systems: ICMEMS-2004*, pp. 322.

[12] Holst, G. C., 1998, *CCD Arrays*, *Cameras*, *and Displays*, SPIE Optical Engineering Press, Bellingham (WA).

[13] Burke, M. W., 1996, *Image Acquisition*, Chapman and Hall, London (UK).

[14] Theuwissen, A. J. P., 1995, *Solid-State Imaging with Charge-Coupled Devices*, Kluwer Academic Publishers, Dordrecht (Nl).

[15] Fossum, E.R.; Theuwissen, A.J.P, et al., 1997, IEEE transactions on electron devices, Special Issue on Solid-State Image Sensors, **44**, 1576–1801, IEEE Press, Piscataway (NJ).

[16] Fossum, E.R.; Teranishi, N.; Theuwissen, A.J.P.; Hynecek, J., et al., 2003, IEEE transactions on electron devices, Special Issue on Solid-State Image Sensors, **50,** 1–265, IEEE Press, Piscataway (NJ).

[17] Interfacing a CMOS image sensor to a d module, application note, Kane Computing Ltd. UK.
[18] McBader, S. and Lee, P., 2002, A programmable image signal processing architecture for embedded vision systems, *Digital Signal Processing*, **2**, 1269.
[19] Theuwissen, A. J. P., 2004, Image processing chain in digital still cameras, *Proc. Symp. VLSI Circuits*, pp. 2.
[20] Lee, S. H., Kim, S. W. and Kim, S., 2004, Implementation of a low power motion detection camera processor using a CMOS image sensor, *Proc. 2004 Int. Symp. Circuits and Systems*, **2**, 23.
[21] Yoon, K., Kim, C., Lee, B. and Lee, D., 2002, Single-chip CMOS image sensor for mobile applications, *IEEE J. Solid-State Circuits*, **37**, 1839.
[22] *OV7640 Color CMOS VGA CameraChip*, Omnivision Technologies.
[23] Brooks, D., Tiwari, V. and Martonosi, M., 2000, Wattch: framework for architectural-level power analysis and optimizations, *Proc. 27th Int. Symp. Comp. Architecture*, pp. 83.
[24] Irwin, M. J. and Narayanan, V., 1999, Energy issues in multimedia systems, *IEEE Workshop Signal Processing Systems,* **SiPS 99**, 24.
[25] McIlrath, L. G., 2001, A low-power low-noise ultra wide-dynamic-range CMOS imager with pixel-parallel A/D conversion, *IEEE Solid-State Circuits*, **36**, 846.
[26] -Fong Yung, Y. and Bermak, A., 2004, A digital CMOS imager with pixel level analog-to-digital converter and reconfigurable SRAM/counter, *Proc. 4th IEEE Int. Workshop System-on-Chip for Real-Time Applications*, pp. 33.
[27] Blanksby, A. J. and Loinaz, M. J., 2000, Performance analysis of a color CMOS photogate image sensor, *IEEE Trans. Electron Devices*, **47**, 55.
[28] Sugiki, T., Ohsawa, S., Miura, H., Sasaki, M., Nakamura, N., Inoue, I., Hoshino, M., Tomizawa, Y. and Arakawa, T., 2000, A 60 mW 10b CMOS image sensor with column-to-column FPN reduction, *IEEE Int. Solid-State Circuits Conf.*, **43**, 108.
[29] Findlater, K., Henderson, R., Baxter, D., Hurwitz, J. E. D., Grant, L., Cazaux, Y., Roy, F., Herault, D. and Marcellier, Y., 2003, SXGA pinned photodiode CMOS image sensor in 0.35 μm technology'' *IEEE Int. Solid-State Circuits Conf.*, **46**, 218.
[30] Takayanagi, I., Shirakawa, M., Mitani, K., Sugawara, M., Iversen, S., Moholt, J., Nakamura, J. and Fossum, E. R., 2003, A 1¼ inch 8.3 M pixel digital output CMOS APS for UDTV application, *IEEE Int. Solid-State Circuits Conf.*, **46**, 216.
[31] Yang, D. X. D., El Gamal, A., Fowler, B. and Tian, H., 1999, A 640 × 512 CMOS image sensor with ultra wide dynamic range floating-point pixel-level ADC, *IEEE Int. Solid-State Circuits Conf.*, **42**, 308.
[32] Kleinfelder, S., Lim, S., Liu, X. and El Gamal, A., 2001, A 10 kframe/s 0.18 μm CMOS digital pixel sensor with pixel-level memory, *IEEE Int. Solid-State Circuits Conf.*, **64**, 434.
[33] van der Poel, C., Pessolano, F., Roovers, R., Widdershoven, F., van de Walle, G., Aarts, E. and Christie, P., 2004, On ambient intelligence, needful

things and process technologies, *Proc. 34th European Solid-State Dev. Res. Conf.*, pp. 3.

[34] Kavadias, S., Dierickx, B., Scheffer, D., Alaerts, A., Uwaerts, D. and Bogaerts, J., 2000, A logarithmic response CMOS image sensor with on-chip calibration, *IEEE J. Solid-State Circuits*, **35**, 1146.

[35] Fox, E. C., Hynecek, J. and Dykaar, D. R., 2000, Wide-dynamic-range pixel with combined linear and logarithmic response and increased signal swing, *IS & T/SPIE 12th Int. Symp. Electronic Imaging*, **3965A**, 4.

[36] McIlrath, L. G., Clark, V. S., Duane, P. K., McGrath, R. D. and Waskurak, W. D., 1997, Design and analysis of a 512 × 768 current-mediated active pixel array image sensor, *IEEE Trans. Elect. Dev.*, **44**, 1706.

[37] Boussaid, F., Bermak, A. and Bouzerdoum, A., 2004, A novel ultra-low power reset/read-out technique for megapixels current-mode CMOS imagers, *IEEE Trans. Consumer Electronics*, **50**, 46.

[38] Wodnicki, R., Roberts, G. W. and Levine, M. D., 1995, A foveated image sensor in standard CMOS technology, *Custom Integrated Circuits Conf.*, **15**, 357.

[39] Shi, B. E., 2000, A low-power orientation-selective vision sensor, *IEEE Trans. Circuits and Systems—II, Analog and Digital Signal Processing*, **47**, 435.

[40] Xu, C., Zhang, W. and Chan, M., 2001, A low voltage hybrid bulk/SOI CMOS active pixel image sensor, *IEEE Elec. Dev. Lett.*, **22**, 248.

Chapter 3.3

MICROSYSTEM TECHNOLOGY FOR AMBIENT INTELLIGENCE

Geert Langereis
Philips Research Eindhoven
geert.langereis@philips.com

Abstract Sensor systems using microsystem technology are essential components to make our daily environment responsive. Microsystem technology enables widespread use of distributed systems with sensory inputs due to intrinsic low costs, low power, a high integration level, and small size.

Keywords ambient intelligence (AmI); micro electromechanical systems (MEMS); microsystem technology (MST); miniaturised total analysis systems (μTAS); system-in-package (SiP)

1. DEFINITION OF MICROSYSTEMS

Before starting a discussion on the impact of microsystem technology (MST) for ambient intelligence (AmI), it is essential to define the field of technology. The original concept of micromachining has evolved towards a broad scope, which could not have been foreseen in the early days, not even by the visionary survey of Richard Feynman [1, 2].

At the end of the 1970s and the beginning of the 1980s, the field of micro electromechanical systems (MEMS) originated from silicon technology. It was seen as the next logical step from integrated electronics towards surface-micromachining and bulk-micromachining integrated with control electronics. Already in an early phase, people managed to make micropumps in silicon for moving fluids. This initiated the trend towards microplumbing on wafer-scale by the creation of fluidic channels. Since then the term *microsystem technology* (MST) has become more common than solely MEMS.

S. Mukherjee et al. (eds.), AmIware, 151–177.

Scientists realized that fluid handling became even more interesting when (bio)chemical and physical sensors were integrated with the fluid handling system. The result is the trend towards a *Lab-on–a-chip* also known as *miniaturized total analysis systems* (μTAS).

Current problems in the field of microsystems are with respect to packaging the system, hence the commonly used name *system-in-package* (SiP). SiPs are the solution for chemical–mechanical sensor and actuator systems including industrial manufacturing and packaging concepts.

Even besides the broadening from the mechanical domain towards chemical and fluidic systems, I would like to widen the scope further. The sensory enablers for AmI do not necessarily have to be created using silicon technology. The SiP approach gives us the opportunity to package different types of technologies. Technologies which are sometimes more suitable with respect to costs, manufacturability, and performance can easily be integrated into a single package. The result is an omni-disciplinary field of physical domains and manufacturing technologies enabling the creation of smart interfacing nodes between microelectronics and our environment.

2. WHAT DO MICROSYSTEMS MEAN FOR AmI?

According to the Cambridge International Dictionary of English [3] the word *intelligence* means *The ability to understand and learn and make judgments or have opinions that are based on reason*. Let us have a closer look at the properties this dictionary attributes to the concept of intelligence.

Obviously, the ability to learn refers to *sensors* for collecting data. However, the definition is more specific than just referring to the collection of data. It requires an intelligent system to be capable of *making judgments* based on the sensor readings. For ambient intelligence this implies that sensor systems must be integrated with electronics for *data processing*. System designers will read *making judgments after reasoning* as retrieving signals from a noisy background, pretreatment of the data, and combining sensor readings in order to output reliable information.

But what is the consequence of putting intelligence in our environment, for example to create AmI*? When integrating reasoning sensors in our equipment and surroundings of daily life, these sensors should be small, numerous, and cheap. Therefore, miniaturization is essential, just as multiplicity and a high level of integration of all system components. Integration of sensor

*When using the word "intelligence" to refer to the level of adaptation in our environment, a direct consequence is that an "Ambient Intelligence Quotient" (AIQ) can be defined. The regime should be clear: when an igloo has the AIQ of a moron, the Home-lab at the High-Tech Campus in Eindhoven comes close to the 150.

systems in all types of materials, from fabrics to plastics, in all thinkable form factors requires the development of new fabrication technologies in new materials. Besides a direct consequence on the sensing element, the packaging method is exposed to specific AmI constraints as well. So, for AmI we need small devices, integrated with electronics into SiP solutions. This is exactly what was defined as being the microsystem mission in Section 1.

3. MICROCOSMOS

An amazing view on the boundless world of insects can be seen in the movie *Microcosmos* by Jacques Perrin [4]. This movie shows the industrious life of tiny little creatures filmed with macro-lenses in order to create a bug's eye view.

While observing the activities of insects on a millimeter scale, we may notice some remarkable things. The body of a swallowtail butterfly is huge with respect to its fragile legs. We can see a beetle rolling a stone of twice its size at tremendous speed. Surface tension makes water boatmen walking on the water and helps the carnivorous sundew plant to cover a grasshopper completely with a film of digestive juices. We can actually see a droplet of water evaporating. But the most surprising, however, is the water spider who collects air bubbles to make an under water nest in which it nibbles its freshly caught water flea.

Although they live in the same physical world as ours (i.e., where all the laws of physics are equal to ours), the viewer is over and over again deceived by his mechanical expectations. Our preciously built-up common sense of mechanics is fooled by ants and bugs!

What happens is that, although the physical laws are the same, the proportionality of physical effects differs. While surface tension can roughly be ignored in the centimeter and meter scale, it cannot be done so in the millimeter scale. A small water droplet evaporates at a rate that is significant with respect to its volume. The large stone appears to be lighter to the beetle than a rock of two meters would be to a human.

This movie can be the inspiration for every MEMS designer. It proves that while designing MEMS, we should abandon our common sense and reconsider the laws of physics.

4. SCALING AND MINIATURIZATION

The mismatch between our macroscopic mechanical sense and what is observed in the millimeter range can be physically explained by evaluating the art of scaling [5].

An example is shown in Figure 3.3-1. A cubic mass of size R^3 is suspended by a beam of length L, width b and thickness d. Due to gravity, the beam deflects and the mass displaces by a distance Δy. The mass of the cube is given by $m = R^3\rho$ with ρ the density of the cube material (we neglect the mass of the beam for simplicity). The spring constant of the beam is given by [6]

$$k = \frac{Ebd^3}{4L^3} \tag{3.3-1}$$

with E the Young's modulus. The force on the spring is equal to the force induced by gravity $F = k\Delta y = mg$ resulting into

$$\Delta y = \frac{\mathrm{m} \cdot \mathrm{g}}{k} = \frac{4\mathrm{g}\rho}{E}\frac{R^3L^3}{bd^3} \tag{3.3-2}$$

with g the gravity constant. When reducing the size of the structure of Figure 3.3-1, which means that each linear dimension is decreased by a factor of for example 0.1, the mass reduces by the power of three and the spring constant by the power of one. The result is that the deflection Δy does not reduce by a factor of 0.1, but by 0.1^2. Apparently, small structures appear to be stiffer. This is the reason that the body of a butterfly can be supported by legs, which are relatively thin with respect to the legs of larger animals.

A more general overview of scaling can be given by considering a certain relevant length S in physical structures. This length can be the length of an arm, the distance of an air-gap, or the thickness of a mem-

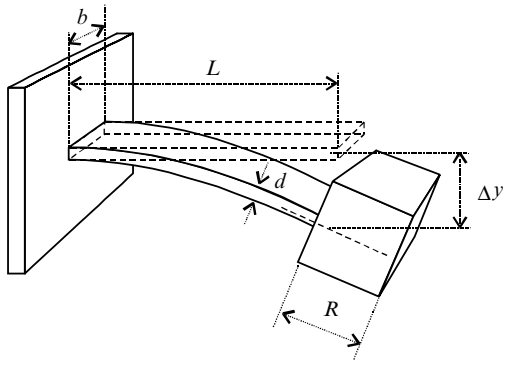


Figure 3.3-1. Deflection of a mass on a cantilever beam.

Table 3.3-1. Scaling laws.

Physical phenomenon	Scales with size S
Capacitor electric field	S^{-1}
Time	S^{0}
Van der Waal's forces	$S^{1/4}$
Diffusion	$S^{1/2}$
Size/velocity	S^{1}
Bending stiffness	S^{1}
Surface	S^{2}
Thermal loss	S^{2}
(Muscle) strength	S^{2}
Electrostatic force	S^{2}
Friction	S^{2}
Volume/Mass	S^{3}
Inertia	S^{3}
Magnetism	S^{3}

brane. As seen in the example above, masses of objects are related to volumes and therefore proportional to S^3. Bending strengths of beams are proportional to S as we have seen from Equation (3.3-1). The pulling strengths of beams and muscles are commonly proportional to their cross section S^2.

Capacitive transducers have the advantage that when reducing the air-gap, the capacitance and plate-to-plate force increases as S^{-1}. Very suitable for miniaturization, even when realizing that the capacitance decreases with S^2 as a function of the plate size. An example is the capacitive microphone, which needs an electret at normal scale to supply the hundreds of volts of biasing voltage. At micron scale the operational point can be biased from a simple low voltage power supply.

For magnetic transducers the situation is worse [7]. Assuming a constant current density through a wire or coil, the magnetic flux is proportional to S^3. For permanent magnets the scaling is proportional to the volume S^2.

More scaling rates are summarized in Table 3.3-1[†]. From top to bottom the phenomena are more dominant at a larger scale. MEMS devices are typically characterized by phenomena in the S^2 and lower domains. This means that magnetism, inertial forces, and masses are of lesser significance than surface tension, electrostatic forces, friction, and diffusion.

However, Table 3.3-1 does not say exactly at what dimension a certain phenomenon becomes prevailing. This depends on all specific geometries

[†] Note that some scaling powers are arguable depending on the configuration. For example, a force due to magnetic flux can be said to scale with S^3, although the magnetic force between two current carrying wires scales with S^4

and material properties. For certain phenomena there are some guidelines. In the field of microfluidics, the dimensionless Reynold's number is an indicator of whether flow is laminar or turbulent [5]. For large Reynold's numbers, convective and inertial forces dominate as we are used to with large objects in water. For small Reynold's numbers, on the other hand, viscosity is so large that transport of, for example, heat in the medium depends on diffusion rather than on convection. This will be the case in channels in the micron range.

Although time does not scale with size as a first order approximation, a smaller device will have a larger throughput. Microsystems will be faster in their response and consume less analyte in case of chemical systems. Diffusion based transport enables quick responses without the need for mechanical convective systems. This has resulted in static micromixers without moving parts [8]. Due to the fast diffusion processes, thermal and electrochemical actuation are new options for the creation of mechanical actions [9]. Another example of an application profiting from fast diffusion due to downscaling is the amperometric ultramicroelectrode [10].

5. OVERVIEW OF MST SENSOR TECHNOLOGY

Not all MEMS and SiP technology was created with AmI in mind. Whatever the motivation for the development was, miniaturization, manufacturability, and cost reduction must have been part of the goals. The result is a broad portfolio of sensory systems, which are available for giving our environment its sensing organs. All equivalences to the human sensory inputs are available: smell (e-nose), hearing (microphones), taste (chemical sensors), sight (camera), touch (temperature, pressure, and movement sensors). The following sections do not give an exhaustive list of devices covering all sensory inputs, just some highlights are picked out which indicate trends in this field.

5.1. Silicon Micromachining

Bulk silicon micromachining started with the development of anisotropic silicon etching in the 1960s to create free hanging masses and membranes. It took until the early 1980s before surface micromachining using sacrificial layer etching was developed. The step from bulk micromachining to surface micromachining enabled the fabrication of mechanical structures in the micron range and so the concept of MEMS was born. In the 1980s, MEMS using surface micromachining became a hype and

this period was characterized by fancy SEM pictures of micromotors and gearwheels created using silicon sculpturing. Practical usefulness was initiated by some examples of MEMS dies with integrated electronics. A striking illustration of the trend in early MEMS is the difference between the bulk-micromachined accelerometers of 1979 [11] and the integrated surface micromachined equivalent of 1982 [12]. The step towards a commercial product took until 1991 when analog devices revealed its first MEMS-based silicon accelerometer with integrated electronics, the ADXL-50 [13].

Silicon accelerometers with integrated electronics have occupied the automotive market as the sensory input for airbag actuation. The automotive market has been a profitable playground for multiple MEMS applications. Currently various types of gyroscopes, air pressure sensors (airco, tire pressure, etc.) and flow sensors have entered the market.

Besides the automotive market, the second largest market is found in the application of nozzles for inkjet printing. First publications originate from the late 1970's [14]. Some manufacturers include the silicon inkjet nozzles in their cartridges, which include a high-tech technology in a disposable product.

In terms of shipment and revenue, the largest growing business at this moment in the field of MEMS devices is the RF-MEMS business with an annual growth rate of 148.5% for the shipment and 54.9% for the revenues [15]. The reason is that RF-MEMS appear to increase power efficiency and allow for reconfigurable networks in radio-frequency (RF) applications. Designing voltage-controlled oscillators, tuneable filters/antennas and adaptive impedance matching circuits requires the combination of a large tuning range, high quality (Q) factors and low switching voltages [16]. MEMS technology appears to be very suitable for accomplishing this. Common RF-MEMS technology uses metal parts for the conductors and air-gaps for the capacitors. This results in lower losses and better on/off ratios than with semiconductor solutions.

At this moment, 85% of the components in the RF section of a mobile phone are passives. These passives generally consume significant space and are therefore less suitable to be implemented in the same process as the active electronics. As a solution, a technology platform is developed which combines the passive components onto a single high-resistivity silicon substrate [17]. This passive chip can be integrated with an active chip into a single SiP solution, which can have a size reduction of up to 50% with respect to conventional technology. The technology used for this is the PASSI™ technology shown in Figure 3.3-2.

It is the result of optimizing electrical performance, mechanical performance, cost of manufacturing, and process compatibility to the existing

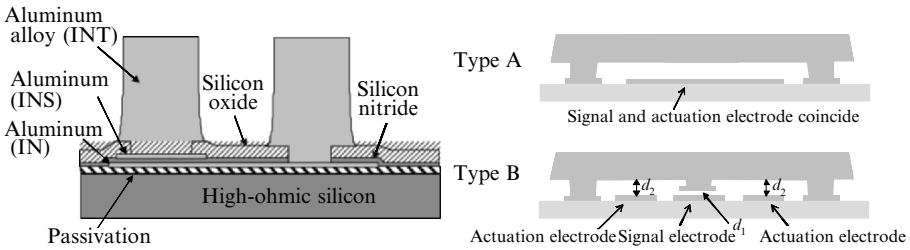


Figure 3.3-2. PASSI™ cross-sectional artist impression (left-hand) and arrangements into two types of capacitive MEMS structures (right-hand).

IC manufacturing infrastructure. The PASSI™ process stack consists of a high-ohmic silicon substrate and three aluminum layers with several silicon-dioxide and silicon-nitride insulation layers. The top aluminum layer is a 5 μm thick alloy and has a very low sheet resistance of 6 mΩ. It is therefore very suitable for defining interconnects, coplanar waveguides and high-Q inductors.

By selective sacrificial layer etching of one or more of the passivation layers, free hanging structures are made with two different air-gap sizes (right-hand of Figure 3.3-2). The lateral geometry is optimized to create good-quality RF-switches and tuneable capacitors as shown in Figure 3.3-3.

When the technology of dry etching macropores into silicon is added to PASSI-like processes, trench capacitors [18] can be realized with a high density of over 30 nF/mm^2. Such local capacitors are very suitable for supply-line decoupling in the GHz-regime in RF wireless communication. Breakdown voltages of over 30 V are realized, which show superior performance over conventional SMD implementations.

The PASSI™ technology is finalized by wafer-level encapsulation using a solder-sealing technique [19]. With such techniques, a cover is soldered

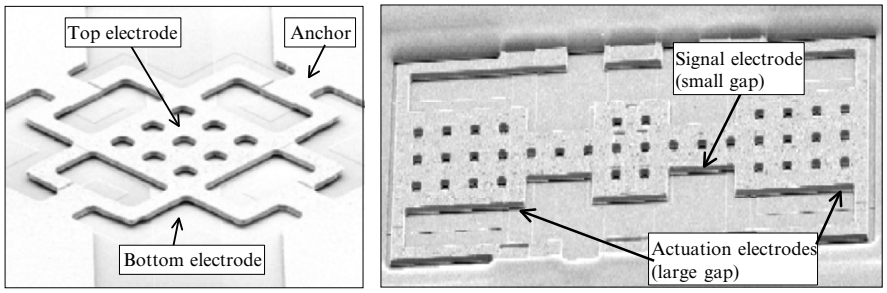


Figure 3.3-3. SEM pictures of a typical MEMS switch (left-hand) and a dual gap tuneable capacitor (right-hand).

on the wafer to cover the devices. In the soldering ring, a small hole is maintained through which air can be pumped out or through which a filling gas can be submitted. In a soldering reflow step the ring is closed subsequently. Thereafter, wafer dicing has become less critical.

So, surface micromachining has evolved into advanced systems like the RF-MEMS platforms for the creation of smaller circuits with better performance than conventional technology, optimized with respect to manufacturability. The PASSI™ technology places the passive components on the optimized PASSI-die while the active control electronics is on a conventional semiconductor substrate. Both dies are integrated into a single package.

5.2. Chemical Sensors

Generally speaking, chemical sensors consist of a selector part and a sensor part. The selector part determines a selectivity for a certain chemical or biological substance. Specific anticipation in the selection part results into modulation of a physical quantity in the sensor part which subsequently converts it into an electrical signal [20].

This two-stage set-up of chemical sensors results in typical key problems. Quantitative data from chemical and biological sensors are hampered by phenomena like drift, temperature changes and other interfering environmental conditions. The change in the selector part of the chemical sensor must be measured with respect to a certain reference. We have to deal with references in time and space. For example, electrochemical measurements require a closed electrical loop. This loop exists partially in the world of electrons and partially in the world of ions. When measuring the phenomena occurring at one ion/electron interface the interferences at the other ion/electron interface should remain unaffected. This is the problem of a reference in space, which requires the design of a proper reference electrode.

On the other hand, the reference in time is subject to drift. Interfering factors like temperature/pressure changes and diffusion of interfering gasses and chemical substances will cause an electric signal, which cannot be distinguished from the signal of interest.

Both types of reference problems can be cancelled out for a short period of time by means of calibration. To do this, the sensor must be placed in a known reference environment. With a one-point calibration we can cancel out effects like drift, with a two-step calibration we can determine the sensitivity as well.

However, there are more options which simplify the problem of drift under certain circumstances. One option is to use a stimulus-response meas-

urement in which the local environment is disturbed deliberately in order to learn from the response due to this disturbance. In that case, the drift of the sensor has become less critical since the problem is shifted to applying a known disturbance, which is generally easier to realize in a microsystem.

The key problems with chemical sensors are illustrated by the biography of the *ion sensitive field effect transistor* (ISFET [21]). From that overview we can read between the lines that the search for market applications of the ISFET is guided by the development of methods to define a decent reference and to calibrate the system. One smart solution is the coulometric microtitrator as developed at the University of Twente [22] where electrochemically generated H^+ or OH^- ions are used to control the local pH in the vicinity of an ISFET. The result is a titration curve from which the end point is detected from a nonlinearity in the pH response. In that case, the pH sensor calibration is less critical since quantitative information is obtained from the integrated current as supplied to the generator electrode.

A more fundamental and systematic description of differential sensor-actuator systems is given in literature [23]. It shows that besides solving reference and calibration problems, differential measurements and stimulus-response measurements give access to the detection of new parameters in some cases.

A review on chemical sensors in general was published by Janata et al [24].

5.3. Fluid Handling and Tubes

The advantage of silicon chemical sensors is that they can be used to analyze tiny volumes. The small volume of analyte results in a fast response. This is especially convenient for DNA analysis where the analyte is scarce and expensive and where many samples have to be analyzed sequentially or in parallel. Operation of the analysis should be simplified by implementing the fluid handling in a microsystem. Such a microsystem is referred to as μTAS or *lab on a chip* [25].

A typical μTAS comprises:

- A fluidic channel system for analyte transport and sample manipulation.
- Reaction chambers, separation columns, calibration and washing liquid storage containers.
- An electrical layer.
- (Bio)chemical sensors and physical sensors (pressure, temperature, etc.).
- Sample insertion system.

- A user friendly and easy to manufacture packaging concept.

Channels and reaction chambers can be made by dry or wet etching on silicon substrates or glass [26]. In the early days of fluid handling, MEMS micropumps were proposed [27]. Later on, the advantages of miniaturization were adopted by using the electroosmotic flow principle [28]. Electroosmotic flow uses the charge double layer, which is present on the interface of a fluid and the surrounding walls to apply an electric force to the bulk fluid. One advantage is that the flow profile is uniform while it is parabolic with a pressure initiated flow. Electroosmotic fluidic systems can be used to split and join streams without using fragile moving parts.

All types of chemical assay can be carried out in fTAS approaches. Separation techniques like electrophoresis and chromatography can be miniaturized on a wafer. Sample insertion is shown in various ways, for example using microdialysis probes [29].

5.4. Optical Systems

The classical example of optical applications for MEMS are the digital micromirror devices (DMD's). These devices consist of 800 up to 1 million reflective plates having a size of $16 \times 16\,\mu m^2$ each. The mirrors can be tilted electrostatically by angles of up to 10° in order to reflect light from an incident source. The result is that pixels can be projected on a screen using nonpolarizing optics at a very high speed. The product is on the market by DLP™ products, a division of Texas Instruments [30].

Micromirrors benefit from miniaturization by means of the speed of the mechanical actuation. Mechanical actuation, however, suffers from wearing and material fatigue. This is not the case when mechanical motion is initiated by electrowetting principles, which appear to be very suitable for optical applications.

Electrowetting is the principle of liquid manipulation using electrostatic forces. If a volume of an aqueous liquid has a cross-section of the order of millimeters, it can be manipulated in shape and position. The Duke University in North Carolina has shown some interesting results with fluid transport using electrowetting [31, 32].

An example of an optical application of electrowetting is the variable-focus liquid lens [33] as shown in Figure 3.3-4. A cylindrical housing is used which is coated on the inner side with a transparent conductor, shielded from the liquid by an insulating hydrophobic coating. The cylinder is filled with two immiscible liquids having a different refractive index. One of the liquids is electrically conducting, for example an aqueous salt solution, the other is insulating, for example a nonpolar oil. If both liquids

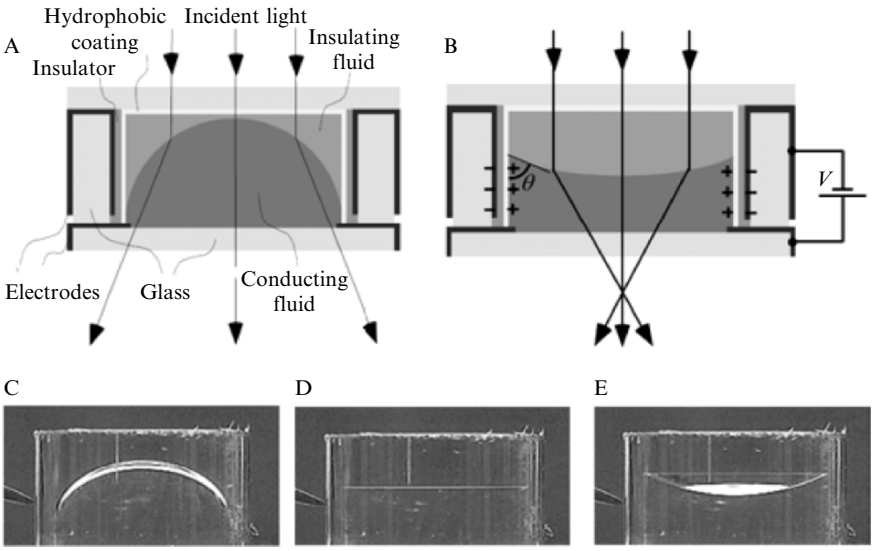


Figure 3.3-4. Schematic representation of the variable-focus liquid lens (A)–(B) and photographs showing the controllable meniscus (C)–(E).

have equal densities the shape of the meniscus is perfectly spherical, independent of orientation. A second electrical contact is made in the transparent bottom, which contacts the conductive fluid.

When a voltage is applied, charges accumulate in the wall electrode and opposite charges collect near the solid/liquid interface in the conducting liquid. The resulting electrostatic force effectively lowers the solid–liquid interfacial tension and with that the contact angle θ. Figure 3.3-4 (C) to (E) are snapshots of video frames of a 6 mm diameter lens taken at voltages of 0, 100, and 120 V. The switching speed is in the milliseconds range.

For camera applications, the adjustable lens can be integrated with corrector lenses, a CCD chip and control logic into a packaged image sensor. Typical applications of such devices are in the field of portable personal consumer electronics like PDA's and mobile phones where the increasing number of pixels requires lenses with adjustable focal distances.

Another example of the electrowetting principle for optical systems is the electrowetting display [34]. As shown in Figure 3.3-5, a colored oil film can be contracted into a localized droplet by applying a voltage. The transparent electrode yields a perfect reflective pixel principle when placed on a white background. When used in a full color system, the reflectivity is four times higher than in an LCD (67% versus 17%). A second advantage

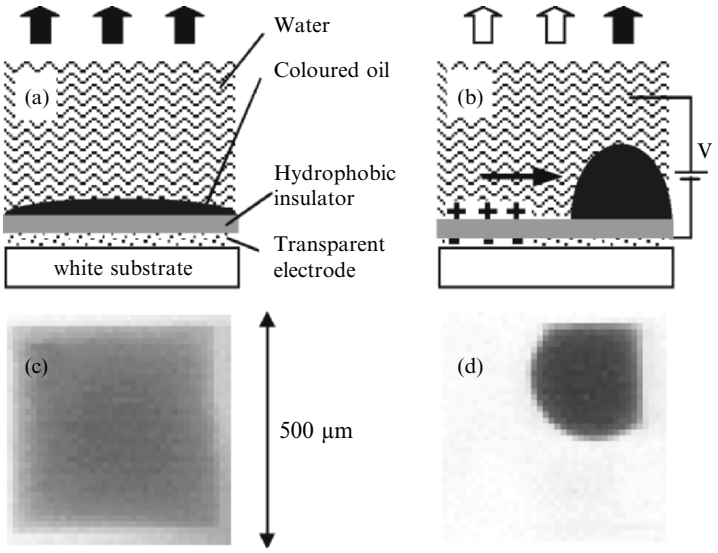


Figure 3.3-5. Electrowetting display principle. (a) No voltage applied, therefore a colored homogeneous oil film is present. (b) Voltage applied, causing the oil film to contract. Top row, diagrams; bottom row, photographs.

over an LCD is the wide viewing angle. Typical switching speeds below 10 ms are measured which is sufficient for a full motion display.

5.5. System Approaches

With respect to the academic MEMS work in the early days, publications in the field of MEMS have grown with respect to adaptation to the application. A good impression of present day work on MEMS can be obtained by scanning the special issue of the proceedings of the IEEE on biomedical applications for MEMS and microfluidics [35]. The presented microsystems are still illustrated by ingeniously fabricated three-dimensional structures, but the articles describe a real problem in the medical world and present an integrated microsystem providing a solution taking manufacturability and operability into consideration.

Consider the field of MEMS microphones. Common MEMS acoustical transducers are based on capacitive principles. An intrinsic problem of a capacitive transducer is the high output impedance, which requires dedicated preamplifiers placed close to the sensing device. The obvious solution is to design a technology in which the microphone and the preamplifier can be realized in CMOS technology on the same silicon

die. Several examples are published [36, 37], where the latter one is a comprehensive approach to realize acoustical systems capable of measuring sound, on-chip data processing and sound radiation.

However, the inconsistency of these approaches is the mismatch of technologies. A CMOS process is a high yield process with over thirty masks optimized for small area microelectronics. A silicon MEMS microphone is expected to have a lower yield due to the anisotropic or dry backetch and sacrificial layer etching. In addition, microphones consume considerable wafer surface which is acceptable because they are realized using typically five mask steps. The combination of a MEMS microphone with CMOS electronics does not combine the best of two worlds, but adds the weaknesses of both. The result is a relatively risky microphone process with the costs of CMOS electronics per unit of surface.

A better solution would be to choose optimized technology for both the microphone and active chip. Both dies can be packaged into a SiP, where the separation of high-impedance sensing element and preamplifier is still small. This is shown by Microtronic [38] where a CMOS die and a silicon MEMS microphone are packaged by flip-chip mounting onto a silicon carrier. The matching of technology to application (a separate MEMS and CMOS process) and separating technologies over two dedicated foundries is commercially much more attractive than the single-chip approach.

So, the SiP approach is a method to match technologies. We have already seen another example in Section 5.1 where the PASSI™ process was introduced as a base layer in which passive components and MEMS structures are integrated in order to flip-chip a CMOS dye onto it.

Combining several systems in a single package is not only interesting because of cost reduction but also creates new operational possibilities. In Section 5.2 the advantage of stimulus-response measurements for solving reference problems in chemical sensors was clarified. In that case, joining sensors and actuators added a value to the individual devices. Using multiple sensors in a system and combining the readings of the sensors might even yield more information. For example, N sensors can give information on $N + M$ parameters in case the M parameters have some correlation with the measured N parameters. In some cases the new parameters couldn't even be measured directly. This principle is called *polygraphy* or *multivariant analysis*. A descriptive example is the situation where multiple simple human-body sensors are measured (respiratory rate, skin resistance, skin temperature, skin potential, and heart rate) while the person under test is asked to shoot a target [39]. From the combined sensor readings we can see whether the test subject hit or missed the target!

Sensors in a miniaturized system usually give very local information, which reduces some disturbing effects. By adding actuators in the same

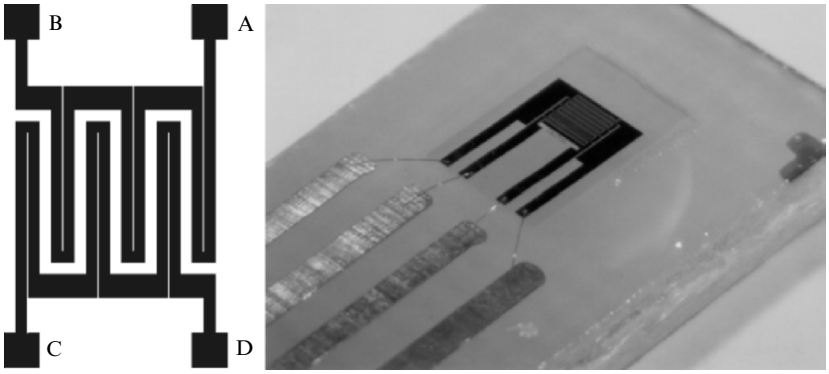


Figure 3.3-6. Integrated sensor-actuator structure: schematic functional representation (left-hand) and photograph of a packaged realization (right-hand).

local environment polygraphic sensor-actuator systems are made. The benefits of polygraphy, localization and sensor fusion are illustrated by the simple single sensor-actuator device [40] of Figure 3.3-6. The multi-purpose sensor structure consists of two meandered platinum wires on a glass substrate. The active area is $1 \times 1\,\mathrm{mm}^2$ while the length of the meanders can be up to 10 mm. The glass substrate with the metal structure is packaged onto a PCB and covered by an epoxy resin, except for the active area of the sensor, to create a dipstick probe. The probe was intended to be used to characterize aqueous solutions, especially laundry washing processes where rigid sensor structures are essential.

Three direct sensor modes can be distinguished. Between pads A and B (or C and D) a resistive path is present which functions as a temperature sensor. The two separated meanders can be switched as an interdigitated finger structure, which is a common geometry for electrolyte conductivity measurements. Finally, when using the whole surface of $1 \times 1\,\mathrm{mm}^2$ relative to an external large counter electrode, chrono-amperometric bleach detection is possible or potentiometric titration of weak acids.

Besides the three direct sensor modes, there are two actuator modes. The first one is the electrolysis mode where the electrode surface is used to generate H^+ or OH^- ions from water in order to control the local pH. The second one is the use of the resistive meanders for local heating.

The possible combinations of actuator modes with direct sensing modes result into a matrix of stimulus-response measurements. We can think of anemometric measurements to determine flow or turbulence by a time-of-flight or cooling down principle. From the relation between conductivity and temperature we can deduce information on specific ion

concentrations under certain conditions [29]. It might be even possible to do a coulometric precipitation titration of Ca^{2+} with conductometric endpoint detection in order to determine the water hardness.

So, by applying several signals (DC, AC, transient, etc.) in different configurations to the single structure of Figure 3.3-6 and extracting the proper signal we can deduce multiple parameters from the polygraphic data set.

6. MICROSYSTEM PACKAGING

The academic world has shown many inventive creations of silicon sensors [41]. What is omitted in almost all cases is a solution for the packaging. Nowadays, more and more solutions are provided for packaging devices with entries to the outside world [42, 43].

In some integrated systems, the package is part of the system. This is the case with the integrated track pointer as developed at Philips Research. On the left hand side of Figure 3.3-7 a cross-sectional drawing of the track pointer is given. A silicon substrate comprises signal processing electronics and planar integrated magnetoresistive (MR) sensors. The tilt of a ferrite stick can be measured with these MR sensors.

Conventional technologies suffer from drift and hysteresis when using piezoresistive sensors, or metal fatigue when using the capacitive principle. The packaging solution of the MR principle is robust and can withstand the tremendous horizontal forces it is exposed to by the user. The integra-

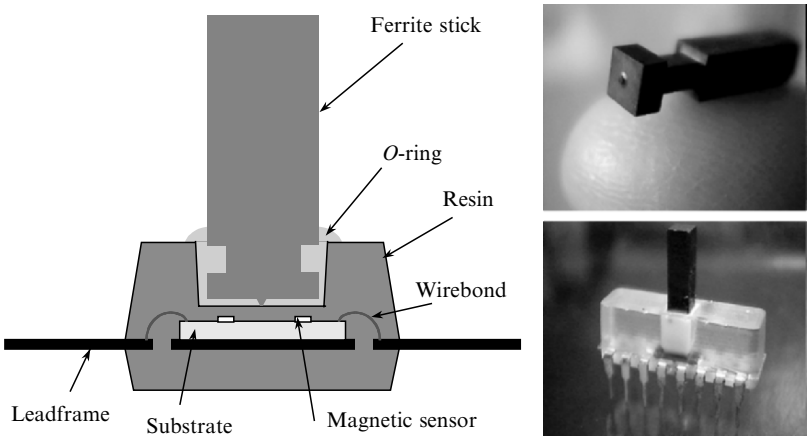


Figure 3.3-7. Integrated track pointer in DIL housing.

tion of the magnetic sensors on a chip allows us to integrate the preprocessing electronics in the same SiP product.

The chip of the integrated trackpointer does not have any free hanging structures like some other mechanical MEMS. Therefore the chip can be cut and handled just like integrated electronic chips. This is less trivial with chips containing free hanging structures. Those structures are damaged easily during sawing and handling of the wafer.

If possible, micromachined devices are packaged on a wafer scale, before separation. Waferscale packaging can be used to fill enclosures with special gasses or to create vacuum chambers [19]. Packaging on wafer scale solves the problem of sawing and handling fragile micromechanical chips.

7. MATERIALS AND PRODUCTION: WHAT ENABLES MST?

The applications of microsystems mentioned above are very diverse with respect to the processes used to realize them. Even two virtually similar surface micromachining processes will differ a lot due to individual constraints on the construction. Without standardization of microsystem processes and packaging, we have to reinvent the basics over and over again. Although batch-processable micromachined devices have the opportunity to become cheap components in all AmI, most applications are low volume products at this moment. Due to the bad uniformity in micromachining silicon processes it is very hard to ramp the applications up to the levels AmI requires. With more standardization we could use multiproject wafers to reduce development time.

However, there are some MEMS technology platforms that are more or less accepted as basic technology, at least within certain fields of application. First is the use of silicon on insulator (SOI) wafers, which are very suitable for surface micromachining. A SOI wafer is preprocessed to have a buried oxide layer and a relatively thick, up to 20 μm, epitaxial silicon layer on top. By etching trenches and holes in the epilayer and subsequently removing parts of the buried oxide layer by wet or dry etching, free hanging structures are created easily. Since the free hanging structures on SOI wafers have considerable mass they are popular for lateral accelerometers [44].

An alternative is the HEXSIL [45, 46] technology that is specifically designed for MEMS devices. With this technology, trenches are etched in a silicon wafer and filled with nickel by electroless plating. A sacrificial oxide layer enables release of the nickel structure to obtain free hanging struc-

tures. High resistive parts are made by using undoped polysilicon instead of the nickel. The silicon wafer acts as a mould and can be reused.

Both bulk-micromachining and surface micromachining processes are offered by foundry services like Europractice [47]. Experiments can be done on multiproject wafers and the design and development process can be completely outsourced.

Substrate transfer technology [48] (STT) is the technology where a SOI wafer, including electronics in the epilayer, is bonded onto a glass substrate. The resistive substrate of the SOI wafer is subsequently removed in order to have electronic circuits on a low loss carrier. This is a very beneficial method for low-loss RF applications and is illustrated in Figure 3.3-8.

A subsequent development of STT is the technology where the adhesion to the glass substrate is made by a polyimide layer with a weak adhesion layer facing the glass. In this case, the stack can be peeled off of the glass substrate and a flexible foil containing the electronics remain.

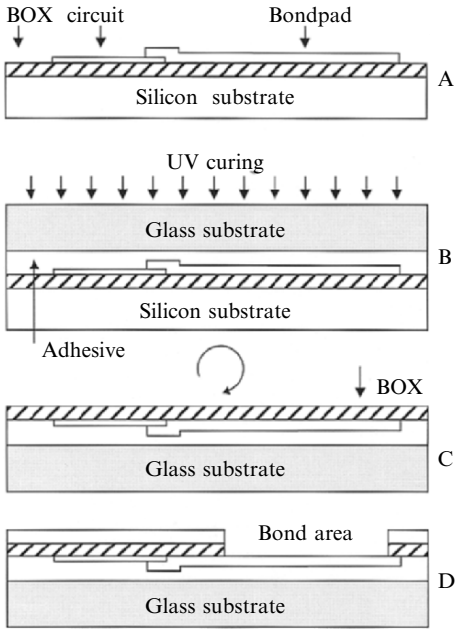


Figure 3.3-8. Schematic cross-section of the transfer of an SOI wafer to a glass substrate. (a) Fully processed SOI wafer. (b) Curing of adhesive by UV exposure through the glass substrate. (c) Removal of the silicon substrate selective toward the buried oxide. (d) Optional deposition of a scratch protection layer and opening of bondpads.

An active RF-tag of size $3 \times 3\,\text{mm}^3$ that can be coiled and bent was demonstrated [49]. This technology, which is not yet optimized for MEMS applications in the sense of micromechanical devices, is extremely valuable for AmI. It combines the opportunity of batch processing with flexible solutions, which can be placed on and in almost any application.

8. MEMS MODELLING

Besides MEMS technology for packaging and realization, successful development depends on the quality of design and test equipment. In this section, a brief overview is given on the method of MEMS modeling.

8.1. Lumped Elements Method

The best method to start calculations on MEMS structures is to use lumped element methods. With lumped element methods, entities of all physical domains are represented in a single formalized network. This yields optimum transparency of transduction principles, where electrical circuit theory can be applied to optimize and calculate performance parameters. The theory is known as *dynamical analogies* or *systems theory* and is as old as the existence of electrodynamic transducers [50–52].

This method will be illustrated using the example of the mass on a cantilever (see Figure 3.3-1). To make it a two-domain physical transducer, an electrostatic actuator is added as shown in Figure 3.3-9. Assume the mass and the cantilevers to be good conductors. The counter electrode is placed at a distance y_0 to create a parallel plate capacitor.

The starting point is the definition of a state variable, a *flow* and an *effort* in each physical domain. The flow is defined as the derivative in time of the state variable and the effort is defined as the cause of the flow. In the electrical domain, the state variable is the charge q. This makes the flow to

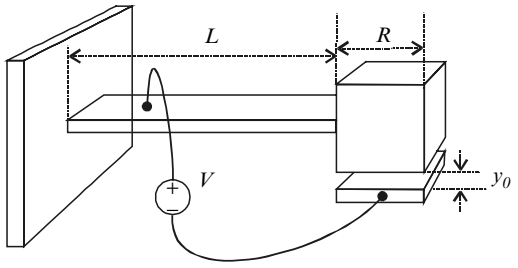


Figure 3.3-9. An electrostatically actuated mass on a spring.

be the electrical current $i = \mathrm{d}q/\mathrm{d}t$ and the potential the effort since it is the cause of an electrical current. In the mechanical domain we may consider translation as the state variable, resulting into force as the flow and speed as the effort[‡].

For a selection of physical domains the state variables, flows, and efforts are given in Table 3.3-2.

In the SI system, the product of effort and flow is the power in Watts. Energy can be stored in either a capacitive or an inertial buffer or be dissipated in a resistive element, satisfying the equations:

$$\text{flow} = C\frac{\mathrm{d}(\text{effort})}{dt},\ \text{effort} = L\frac{\mathrm{d}(\text{flow})}{dt} \text{ or effort} = R \cdot \text{flow} \qquad (3.3\text{-}3)$$

respectively. Besides the one-port elements R, C, and L, there are also two-port elements like transformers and gyrators.

Now we can derive an equivalent circuit for the electrostatically actuated mass on a spring of Figure 3.3-9. The system is driven by an electric potential source $u(t)$. The source has an electrical internal resistance R_e that is loaded by the impedance of the capacitance C_e. This is shown at the electrical side of Figure 3.3-10.

Transduction from the electrical to the mechanical domain is modeled by an ideal transformer. In a certain regime, where the system is biased by a voltage V_Bias, the transduction is linear and satisfies $u = T_\mathrm{EM} \cdot F$ and $v = T_\mathrm{EM} \cdot i$. The transduction coefficient $T_{\mathrm{EM\ equals}} y_0^2/(\varepsilon_0 R^2 V_\mathrm{Bias})$.

On the mechanical side, the compliance of the beam C_M [equal to k^{-1} as given in Equation (3.3-1)], the mass M_M, and a certain mechanical friction with air R_M are incorporated into the model. These components are placed in series since they are exposed to the same velocity v.

From a circuit analysis we can understand the behavior of the system. The voltage source is not only loaded by the impedance due to the electrical parts R_e and C_e, but also by the mechanical parts R_M, C_M, and M_M. This is due to (electromechanical) coupling acting in two directions, back and forth. The idealized part of the coupling is modeled by the ideal transformer. The frequency response can be derived from the model. There will be a resonance due to the mass and the spring at $f_\mathrm{res} = 1/[2\pi\sqrt{(C_\mathrm{M} M_\mathrm{M})}]$. At lower frequencies where the effect of C_e is negligible, the transduction from input voltage $u(\omega)$ to mass velocity $v(\omega)$ is given by

[‡]This is the *impedance* equivalence which opposes the more commonly used *mobility* analogy where it is just the opposite way (speed is *flow* and is force is *effort*) since mechanical impulse is considered as the state variable. These conventions are called *dual* and result in mathematical interchangeability of the associated buffers.

Table 3.3-2. Summary of state variables, signals and components in several physical domains.

Physical domain	State variable	Effort	Flow	Resistive	Inertial buffer	Capacitive buffer
Electrical	Charge $\boldsymbol{q}$ **[C]**	Potential $\boldsymbol{u}$ **[V]**	Current $\boldsymbol{i}$ [*A*]	Resistor $\boldsymbol{R}$ **[Ω]**	Coil $\boldsymbol{L}$ **[H]**	Capacitor $\boldsymbol{C}$ **[F]**
Mechanical	Translation $\boldsymbol{x}$ **[m]**	Force $\boldsymbol{F}$ **[N]**	Velocity $\boldsymbol{v}$ **[m/s]**	Friction **[N·s/m]**	Mass **[kg]**	Compliance **[m/N]**
Acoustical	vol. displ. $\boldsymbol{x}$ **[m]**	Pressure $\boldsymbol{p}$ **[Pa]**	Vol. velocity $\boldsymbol{U}$ **[m³/s]**	Resistor **[Pa · s/m³]**	Mass **[kg/m⁴]**	Compliance **[m³/Pa]**
Magnetic	Flux $\boldsymbol{\Phi}$ **[VS]**	Mmf $\boldsymbol{M}$ **[A]**	Potential $\boldsymbol{u}$ **[V]**			
Thermodynamical	Entropy $\boldsymbol{S}$ **[J/K]**	Temp. $\boldsymbol{T}$ **[K]**	Entr. flow $\boldsymbol{f}_S$ **[J/s · K]**			

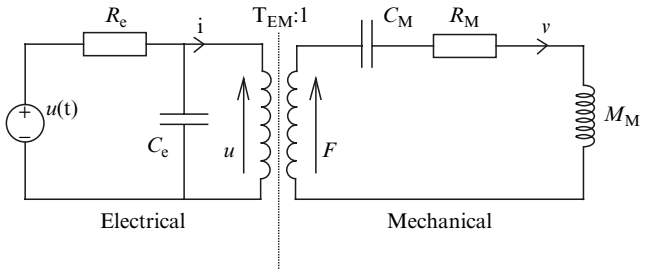


Figure 3.3-10 Equivalent model for the electrostatically actuated mass on a spring.

$$v(\omega) = T_{EM}^{-1}\left(\frac{R_e}{T_{EM}^2} + \frac{1}{j\omega C_M} + R_M + j\omega M_M\right)^{-1} u(\omega) \tag{3.3-4}$$

which shows how the method of lumped elements easily helps us to develop mathematical models.

When the model of Figure 3.3-10 shows too much disagreement with observations, we have to refine the model by adding more components. For example, oscillating modes in the cantilever are not implemented in the model, but could be included by adding *LC* circuits. At any time, the lumped element model is linked phenomenologically to reality.

8.2. Finite Elements Method

The values of the lumped elements are not straightforward to be determined in all cases. This can be due to the fact that an element represents a phenomenon, which cannot be localized intuitively and that it represents a nonlinear effect, or that it is difficult to be represented in a single element. Examples are cases where all the vibrational modes of a plate have to be considered, when the shape of a structure is very complicated or when there is no analytical solution to a certain nonlinear differential equation. In that case modeling must be based on the motional differential equations directly.

The finite element method (FEM) is based on spatial segmentation of a structure where each segment is described by its coupled differential equations. Software packages like ANSYS, COSMOS, CoventorWare, and FEMLAB are available for numerical evaluations and provide graphical representations of the results.

However, this method has several drawbacks. The finite element method is a low-level interpretation of the physical behavior. To combine them into macroscopic overviews of device characteristics, large capacity with respect

to memory and calculation speed is required. It is very hard to maintain the systems overview since the links between physical domains have become invisible. In many cases, the designer has to limit himself to a FEM analysis of only a part, one or two domains, of the problem to limit the complexity of the problem. To understand the relation between design parameters and performance in FEM analysis it has become the skill of the scientist, rather than the power of the tool as we have seen with the lumped element method.

9. SUMMARY AND CONCLUSIONS

Based on the overview of technology presented in this chapter, we can summarize the advantages of microsystems with respect to AmI as six key benefits:

- Microsystems are the small domain interfaces enabling AmI. They are small, will *fit in anything* and act in their position as a reasoning transduction node.
- The devices are realized using batch processing which results in cost reduction and better uniformity. Since this *multiplicity* reduces costs, we can put even more of them in everything.
- New features are facilitated by *scaling* physical phenomena. Think of the static micromixers without moving parts, the ultramicroelectrodes and all the new things enabled by electrowetting.
- Scaling down and batch processing does not necessarily result in poor or fragile devices. On the contrary: MEMS devices especially are *high-end*, consume less power and are robust when designed properly.
- An array of sensor elements and sensor/actuator fusion enables multivariant analysis and *stimulus response measurements*, which is beneficial with respect to deriving new and otherwise inaccessible parameters and increases measurement reliability.
- The SiP approach considers the matching of technologies in order to optimize development time, yield, cost efficiency, and functionality. A sensor has become more than a single sensing element, it comprises signal conditioning and a package suitable for PCB mounting.

Considering these benefits, one might wonder why microsystem technology, especially MEMS sensors, has penetrated only in limited fields of application. The question is what the limiting factors are for the success of micromachined products. A cautious onset for a list of explanations is:

- Silicon foundries have to make structural changes in their fleet of equipment for the relatively new technologies like surface micromachining and microfluidics. Some typical MEMS materials are not appreciated in existing plants, which requires additional structural changes.
- To have a freshly designed MEMS product welcomed at one of the silicon-foundries, persuasive precalculations are needed. However, the MEMS technology platforms are relatively new resulting in ambiguous cost estimations, which hamper the process of acceptation.
- There is no satisfying set of universal microsystem processing technologies. Unification would speed up the design process and enable more flexible use of foundry services.
- Microsystem technology requires multidisciplinary skills from the design, development, industrialization, and manufacturing teams. In the first three teams, dealing with the edge of several physical domains has become common practice. At production sites, however, introducing disciplines besides microelectronics require adaptation of a mindset and the installation of equipment to check standards in the field of the application.

As shown in Section 2, the definition of AmI maps perfectly onto the aim of the SiP approach. From this consideration we may conclude that microsystem technology, besides CMOS technology, is the most important enabler for AmI. However, it is fair enough to say that AmI is a challenge for microsystem technology. This appears to be a more reasonable view since at this moment the development and acceptance of AmI is partially limited by the limited availability of cheap sensor and transducer technologies. Nevertheless, we must realize that microsystem technology, like MEMS, is not the goal but is the method to design the consumer products of the future.

ACKNOWLEDGEMENTS

Special thanks Jaap Ruigrok and John Mills for reviewing this chapter. In addition, I would like to acknowledge Joost van Beek, Stein Kuiper, Johan Feenstra, Fred Roozeboom, and Ronald Dekker for reviewing the specific parts on their projects.

REFERENCES

[1] Feynman, R. P., 1992, There's plenty of room at the bottom (1959), *J. Microelectromech. Syst.*, **1**, 60–66.

[2] Feynman, R. P., 1993, Infinitesimal machinery (1983), *J. Microelectromech. Syst*., **2**(1), 4–14.
[3] Cambridge Intenational Dictionary of English, 1995, Cambridge University Press.
[4] *Microcosmos*, A Jacques Perrin film by Claude Nuridsany and Marie Perennou (Galatée Films, 1996)
[5] Madou, M., 2002, Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edition, CRC Press, Boca Raton.
[6] Gere, J. M. and Timoshenko, S. P., 1990, Mechanics of Materials, 3rd edition, PWS-KENT Publishing Company, Elsevier Science BV, Amsterdam.
[7] Trimmer, W. S. N., 1989, Microrobots and micromechanical systems, *Sensors and Actuators*, **19**, 267–287.
[8] Möbius, H., Ehrfeld, W., Hessel V. and Richter, Th., 1995, Sensor controlled processes in chemical microreactors, *The 8th International Conference on Solid-State Sensors and Actuators*, *Transducers '95 and Eurosensors IX*, Stockholm, Sweden, June 25–29, 1995, pp. 775–778.
[9] Neagu, C., 1998, A medical microactuator based on an electrochemical principle, Thesis, University of Twente.
[10] Morf, W. E. and de Rooij, N. F., 1996, Performance of amperometric sensors based on multiple microelectrode arrays, *Eurosensors X*, Leuven, Belgium, September 8–11, 1996.
[11] Roylance, L. M. and Angell, J. B., 1979, A batch-fabricated silicon accelerometer, *IEEE Trans. on Electron Devices*, ED-26, 1911–1917.
[12] Petersen, K. E., Shartel, A. and Raley, N. F., 1982, Micromechanical accelerometer integrated with MOS detection circuitry, *IEEE Trans. Electron Devices*, ED-29, 23–27.
[13] Analog Devices, ADXL50: Monolithic accelerometer with signal conditioning, 1996 obsolete datasheet 2044696ADXL50.pdf, www.analog.com.
[14] Bassous, E. and Baran, E. F., 1978, The fabrication of high precision nozzles by the anisotropic etching of (100) silicon, *J. Electrochem. Soc*., **125**, 132.
[15] In-Stat/MDR market overview report IN030601EA, Got MEMS? 2003 Industry overview and forecast.
[16] Yao, J. J., 2000, RF-MEMS from a device perspective, *J. Micromech. and Microeng*., **10**, R9–R38.
[17] van Beek, J. T. M., et al., 2004, High-Q integrated RF passives and RF-MEMS on silicon: Materials, integration and packaging issues for high-frequency devices, *Symp*. Boston 2003, *in* P. Muralt, et al. (eds), *MRS Fall Meeting*, Warrendale, MRS Materials Research Soc., **783**, 97–108.
[18] Roozeboom, F., Kemmeren, A., Verhoeven, J., van den Heuvel, F., Kretschman, H. and Frič, T., 2003, High density, low-loss MOS decoupling capacitors integrated in a GSM power amplifier, *Procs. Mat. Res. Soc. Symp*., **783**, 157–162.
[19] Rijks, T. G. S., et al., 2004, MEMS tunable capacitors and switches for RF applications, MIEL 2004, Nis, Serbia, *24th Int. Conf. Microelectronics*, Conference Proceedings.

[20] Middelhoek, S. and Audet, S., 1989, *Silicon Sensors, Microelectronics and Signal Processing*, Academic Press, London.
[21] Bergveld, P., 2003, Thirty years of ISFETOLOGY: What happened in the past 30 years and what may happen in the next 30 years, *Sensors and Actuators*, **B88**, 1–20.
[22] van der Schoot, B. H. and Bergveld, P., 1985, An ISFET-based microlitre titrator: Integration of a chemical sensor-actuator system, *Sensors and Actuators*, **8**, 143–151.
[23] Olthuis, W., Langereis, G. and Bergveld, P., 2001, The merits of differential measuring in time and space, *Biocybernetics and Biomedical Engineering*, **21**(3), 5–26.
[24] Janata, J., Josowicz, M., Vanýsek, P. and DeVaney, D. M., 1998, Chemical sensors, *Anal. Chem*., **70**, 179R–208R.
[25] Manz, A., Graber, N. and Widmer, H. M., 1990, Miniaturized total chemical analysis systems: A novel concept for chemical sensing, *Sensors and Actuators*, **B1**, 244–248.
[26] Laurell, T. and Drott, J., 1995, Silicon wafer integrated enzyme reactors, *Biosensors and Bioelectronics*, **10**, 289–299.
[27] Elwenspoek, M., Lammerink, T. S. J., Miyake, R. and Fluitman, J. H. J., 1994, Towards integrated microliquid handling systems, *J. Micromech. Microeng*., **4**, 227–245.
[28] Harrison, D. J., Manz, A., and Glavina, P. G., 1991, Electroosmotic pumping within a chemical sensor system integrated on silicon, *International Conference on Solid-State Sensors and Actuators*, *1991*, *Transducers '91*, June 24–27, 1991, pp. 792–795.
[29] Olthuis, W., Böhm, S., Langereis, G. R. and Bergveld, P., 2000, Selection in system and sensor, *in* A. Mulchandani and O. A. Sadik (eds), *Chemical and Biological Sensors for Environmental Monitoring*, *ACS Symposium Series 762*, Am. Chem. Soc., Chapter 5, pp. 60–85, Oxford University Press, Washington, DC.
[30] http://www.dlp.com, DLP: A Texas Instruments technology.
[31] Paik, P. Pamula, V. K. and Fair, R. B., Rapid droplet mixers for digital microfluidic systems, *Lab on a Chip*, **3**, 253–259.
[32] http://www.ece.duke.edu/Research/microfluidics, Digital microfluidics, Duke University, Durham, NC.
[33] Kuiper, S. and Hendriks, B. H. W., 2004, Variable-focus liquid lens for miniature cameras, *Appl. Phys. letters*, **85**(7), 1128–1130.
[34] Hayes, R. A. and Feenstra, B. J., 2003, Video speed electronic paper based on electrowetting, *Nature*, **425**(25), 383–385.
[35] Special issue on biomedical applications for MEMS and microfluidics, 2004, *Proc. IEEE*, **92**(1), 1–184.
[36] Pedersen, M., Olthuis, W. and Bergveld, P., 1998, An Integrated silicon capacitive microphone with frequency-modulated digital output, *Sensors and Actuators*, **A69**, 267–275.

[37] Neumann, Jr., J. J. and Gabriel, K. J., 2002, CMOSMEMS membrane for audio-frequency acoustic actuation, *Sensors and Actuators*, **A95**, 175–182.
[38] Klein, U., Müllenborn, M. and Rombach, P., The advent of silicon microphones in high-volume applications, *MSTnews* 02/1, pp. 40–41.
[39] Dittmar, A., et al., 1995, A multi-sensor system for the non invasive measurement of the activity of the autonomic nervous system, *Sensors and Actuators*, **B27**, 461–464.
[40] Langereis, G. R., Olthuis, W. and Bergveld, P., 1999, Using a single structure for three sensor operations and two actuator operations, *Sensors and Actuators*, **B53**, 197–203.
[41] Middelhoek, S., 2000, Celebration of the tenth transducer conference: The past, present and future of transducer research and development, *Sensors and Actuators*, **A82**, 2–23.
[42] Hippert, M. A., 2004, Board to board interconnects: Innovative new possibilities with MID packages, *Procs. 6th International Congress Molded Interconnect Devices*, *MID 2004*, Erlangen, Germany, September 22–23, 2004, pp. 41–47.
[43] Peels, W., van Montfoort, V., Verweg, F. and Weekamp, W., 2004, Integrated display modules in I^2MC-technology: 3D-MID/over-moulding of electronics, *Procs. 6th International Congress Molded Interconnect Devices*, *MID 2004*, Erlangen, Germany, September 22–23, 2004, pp. 49–56.
[44] Yazdi, N., et al., 1998, Micromachined inertial sensors, *Procs. IEEE*, pp. 1640–1659.
[45] http://www.memspi.com, MEMS Precision Instruments.
[46] Keller, C. G. and Howe, R. T., 1995, Nickel-filled Hexsil thermally actuated tweezers, *8th International Conference on Solid-state Sensors and Actuators (Transducers '95)*, June 1995, Stockholm, Sweden, pp. 99–102.
[47] http://www.europractice.bosch.com/en/foundry/index.htm, Europractice Surface Micromachining foundry.
[48] Dekker, R., Baltus, P. G. M. and Maas, H. G. R., 2003, Substrate transfer for RF technologies, *IEEE Trans. on Electron Dev.*, **50**(3), 747–757.
[49] Dekker, R., et al., 2003, Substrate transfer: Enabling technology for RF applications, *Electron Devices Meeting*, *2003*, *IEDM '03 Technical Digest*, *IEEE International*, December 8–10, 2003, pp. 15.4.1–15.4.4.
[50] Firestone, F. A., 1993, A new analogy between mechanical and electrical systems, *J. Acoust. Soc. Amer.*, **4**, 249–267.
[51] Firestone, F. A., 1938, The mobility method for computing the vibrations of linear mechanical and acoustical systems: Mechanical-electrical analogies, *J. Appl. Phys.*, **9**, 373–387.
[52] Beranek, L. L., 1954, *Acoustics*, McGraw-Hill, New York.

Section 4

Low Power Electronics and System Architecture

Chapter 4.1

LOW ENERGY DIGITAL CIRCUIT DESIGN

Benton H. Calhoun
University of Virginia
bcalhoun@virginia.edu

Curt Schurgers
University of California, San Diego
curts@ece.ucsd.edu

Alice Wang
Texas Instruments
aliwang@ti.com

Anantha Chandrakasan
Massachusetts Institute of Technology
anantha@mtl.mit.edu

Abstract Most ambient intelligence (AmI) applications rely to some degree on digital processing. Since many AmI applications also are energy-constrained, reducing energy consumption in digital blocks lengthens the lifetimes of these systems and helps to enable self-sustained energy harvesting from the environment. This chapter examines algorithmic, architectural, and circuit design techniques for reducing the energy consumption of digital blocks. Variations to system architecture or to algorithms themselves can provide lower energy implementations of a given basic specification, as we illustrate with an example of turbo decoding. Likewise, flexible architectures that adjust to variable requirements improve the energy efficiency of a system. We present an energy-efficient design for an FFT processor. At the circuit level, sub-threshold operation has proven to minimize energy in digital systems. We show an analytical solution for the optimum V_{DD} to minimize energy for a given block and describe the key parameters that affect the minimum energy solution. We also examine sizing issues for sub-threshold circuits.

Key words low-power design; minimum energy; energy-aware; sub-threshold operation

S. Mukherjee et al. (eds.), AmIware, 181–202.

1. INTRODUCTION

Digital blocks are critical to ambient intelligence (AmI) applications, and CMOS technology scaling has made power a primary concern in modern digital circuits. For portable, battery-operated systems like those common to AmI, energy and power often are the driving constraints. Microsensors are one application that is typical of AmI systems. A microsensor node operates from a stored energy source like a battery. Since its energy supply is limited, efficient approaches to operation are required to ensure adequate lifetime for the node. Figure 4.1-1 shows an example architecture for a microsensor node. The basic function of the node is receiving information from its environment through the sensor subsystem, processing the data using different digital processing engines, and communicating the data to other nodes using the radio subsystem. The figure clearly shows that the node consists of multiple blocks that require optimization for energy efficiency on the individual level and collectively [1]. This chapter describes techniques for reducing the energy of digital blocks that cross the vertical hierarchical layers of design.

First, reassessing digital processing algorithms with attention to the hardware that must implement them can produce lower energy solutions. Section 2 demonstrates that exploring different algorithmic variations of the same specification can produce energy savings. This is illustrated for the example case of a turbo decoder. Secondly, cleverly designed digital architectures can take advantage of changing user requirements intrinsic to a given application to save energy. This ability to adapt energy consumption to different operating conditions is called energy awareness.

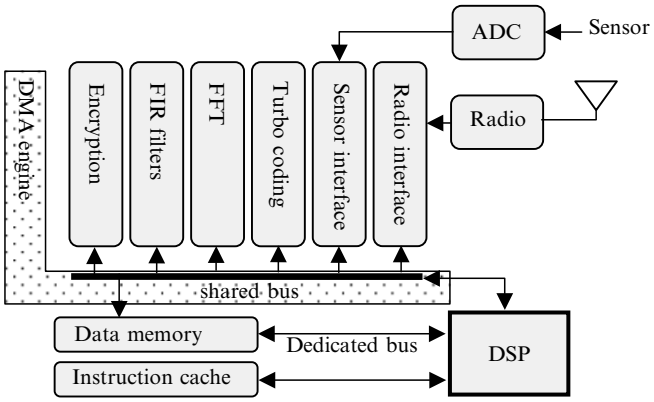


Figure 4.1–1. Example architecture of a microsensor node.

Section 3 applies the concept of energy-aware design to a real-valued FFT implementation. Finally, at the circuit level, sub-threshold operation allows digital circuits to perform their calculations with the minimum energy per operation. Section 4 explains the dependencies of the minimum energy operating point for generic digital circuits operating in the sub-threshold region. Building from well-known equations for MOSFETs in the sub-threshold region, we demonstrate a model for total energy consumption of a digital block. We derive from the model a solution for the optimum V_{DD} to minimize energy. We also examine both theoretically optimum sizing for energy minimization and standard cell performance relative to the optimum.

2. ALGORITHMIC AND ARCHITECTURAL TECHNIQUES

Optimizing the high-level architecture of digital hardware is an excellent technique for reducing power. Often there are multiple hardware structures or architectures, capable of implementing the same basic algorithm. A well-known example is the use of various adder structures, such as ripple-carry or carry-save, which essentially are different architectural variants to add two numbers. Often, the line between architectural variations of the same algorithm and modifications to the algorithm directly is rather vague. One could, for example, argue that the carry–save adder is an algorithmic variant to calculate an addition. Whether we wish to classify our technique as an architectural or algorithmic optimization, the bottom line is that the hardware designer usually is provided with a behavioral specification and has the freedom to optimize the algorithmic/architectural organization that implements this specification. This flexibility is not limited to hardware building blocks, such as adders, but extends to the level of subsystems as well. When implementing a design, which is often given as an algorithmic specification, investigating various algorithmic alternatives and architectural variants can lead to very significant energy savings, which are above and beyond those achievable by circuit and device level design techniques. This principle of algorithmic/architectural optimization is explored here for the example of a turbo decoder implementation.

2.1. Turbo Coding and Decoding

In communication systems, channel coding is used to correct errors that are introduced during transmission. One of the most powerful such coding schemes belongs to the class of turbo coding [2]. When considering

energy efficiency, the main bottleneck is the decoder part of the system [3]. This decoder resides at the receiver-side of the communication link, and processes the received data to eliminate the errors that might have been introduced during transmission. Several variants of turbo coding exist. A common one is known as parallel concatenated convolutional codes, which can be decoded with the maximum a posteriori (MAP) algorithm [2]. The case study in this section focuses on the optimization of this important algorithm.

The MAP decoding for turbo codes is iterative in nature. One of most powerful ways to reduce the energy consumption is by limiting the number of decoding iterations by means of a stop criterion, ending the iteration process when all errors have been corrected [4, 5]. It is possible to further reduce the energy consumption of a single iteration by appropriate algorithmic transformations.

At an abstract level, one iteration can be decomposed into a three-step process. For a data block of length N, the goal is to calculate the likelihood of each bit in the block being equal to $+1$ or -1. While encoding the data at the transmitter side, the encoder goes through a sequence of states. Each encoding of a bit of the block causes a state transition. The number of states, S, is a parameter that also describes the strength of the code (where a higher S corresponds to a more powerful turbo code). The implementation complexity also increases when S goes up. The decoder estimates the likelihood of each received bit being equal to $+1$ or -1, by calculating the likelihood that the encoder made a state transition corresponding to such a $+1$ or -1 bit. The three-step decoding process is described below. More details can be found in the literature [2, 3, 6].

(1) **Calculation of α metrics**: There are $S\,\alpha$ metrics $\alpha_k(s)(s = 1 \ldots S)$ for each bit k of the block ($k = 1 \ldots N$). Each $\alpha_k(s)$ gives the likelihood that the encoder was in state s before bit k was encoded. This information can be calculated with a forward recursion (i.e., all $\alpha_k(s)$ for bit k can be obtained from those for bit $k-1$). Formally, the recursion for the MAX-LOG-MAP [6, 7], which is a particular instantiation of the MAP algorithm, is shown in Equation (1). Here, $\gamma_{k-1}(s', s)$ is the probability of a state transition from state s' for bit $k-1$ to state s for bit k. The max-operation selects the most likely event from all achievable possibilities. Most often, there are two possibilities to get to a state s, and the max-operation is therefore between two values only [6]. This principle can be extended directly towards more general scenarios.

$$\alpha_k(s) = \max_{s' \in S} \left(\alpha_{k-1}(s') + \gamma_{k-1}(s', s)\right) \tag{1}$$

(2) **Calculation of β metrics**: Similar to the α metrics, there are also $S\,\beta$ metrics $\beta_k(s)$ ($s = 1 \ldots S$) for each bit k of the block ($k = 1 \ldots N$). Each $\beta_k(s)$ gives the likelihood that the encoder was in state s after bit k was encoded. This information can be calculated with a backward recursion (i.e., all $\beta_k(s)$ for bit k can be obtained from those for bit $k+1$). This recursion for the MAX-LOG-MAP is shown in Equation (2). The β recursion is exactly the same as the α recursion; only the direction is changed from forward to backward.

$$\beta_k(s) = \max_{s' \in S} \left(\beta_{k+1}(s') + \gamma_k(s, s')\right) \tag{2}$$

(3) **Calculation of $L(u_k)$**: The final step combines the α and β metrics to find the normalized likelihood $L(u_k)$, and that bit k is equal to $+1$ instead of -1. This $L(u_k)$ is equal to the difference of Q_k^1 and Q_k^0, where Q_k^i ($i = 0$ or 1) is calculated from Equation (3). The max-operation is over all state transitions from s' to s that correspond to a bit being equal to i. In the case considered here, there are S such transitions for both $i = 0$ and $i = 1$.

$$Q_k^i = \max_{u_k = i} \left[\alpha_k(s') + \gamma_k(s', s) + \beta_{k+1}(s)\right] \tag{3}$$

2.2. Algorithmic/Architectural Optimizations

The three-step process described above constitutes the behavioral algorithmic specification of the desired decoder hardware. In principle, the forward α recursion and backward β recursion need to be executed for the entire block before the calculation of $L(u_k)$ can proceed. This straightforward implementation will be referred to as *option 1*. It would require storing all S α and β metrics for the entire block of length N. If each such metric is quantized with w bits, this requires two memories of $w{\cdot}S{\cdot}N$ bits each [4]. For most typical turbo coders ($w = 8 \ldots 13$, $S = 4 \ldots 16$, and $N = 400 \ldots 32{,}000$ or more), this is prohibitively large. However, one of these memories can easily be eliminated by observing that $L(u_k)$ can be calculated together with one of the metric recursions (*option 2*). For example, α metrics are first calculated and stored. Next, β metrics are calculated, and since the α metrics are already available, all the information is there to generate $L(u_k)$. The forward recursion of β metrics can then proceed without the need for storing these values, since they can be consumed directly.

A further reduction of memory size is possible by a slight algorithmic modification. It has been realized that it is possible to initialize the β metrics at a random point k in the block. If the recursion is then run for L steps, the metrics converge towards the ones that would have resulted from starting the recursion at the end of the block. This L is called the sliding window length and a good value can be found experimentally to be $5 \cdot [\log_2(S) + 1]$ [3, 5]. After a random initialization, the β recursion is executed for L steps to ensure conversion, but the metrics do not need to be stored since they will never be used. Once conversion is achieved, valid β metrics are calculated and stored. Figure 4.1–2(a) shows in a graphical way how this approach, called the sliding windows, works (*option 3*). The horizontal axis indicates the time progression of the algorithm. The vertical axis denotes the bit index k, for which the metrics or $L(u_k)$ are calculated. This approach further reduces the memory requirements to $w \cdot S \cdot L$ bits at the cost of extra calculations (extra beta metrics are calculated to ensure convergence).

A more intricate modification of the algorithm uses the notion of traceback (*option 4*) [7]. During the forward α recursion, it is possible to store which of the two options was chosen in the max-operation of Equation (1). This information, called the traceback information, can be used to simplify the calculation of $L(u_k)$. Specifically, for each bit k, either Q_k^1 or Q_k^0 can be simplified to just an addition of three numbers. The corresponding algorithmic representation is shown in Figure 4.1–2 (b). At the start of each window, the traceback operation is initialized based on the available α and β metrics [7].

Another level of flexibility available to the designer is varying the number of $L(u_k)$ that are calculated after each β recursion of L steps (which was needed to ensure conversion of the sliding window approach) [3]. Figure 4.1–3 shows different alternatives that are characterized by one parameter η. The number of $L(u_k)$ calculated in each window is equal to

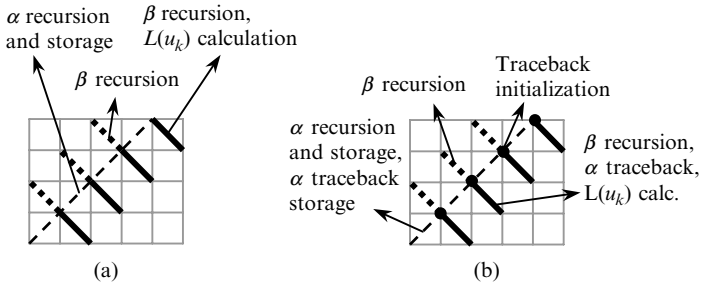


Figure 4.1–2. Architectural alternatives for trace-back enhancement

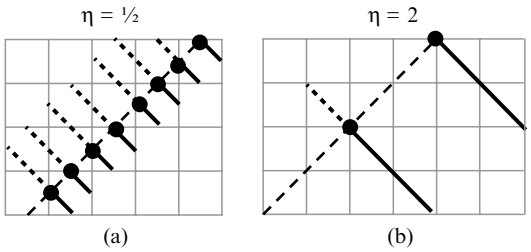


Figure 4.1–3. Architectural parameter η

$\eta \cdot L$. This technique resides on the architectural level. Its benefit is a reduction of β metric calculations at the expense of increased storage for the traceback and the α metrics.

Finally, it is also possible to start the recursions concurrently at both sides of the block to improve the throughput [3]. This architectural variant is shown in Figure 4.1–4(a) (*option 5*). By offsetting the two recursions as shown in Figure 4.1–4(b) (*option 6*), it is possible to reduce the required memory compared to the option in Figure 4.1 (a).

Table 4.1-1 compares all six possible architectural/algorithmic options. The reported number of operations is in terms of number of additions, where it is assumed a max-operation and a traceback operation are equally complex as an addition. The memory size in bits accounts for the metric memory and the storage required for the traceback information. For both the number of operations and the memory size, the first subcolumn lists the parametric expressions [3, 7], and the second subcolumn gives numeric values for an example design. For this example design, $N = 1000$, $S = 8$, $L = 16$, and $w = 10$. Although η is a parameter that is under the designer's control, it was set equal to 1 in this case. From the parametric expressions, it can be seen that by varying η, the number of operations

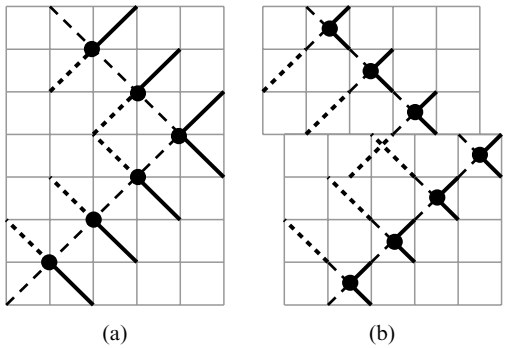


Figure 4.1–4. Double flow structures.

Table 4.1–1. Comparison of the various alternatives.

	# Operations		Memory size (bits)	
Option 1	$12{\cdot}S + 2$	98	$2{\cdot}w{\cdot}S{\cdot}N$	$160 \cdot 10^3$
Option 2	$12{\cdot}S + 2$	98	$w{\cdot}S{\cdot}N$	$80 \cdot 10^3$
Option 3	$[12 + 3/\eta]{\cdot}S + 2$	122	$[w{\cdot}S{\cdot}L]{\cdot}\eta$	1280
Option 4	$[9 + (3 + 2/\mathrm{L})/\eta]{\cdot}S + 6$	103	$[w{\cdot}S{\cdot}L + S{\cdot}L]{\cdot}\eta$	1408
Option 5	$[9 + (3 + 2/\mathrm{L})/\eta]{\cdot}S + 6$	103	$2{\cdot}[w{\cdot}S{\cdot}L + S{\cdot}L]{\cdot}\eta$	2816
Option 6	$[9 + (3 + 2/\mathrm{L})/\eta]{\cdot}S + 6$	103	$[w{\cdot}S{\cdot}L + S{\cdot}L] \cdot \eta$	1408

(and thus the energy thereof), can be traded off with memory size. Note that the number of memory accesses, which is not listed here, also figures into the overall energy consumption [3].

From Table 4.1–1, it can be observed that some algorithmic/architectural optimizations are fundamental improvements (i.e., they reduce both the number of operations and the memory size). Most of the time, however, there is a tradeoff between the two. In typical practical systems, options 1 and 2 are prohibitively large in terms of memory area. Of the remaining ones, option 6 is preferable in terms of energy consumption. Compared to option 4, it behaves similar, but has a decoding delay that is a factor of two smaller. It is interesting to see that this enables the designer to recuperate some of the speed that was lost by lowering the supply voltage of the circuit. For a target throughput, option 6 will therefore be more energy efficient as it is able to operate with a smaller supply voltage. This illustrates that the algorithmic/architectural techniques go hand-in-hand with and complement optimizations on the circuit and device level. Not only do they all contribute to the overall energy efficiency of the system, but tradeoffs at one level of abstraction can be exploited at the other to further reduce energy consumption.

3. ENERGY-AWARE ARCHITECTURES

At the architectural level, designing for energy awareness can allow a sensor node to minimize energy consumption in the variable environment of a microsensor network. Energy-aware design is in contrast to low power design, which targets the worst case scenario and may not be globally optimal for systems with varying conditions. The energy awareness of a system can be increased by adding additional hardware to cover functionality over many scenarios of interest and to tune the hardware so that the system is energy-efficient over a range of scenarios. Energy efficiency reduces the average energy per operation under varying performance

requirements and thus relaxes the energy storage requirement for the microsensor node. This section provides an example of an energy-aware implementation of the widely used FFT algorithm.

The FFT calculates the frequency content of time-domain data. It is used in frequency domain beamforming, source tracking, harmonic line association, and classification. To achieve energy awareness, the FFT implementation includes tunable structures, such as memory size and variable bit precision, to handle a variety of scenarios efficiently. This example design implements a real-valued FFT (RVFFT) that scales between 128- and 1024-point FFT lengths and operates at both 8- and 16-bit precision. The Baugh Wooley (BW) multiplier design provides an example of an energy-scalable bit precision datapath in Figure 4.1-5. The RVFFT uses four BW multipliers to perform complex multiplication.

When a 16-bit multiplication is needed, the entire multiplier is used. However, if an 8-bit multiplication is needed, only the MSB quadrant of adders is required. In this case, the 8-bit inputs feed directly to the MSB quadrant, and the LSB inputs are gated to eliminate switching in the unused adders.

Variable FFT length is another hook designed into the FFT processor. The control logic to the FFT scales the number of butterflies with FFT length. The processor stops early to save energy for smaller FFT lengths.

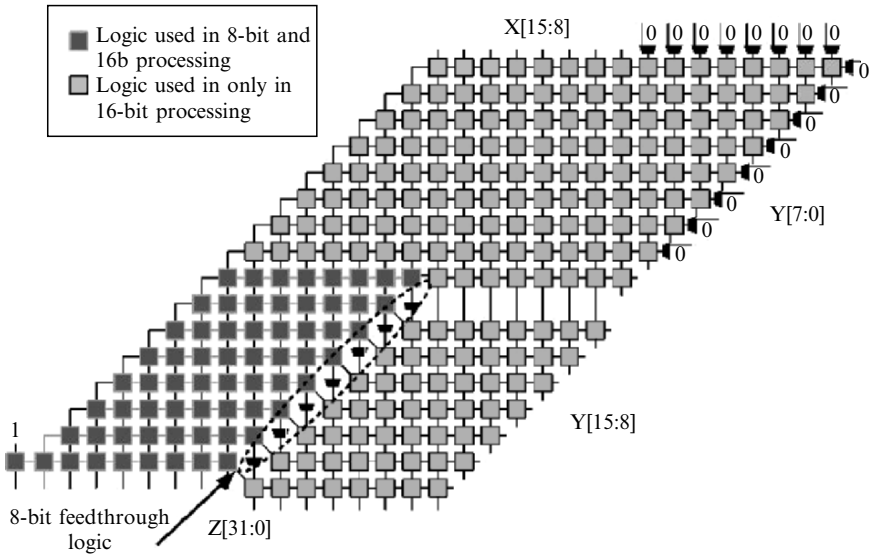


Figure 4.1–5. The 8-bit and 16-bit scalable Baugh Wooley multiplier architecture. The 8-bit multiplier is reused for the 16-bit multiplication, thereby adding scalability without a large area penalty.

Also, the dedicated memory is designed to scale the memory size with FFT lengths. For example, 128-point FFT processing only requires a 128W $\times$ 32b memory. Therefore using a nonscalable memory designed for the 1024-point FFT dissipates additional energy overhead for 128-point processing. A scalable memory that uses the correct memory size for 128-point processing and the entire memory for 1024-point processing is shown in Figure 4.1-6.

The energy scalable FFT architecture was simulated in a 0.18-μm CMOS process at 1.5-V operation, and the simulated energy dissipated is shown in Table 4.1-2.

The simulation results show a definite advantage for an energy-scalable architecture over a nonscalable architecture. The scalable architecture is more energy efficient for all but the high quality point (1024-point, 16-bit). At the high quality point, the scalable design sees a disadvantage due to the overhead logic. However, the scalable implementation uses 2.7 times lower energy at the low quality point (128-point, 8-bit). The scalable FFT processor was fabricated in a standard 0.18-μm CMOS process and standard ASIC flow to demonstrate these energy-scalable architectural techniques. At 1.5-V operation, when compared to a StrongARM SA-1100 implementation, the FFT processor shows over a 350X measured energy reduction. This result is evidence of the significant energy savings that can be achieved by using dedicated hardware modules.

Energy-scalable architectures are designed with many hooks that allow the processor to gracefully scale energy with quality and to achieve global

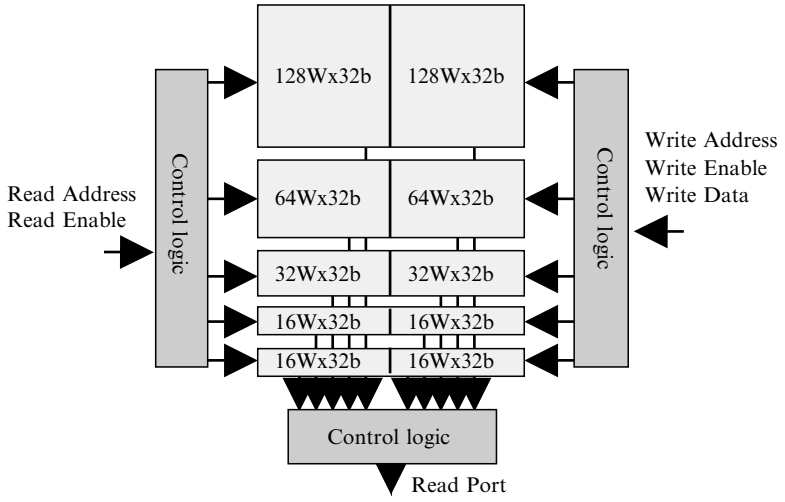


Figure 4.1–6. Scalable FFT memory that enables variable memory size.

Table 4.1–2. Comparing energy/FFT for a nonscalable RVFFT to the scalable RVFFT method.

FFT length	Nonscalable 8-bit	Nonscalable 16-bit	Scalable 8-bit	Scalable 16-bit
1024-point	1320nJ	1448nJ	575nJ	1491nJ
512-point	607nJ	750nJ	240nJ	629nJ
256-point	269nJ	334nJ	103nJ	269nJ
128-point	118nJ	147nJ	44nJ	116nJ

energy efficiency. These techniques enabled variable bit precision and variable FFT lengths in an FFT processor and increased the energy awareness of the system with minimal area and energy overhead.

4. SUB-THRESHOLD OPERATION FOR MINIMUM ENERGY

Emerging applications, such as distributed sensor networks or medical applications have low energy as the primary concern instead of performance, with the eventual goal of harvesting energy from the environment (e.g. [8]). Minimum energy operation for low performance situations occurs in the sub-threshold region [9, 10]. We explain a sub-threshold energy model that shows the dependence of the minimum energy point on design characteristics and operating conditions. The model also provides an analytical solution for the optimum supply voltage (V_{DD}) required for a given circuit to function at its minimum energy point. We also examine the effect of device sizing on minimum energy operation. After considering theoretically optimal sizing, we evaluate standard cell designs for minimum energy operation. A fabricated 0.18-μm test chip provides measurements for analysis.

Sub-threshold operation (where $V_T > V_{DD}$) is currently used for some low power applications such as watches [11] and hearing aids [12]. Emerging ultralow power applications, such as distributed sensor networks, are a natural fit with sub-threshold circuits. Special circuit techniques for improving robustness in deep sub-threshold have been explored [13, 14]. Examining the energy-delay contours over V_{DD} and V_T shows that minimum energy operation occurs in the sub-threshold operation regime for low-to-medium performance systems, and the optimum point changes depending on activity factor and threshold variation [9].

4.1. Modeling for Sub-threshold Operation

This section develops the models for sub-threshold energy analysis that give a closed form solution for the optimum V_{DD} and V_T to minimize energy for a given frequency and technology operating in the sub-threshold regime. The basic equation for modeling sub-threshold current, which comprises the total *off* current in the sub-threshold region because other leakage components are negligible, is given in Equation (4):

$$I_{\mathrm{SUB}} = I_o e^{\frac{V_{\mathrm{GS}} - V_T + \eta V_{\mathrm{DS}}}{n V_{th}}} \left(1 - e^{-\frac{V_{\mathrm{DS}}}{V_{th}}}\right),$$

where

$$I_o = \mu_o C_{ox} \frac{W}{L} (n-1) V_{th}^2 \tag{4}$$

where n is the sub-threshold slope factor $(1 + C_d/C_{ox})$, and V_{th} is kT/q. In order to expand this device equation into a model for generic circuits, we use fitting parameters that are normalized to a characteristic inverter in the technology of interest. Equation (5) shows the propagation delay of a characteristic inverter with output capacitance C_g in sub-threshold:

$$\frac{t_d = K C_g V_{\mathrm{DD}}}{I_{o,g} e^{\frac{V_{\mathrm{GS}} - V_{T,g}}{n V_{th}}}} \tag{5}$$

where K is a delay fitting parameter. The expression for current in the denominator of Equation (5) models the *on* current of the characteristic inverter, so it accounts for transitions through both NMOS and PMOS devices. Unless the PMOS and NMOS are perfectly symmetrical, the terms $I_{o,g}$ and $V_{T,g}$ are fitted parameters that do not correspond exactly with the MOSFET parameters of the same name. Operational frequency is simply $f = (t_d L_{DP})^{-1}$, where L_{DP} is the depth of the critical path in characteristic inverter delays. Dynamic (E_{DYN}), leakage (E_L), and total energy (E_T) per cycle are expressed in [6–8, 16], assuming rail-to-rail swing ($V_{GS} = V_{\mathrm{DD}}$ for *on* current).

$$E_{\mathrm{DYN}} = C_{\mathrm{eff}} V_{\mathrm{DD}}^2 \tag{6}$$

$$E_L = W_{\text{eff}} I_{o,g} e^{-\frac{V_{T,g}}{nV_{th}}} t_d L_{\text{DP}} V_{\text{DD}} = W_{\text{eff}} K C_g L_{\text{DP}} V_{\text{DD}}^2 e^{-\frac{V_{\text{DD}}}{nV_{th}}} \tag{7}$$

$$E_{\text{Total}} = C_{\text{eff}} V_{\text{DD}}^2 + W_{\text{eff}} L_{\text{DP}} K C_g V_{\text{DD}}^2 e^{-\frac{V_{\text{DD}}}{nV_{th}}} = V_{\text{DD}}^2 \left(C_{\text{eff}} + W_{\text{eff}} K C_g L_{\text{DP}} e^{-\frac{V_{\text{DD}}}{nV_{th}}} \right) \tag{8}$$

Equations (6–8) extend the expressions for current and delay of an inverter to arbitrary larger circuits. This extension sacrifices accuracy for simplicity since the fitted parameters cannot account for all of the details of every circuit. Thus, C_{eff} is the average effective switched capacitance of the entire circuit, including the average activity factor over all of its nodes, short circuit current, glitching effects, etc. Likewise, W_{eff} estimates the average total width, relative to the characteristic inverter, that contributes to leakage current. Treating this parameter as a constant ignores the state dependence of leakage. Solving this set of equations provides a good estimate of the optimum for the average case and shows how the optimum point depends on the major parameters. To calibrate the model, C_{eff} is estimated by measuring average supply current for an average simulation and solving the expression $C_{\text{eff}} = I_{avg}(fV_{\text{DD}})^{-1}$. Simulating to exercise the circuit's critical path provides the logic depth, L_{DP}. Lastly, W_{eff} is determined by simulating the circuit's steady-state leakage current and normalizing to the characteristic inverter. Since W_{eff} is a function of circuit state, averaging the circuit leakage current for simulations over many states improves the total leakage estimate.

Differentiating Equation (8) and equating to 0 allows us to solve for V_{DDopt}:

$$\frac{\partial E_{\text{TOTAL}}}{\partial V_{\text{DD}}} = 2C_{\text{eff}} V_{\text{DD}} + 2W_{\text{eff}} L_{\text{DP}} K C_g V_{\text{DD}} e^{-\frac{V_{\text{DD}}}{nV_{th}}} + \frac{-W_{\text{eff}} L_{\text{DP}} K C_g V_{\text{DD}}^2}{nV_{th}} e^{-\frac{V_{\text{DD}}}{nV_{th}}} = 0 \tag{9}$$

The analytical solution for V_{DDopt} is in Equation (10), with the constraint in Equation (11):

$$V_{\text{DDopt}} = nV_{th}(2 - \text{lambert}W(\beta)) \tag{10}$$

$$\beta = \frac{-2C_{\text{eff}}}{W_{\text{eff}} L_{\text{DP}} K C_g} e^2 > -e^{-1} \tag{11}$$

The Lambert W Function, $W = \text{lambert}W(x)$, gives the solution to the equation $We^W = x$, just as $W = ln\ x$ is the solution to $e^W = x$ [15]. Now, substituting Equation (5) into $f = (t_d L_{DP})^{-1}$ gives V_{Topt} to achieve a given frequency:

$$V_{\text{Topt}} = V_{\text{DDopt}} - nV_{th} \ln\left(\frac{fKC_g L_{\text{DP}} V_{\text{DDopt}}}{I_{o,g}}\right) \tag{12}$$

If the argument to the natural log in Equation (12) exceeds 1, then the assumption of sub-threshold operation no longer holds because $V_{\text{Topt}} < V_{\text{DDopt}}$. This constraint shows that there is a maximum achievable frequency for a given circuit in the sub-threshold region. Equations (10) and (12) give V_{Topt} and V_{DDopt} for a sub-threshold circuit consuming the minimum energy at a given frequency. Some ultralow power applications, such as energy scavenging sensor nodes, might consider minimizing energy to be more important than any performance requirement. Assuming a standard technology where V_T is fixed (i.e., no triple wells for body biasing), the problem becomes finding the optimum V_{DD} to minimize energy per operation for a given design. The optimum V_{DD} in this scenario still is given by Equation (10), and the frequency at the optimum point is given by $f = (t_d L_{DP})^{-1}$.

Figure 4.1-7 shows the energy profile of an 8-bit, 8-tap parallel programmable FIR filter versus V_{DD}. The contributions of dynamic and leakage energy are both shown. The lines on the plot show the results of numerical equations using a transregional current model [16], and the markers show the simulation values. The analytical solution (small star) matches the numerical model and simulations with less than 0.1% error. The optimum point is $V_{\text{DDopt}} = 250\,mV$ at a frequency of 30 kHz. Equation (10) provides the optimum V_{DD} for the analytical solution, and substituting this value into Equation (8) gives the total energy. The inset in Figure 4.1-7 shows how the delay (t_d) and current (I_{LEAK}) components of leakage energy per cycle (E_L) vary with supply voltage. As V_{DD} reduces, the current decreases due to the DIBL effect. However, the delay increases exponentially in the sub-threshold region, leading to the increase in sub-threshold E_L.

Equation (10) shows that the optimum V_{DD} value is independent of frequency and V_T. Instead, it is set by the relative significance of dynamic and leakage energy components as expressed in Equation (8). E_L increases compared to the characteristic inverter in two ways. First, the ratio of $C_{\text{eff}}/W_{\text{eff}}$ decreases, indicating that a greater fraction of the total width is idle and thereby drawing static current without switching. Secondly, L_{DP}

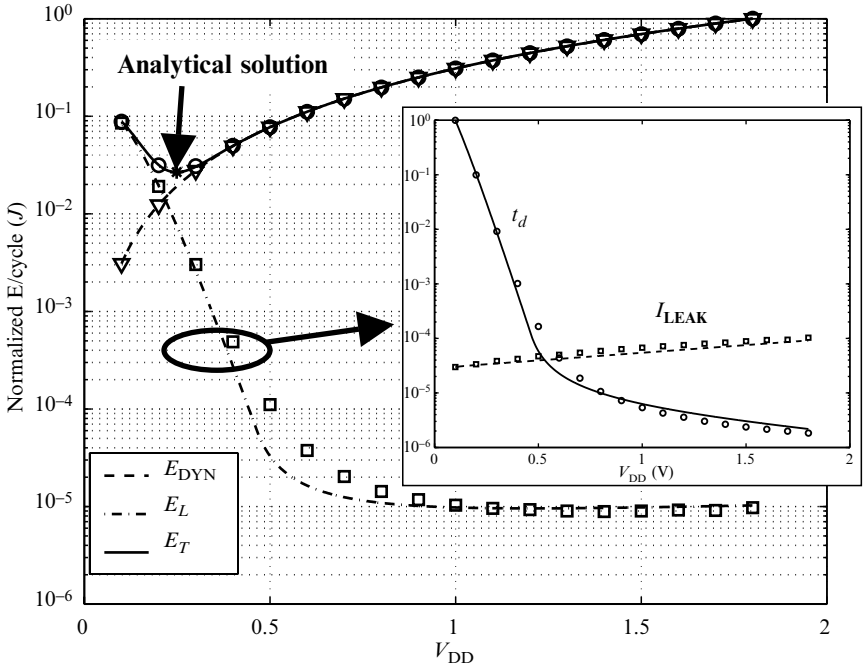


Figure 4.1–7. Model versus simulation of an 8-bit 8-tap FIR filter showing the minimum energy point and active and leakage energy. Inset shows I_{LEAK} and t_d. Markers are simulation values, lines are model (Copyright 2004, IEEE).

can increase. The larger resulting period gives more time for leakage currents to integrate, raising E_L. An FFT processor [14] and the FIR filter previously described have V_{DDopt} at 350 mV and 250 mV, respectively. Circuits with higher relative leakage energy, like the FIR filter or FFT processor, have less negative β and thus higher optimum V_{DD}.

4.2. Sizing and Minimum Operating Voltage

Transistor sizing impacts the functionality of CMOS circuits at low supply voltages. Minimum V_{DD} operation occurs when the PMOS and NMOS devices have the same current (e.g. [17]). Previous efforts have explored well biasing to match the device currents for minimum voltage operation of ring oscillators [13]. Sizing can create the same symmetry in device current. Figure 4.1-8 shows the minimum voltage for which a ring oscillator maintains 10–90% voltage swing. The optimum PMOS/NMOS width across all process corners is 12, because this size matches the sub-threshold currents through the two types of devices.

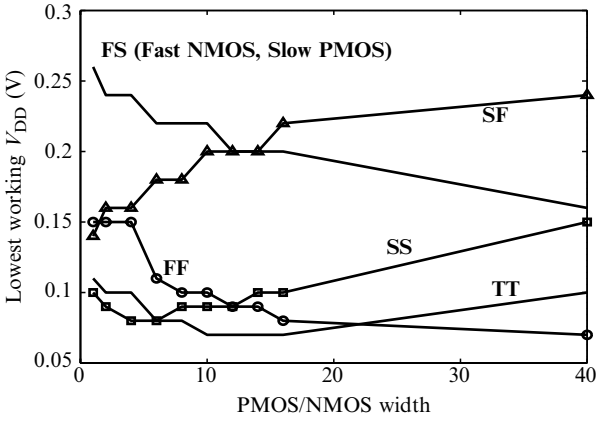


Figure 4.1–8. Minimum V_{DD} to retain 10–90% output swing for $0.18-\mu m$ ring oscillator across process corners (simulation) (Copyright 2004, IEEE).

Sizing according to this ratio allows for operation at lower V_{DD} but increases the energy consumed for a given V_{DD} (Equation (8). The energy savings from lowering V_{DD} are at best proportional to V_{DD}^2 if leakage is still negligible. Figure 4.1-8 shows that the impact of sizing an inverter on the minimum supply voltage is only 60 mV, producing best-case energy savings of $0.20^2/0.26^2 = 0.6X$ due to voltage reduction. This improvement is not worthwhile if all PMOS devices are increased in size by 12X. Thus, minimum sized devices are theoretically optimal for reducing energy per operation when accounting for the impact of sizing on voltage and energy consumed [18]. Process variation in deep submicron processes imposes one restriction to applying this rule blanketly. The sigma for V_T variation due to random doping fluctuations is proportional to $(WL)^{-1/2}$, so minimum sized devices produce the worst case random V_T mismatch. Statistical analysis is necessary to confirm functionality in the face of process variation, and some devices might need to increase in size to ensure acceptable yield.

4.3. Standard Cells and Minimum Energy

Standard cell libraries aid digital circuit designers to reduce the design time for complex circuits through synthesis. Most standard cell libraries focus on high performance, although including low power cells is becoming more popular [19]. Lower power cells generally use smaller sizes. One standard cell library geared specifically for low power uses a reduced set of standard cells and branch-based static logic to reduce parasitic capacitances. Eliminating complicated cells with large stacks of devices and using

a smaller total number of logic functions was shown to reduce power and improve performance [20]. Standard cell libraries have not been designed specifically for sub-threshold operation. This section evaluates the performance of a 0.18-μm standard cell library in sub-threshold operation.

We use the 8-bit, 8-tap FIR filter to compare normal cell selection with cells sized to minimize the operating voltage. Figure 4.1-9 shows the minimum operating voltages for the different standard cells appearing in a normal synthesis of the FIR filter. The typical (TT) and worst-case (FS (fast N, slow P) and SF) process corners are shown. All of the cells operate at 200 mV at the typical corner, showing the robustness of static CMOS logic. Additionally, most of the cells operate at 300 mV in the worst case, which is close to the optimum performance shown in the previous section for a ring oscillator. The cells which exhibit the worst case (failing below 400 mV) are flip-flops and complex logic gates with stacks of series devices (e.g., AOI). We eliminated the problematic cells by preventing the synthesis tool from selecting logic gates with large device stacks and by resizing the offending flip-flop cell [18].

Figure 4.1-10 shows a schematic of the D-flip-flop. In the standard implementation, all of the inverters use small NMOS and only slightly larger PMOS devices except I3, which is several times larger to reduce CQ delay. At the FS corner (fast NMOS, slow PMOS), the narrow PMOS in I6 cannot hold N3 at a one when CK is low. This is because the combined, strong *off* current in the NMOS devices in I6 and I3 (larger sized) overcomes the weakened, narrow PMOS device in I6. The combined NMOS devices create an effective P/N ratio that is less than one. To prevent this,

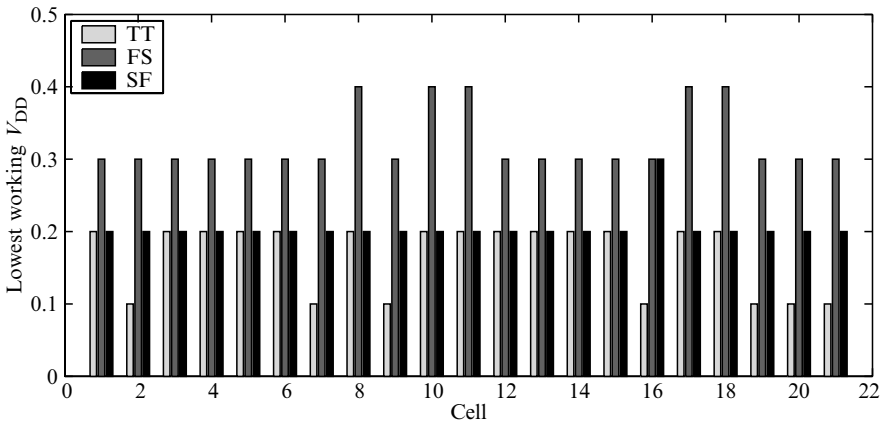


Figure 4.1–9. Functionality of standard cells over process corners in an FIR synthesized using normal cell selection (simulation) (Copyright 2004, IEEE).

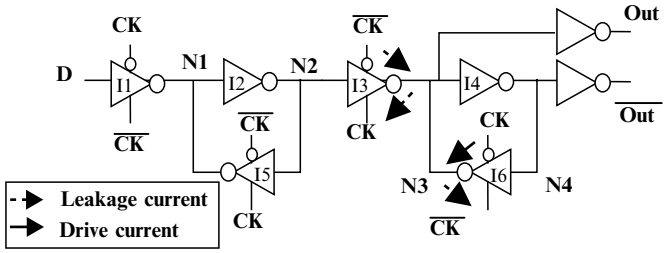


Figure 4.1–10. Standard cell flip-flop at worst-case failure point where CK = 0 at FS corner (fast NMOS, slow PMOS).

we reduced the size of I3 and strengthened I6. The larger feedback inverter creates some energy overhead. However, the resized flip-flop can operate at 300 mV at all process corners in simulation. Figure 4.1–11 shows the lowest operating voltage for the cells in the minimum-V_{DD} FIR filter. The number of cell types has reduced, and all of the cells work to below 300 mV across all corners. The next section uses test chip measurements to compare the filter sized for minimum V_{DD} with the normal filter.

4.4. Measured Results from Test Chip

A 0.18-μm, 6-M layer, 1.8 V, 7 mm^2 test chip was fabricated to measure the impact of sizing on minimum energy operation of standard cells. The test chip features two programmable 8-bit, 8-tap FIR filters. Both filters produce nontruncated 19-bit outputs. The first filter was synthesized using the un-

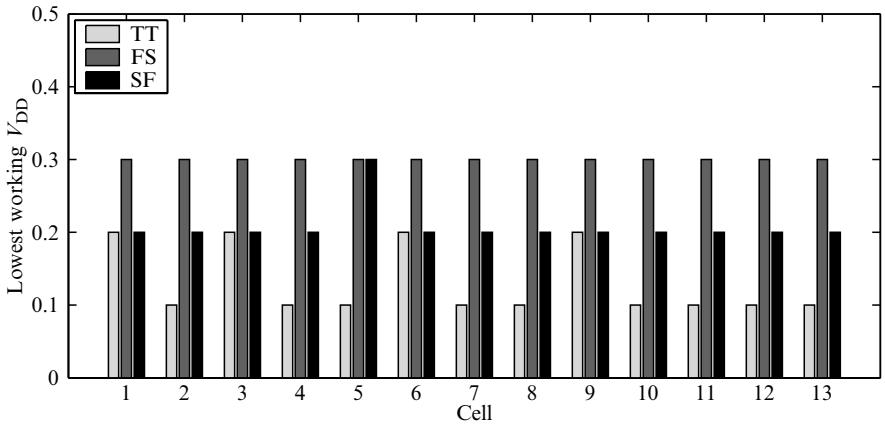


Figure 4.1–11. Functionality of standard cells over process corners for an FIR synthesized using cells sized to minimize V_{DD} (simulation) (Copyright 2004, IEEE).

modified synthesis flow and normal cells (Figure 4.1–9). The second filter was synthesized using the modified flow, in which some cells were omitted and some cells were resized to minimize V_{DD} (Figure 4.1–11). Both filters can operate using an external clock or an on-chip clock generated by a ring oscillator that matches the respective critical path delay of the filters. Filtered data comes from an off-chip source or from an on-chip linear-feedback shift-register. The minimum-V_{DD} filter exhibits a 10% delay penalty over the standard filter. Both filters operate in the range of 3 kHz to 5 MHz over V_{DD} values of 150 mV to 1 V and are fully functional to below 200 mV.

Figure 4.1–12 shows an oscilloscope plot of the standard filter working correctly at $V_{DD} = 150\,\text{mV}$. The clock in this plot is produced by the ring oscillator on-chip. The reduced drive current and large capacitance in the output pads of the chip cause the slow rise and fall times in the clock, but the signal is still full swing. One bit of the output is shown.

Figure 4.1–13 shows the measured total energy per output sample of the two FIR filters versus V_{DD}. The solid line is an extrapolation of $C_{eff} V_{DD}^2$ for each filter, and the dashed lines show the measured leakage energy per cycle. Both filters exhibit an optimum supply voltage for minimizing the total energy per cycle between 250 and 300 mV. There is a measured overhead energy per cycle of 50% in the filter sized for minimum V_{DD}. The figure also shows the worst-case minimum V_{DD} for the two filters (cf. Figures 4.1–9 and 4.1–11). Accounting for overhead at

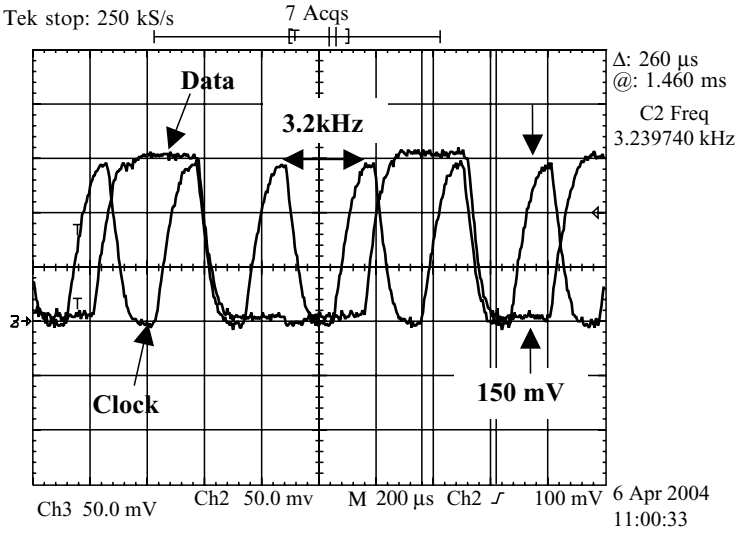


Figure 4.1–12. Oscilloscope plot showing FIR operation at $V_{DD} = 150\,\text{mV}$.

the worst-case minimum V_{DD}, the minimum-V_{DD} FIR offers a reduction in total energy of less than 10% at the worst-case process corner, but this improvement comes at a cost of 50% at the typical corner.

Simulations show that the measured overhead cost in the minimum-V_{DD} filter primarily results from restricting the cell set that the synthesis tool could use. Since the tool was not optimized for the smaller set of cells, we did not see the improvements that are possible through this approach [20]. Using only sizing to create the minimum V_{DD} filter would have decreased the overhead. However, the shallow nature of the optimum point in Figure 4.1–13 shows that the unmodified standard cell library does not use much extra energy by failing at a higher V_{DD} at the worst-case corner. Thus, existing libraries provide good solutions for subthreshold operation. Simulation shows that a minimum-sized implementation of the FIR filter has 2X less switched capacitance than the standard FIR, so a mostly minimum-sized library theoretically would provide minimum energy circuits [18].

5. CONCLUSIONS

This chapter describes energy reduction methods at the algorithmic, architectural, and circuit levels of digital design. System-level architectural and algorithmic modifications allow trading off energy consumption versus other design metrics, such as area or delay at a high level of abstrac-

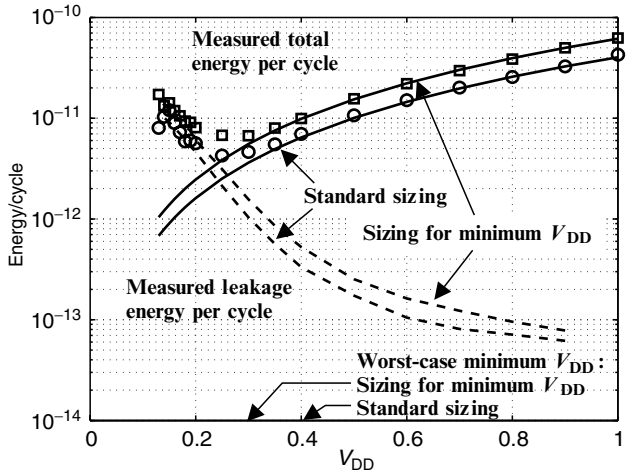


Figure 4.1–13. Measured energy per cycle for the FIR filters on the test chip (Copyright 2004, IEEE).

tion. The addition of architectural knobs that enable smooth adaptation to different operating scenarios can provide energy awareness. An energy-aware system provides significant energy savings by exploiting the architectural knobs as its requirements vary. Modeling and analyzing sub-threshold operation gives a solution for the optimum V_{DD} and V_T to minimize energy for a given frequency in the subthreshold region. For typical circuits and modern technologies, the optimum supply voltage for minimizing power is higher than the failure point for minimum sized devices at the typical corner. Thus, minimum sized devices are theoretically optimal for minimizing power. Measurements from a test chip confirm that existing static CMOS standard cell libraries function well in sub-threshold. Resizing or restricting cell usage in such libraries can lower the worst-case minimum V_{DD}, but the overhead increases energy consumption at the typical corner. In theory, a standard cell library primarily using minimum-sized devices would minimize energy per operation.

ACKNOWLEDGEMENTS

This work was sponsored by the Defense Advanced Research Projects Agency (DARPA) through a subcontract with MIT Lincoln Labs and by Texas Instruments.

REFERENCES

[1] Calhoun, B. H., et al., 2005, Design considerations for ultra-low energy Wireless microsensor nodes, *IEEE Transactions on Computers*, **54**(6), 727–740.
[2] Berrou, C., Glavieux, A. and Thitimajshima, P., 1993, Near Shannon limit error-correcting coding and decoding: Turbo-codes, *ICC*'93, Geneva, Switzerland, 1064–1070.
[3] Schurgers, C., Catthoor, F. and Engels, M., 2001, Memory optimization of MAP turbo decoder algorithms, *IEEE Transactions on VLSI Systems*, **9**(2), 305–312.
[4] Thul, M., Vogt, T., Gilbert, F. and Wehn, N., 2002, Evaluation of algorithm optimizations for low-power turbo-decoder implementations, *ICASSP'02*, Orlando, FL, III-3101–3104.
[5] Garrett, D., Xu, B. and Nicol, C., 2001, Energy efficient turbo decoding for 3G mobile, *ISLPED'01*, Huntington Beach, CA, 328–333.
[6] Benedetto, S., Divsalar, D., Montorsi, G. and Pollara, F., 1997, A soft-input soft-output APP module for iterative decoding of concatenated codes, *IEEE Comm. Letters*, **1**(1), 22–24.

[7] Schurgers, C. and Chandrakasan, A., 2004, Traceback-enhanced MAP decoding algorithm, *ICASSP'04*, Montreal, Canada, pp. IV.645–IV.648, May 17–21, 2004.
[8] Meninger, S., et al., 2001, Vibration-to-electric energy conversion, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 64–76.
[9] Wang, A., Chandrakasan, A. and Kosonocky, S. V., 2002, Optimal supply and threshold scaling for sub-threshold CMOS circuits, *Proc. Symposium on VLSI*, 2002, 5–9.
[10] Burr, J. and Peterson, A., 1991, Ultra low power CMOS technology, *3rd NASA Symposium on VLSI Design*, 4.2.1–4.2.13.
[11] Vittoz, E., 1994, Micropower techniques, *in* J. E. Franca and Y. P. Tsividis (Eds), *Design of VLSI Circuits for Telecommunication and Signal Processing*, chapter 5, Prentice Hall.
[12] Kim, H. and Roy, K., Ultra-low power DLMS adaptive filter for hearing aid applications, *ISLPED*, 2001.
[13] Bryant, A., et al., 2001, Low-power CMOS at Vdd = 4kT/q, *Device Research Conference*, 22–23.
[14] Wang, A., and Chandrakasan, A., A 180mV FFT processor using subthreshold circuit techniques, *ISSCC*, 2004.
[15] Corless, R., et al., 1996, On the Lambert W Function, *Advances in Computational Mathematics*, **5**, 329–359.
[16] Calhoun, B.H., and Chandrakasan, A., Characterizing and modeling minimum energy operation for sub-threshold circuits, *ISLPED*, August 2004.
[17] Schrom, G. and S. Selberherr, 1996, Ultra-low-power CMOS technologies, *Intl. Semiconductor Conf.*, 237–246.
[18] Calhoun, B. H., Wang, A. and Chandrakasan, A., 2004, Device sizing for minimum energy operation in sub-threshold circuits, *CICC*, 95–98.
[19] Piguet, C., 1998, Design of low-power libraries, *Intl. Conf. on Electronics, Circuits and Systems*, **2**, 175–180.
[20] Piguet, C., et al., 2001, Low-power low-voltage library cells and memories, *ICECS*, 1521–1524.

Chapter 4.2

ANALOG INTERFACE CIRCUITS
The Limit for AmI Applications

Michiel Steyaert, Willem Laflere, and Wim Vereecken
ESAT-MICAS, KULeuven
{michiel.steyaert, willem.laflere, wim.vereecken}@esat.kuleuven.be

Abstract Due to further integration towards nanometer technologies, complex functions with a lot of DSP can be single-chip integrated at extreme low powers. However, for the analog interface circuits other limitations are appearing. Due to limited matching specifications, noise limitations, dynamic range requirements, and lower allowed power supply voltages, the analog circuits are becoming the bottleneck in the fully integration of wireless low power ambient intelligence (AmI) devices. Power management in combination with special high efficiency architectures are usually in contrast with integration robustness and fully integration. An overview of limitations towards ultra low power analog RF interface circuits will be discussed.

Keywords analog; CMOS; limitations; receiver; RF; transmitter

1. INTRODUCTION

The enormous growth of the telecommunications market is the driver behind the development of performant RF circuits in low-cost technologies. To increase autonomy, devices are equipped with wireless frond-ends. Since portable devices are battery powered, limited battery lifetime has become an important design limitation. Deep submicron technologies allow to extend this boundary, but new problems are introduced: analog RF front-ends remain the bottleneck in the mainly digital CMOS technology. The goal of this work is to localize new opportunities, but also to uncover potential problems related to power consumption, single-chip integration, and technology scaling.

S. Mukherjee et al. (eds.), AmIware, 203–222.

2. HETERODYNE ARCHITECTURE

The choice of an appropriate transceiver architecture is of major importance in the design flow of an RF front-end. In addition, the selection of the modulation scheme determines the overall performance of the transceiver: not only bandwidth efficiency or throughput, but also power consumption. Frequently, these design parameters are overlooked, resulting in a performant, but power-inefficient design.

The block diagram of the famous heterodyne receiver [1] architecture is shown in Figure 4.2–1. A band select filter (1) removes high power signals *(blockers)* out of the band of interest. In this way the dynamic range requirements of the ensuing analog circuits can be relaxed. The contribution to the equivalent input noise of the remaining receiver part is reduced by a low noise amplifier (LNA). From this point on, the RF signal is converted to an intermediate (IF) frequency. In the mixing process, unwanted signals are folded into the band of interest. The linearity of the mixer prevents this irreparable corruption of the wanted signal. The IF-filter strips off unwanted mirror signals from the mixer. Usually, this filter is also used as the channel select filter. Surface acoustic wave (SAW) filters are the common way to achieve the required selectivity. Power hungry drivers necessarily get the signal off-chip and drive these low-impedance filters. Eventually, a second mixer converts the IF-band down to baseband. From this point, the signal can be digitized by an AD-converter for further processing by DSP logic.

3. LOW-IF ARCHITECTURE

The low-IF architecture [2], shown in Figure 4.2–2 avoids the external IF-filter section. The RF-signal is fed to an in phase (I) and a quadrature path (Q). The mixers of each path are driven by a quadrature LO signal.

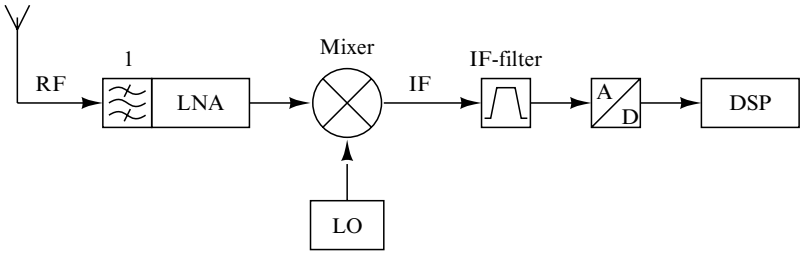


Figure 4.2-1. Heterodyne receiver architecture.

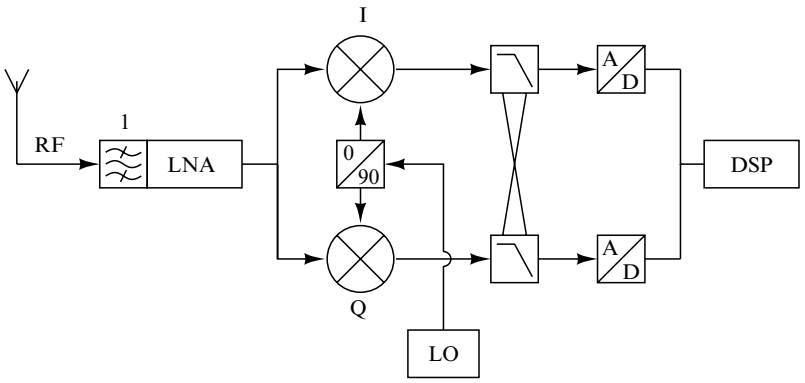


Figure 4.2-2. Low-IF receiver architecture.

In this type of receiver, mirror signals are suppressed by an analog (before conversion), or a digital (after conversion) on-chip complex filter. If an analog filter is used, this filter has the added advantages of being the antialiasing filter for the AD-converter and the low-pass filter for channel selection. The advantages of this architecture are obvious: since the RF signal band is directly converted to baseband, a complete mixer- and IF-circuit can be omitted. However, there is a disadvantage: phase and amplitude errors of the quadrature generator directly result in a bad mirror suppression. A careful design of the quadrature generator is highly recommended. The good power performance and low external component count of the low-IF structure are its trump cards for use in low-cost battery powered systems.

4. RECEIVER VERSUS TRANSMITTER

The same architectural principles apply to both transmitters and recievers, however, the signals involved are quite different. The function of the receiver is to detect a small signal from a dirty environment full of blocking signals. While in the transmit chain, the only signal involved is the one being transmitted. Important specifications in the receive part are noise added to the signal and intermodulation with blocking signals. In the transmitter, important specs are the linearity of the signal and unwanted emissions in nearby channels.

As faster technologies become available for the design of a certain application, some architectural changes can be exploited to gain global efficiency. Some building blocks of the modular design, as described in the previous part can be combined together if speed is not the limiting factor

in a certain technology. For example, all channels in one band can be digitized together, so shifting functions of the PLL to the A/D-convertor and the digital post-processing. But when working on the edge of a technology, the only suitable architecture is the one described in the previous paragraph; so it will be discussed later.

5. CONSEQUENCES OF THE TECHNOLOGY SCALING TRENDS

Technology keeps scaling down in an exponential way, as predicted by Moore's law 40 years ago. [3] Figure 4.2–3(a). The scaling is driven by the demand for ever faster and smaller digital circuitry [4]. As it is the goal to integrate analog and digital circuitry all together, analog designs must follow the same scaling trends. Benefit is taken from this scaling so that higher operating frequencies are possible. Figure 4.2–3(c) shows the increase in f_T (the maximal operating frequency of a transistor) and the f_{3dB} (the maximal frequency in a practical circuit).

A consequence of the smaller line widths is a decrease in maximum allowable supply voltage, in order to keep the electric field strengths in the silicon within the limits. The supply voltage, being 5 V for a long time, has decreased rapidly in recent years, and will keep decreasing in the future. The threshold voltage to put a transistor on and off remains more or less constant for leakage reasons in the off-state. So, the margin between V_t and V_{dd} has disappeared. (Figure 4.2–3(b))

Matching, or the difference in electrical characteristics between devices that are designed behave the same, has also evolved with technology. (Figure 4.2–3(d)) The mismatch of the V_t is becoming smaller in advanced technologies, since the value of V_t has become a more critical parameter with the power supply voltage decreasing. The mismatch on β related to the gain of a transistor does not evolve a lot. It is related to both the mobility in the channel and the gate oxide thickness. The very thin gates in advanced technologies show a larger spread in thickness.

The consequences of these evolutions on RF-transmitters will be discussed based on building block examples.

6. LNA

The low noise amplifier [5] (LNA) is a crucial building block in today's high-performance receivers. Even though the active component count of the LNA is low, it fulfills several complex tasks in the receiver chain. This

section gives an overview of the performance requirements of a modern LNA circuit and discusses several problems and opportunities if an LNA is integrated in a single-chip, low-voltage application.

The main task of the LNA is to amplify the antenna signal to a higher level. The main reason for this is, as described in the previous section, that a high amplification factor in the LNA makes the receiver less sensible to the noise of the *ensuing* receiver circuits.

However, high signal levels will result in distortion in both the LNA and the ensuing receiver circuits. Due to third-order intermodulation products, unwanted signals will be mixed into the band of interest, again increasing the noise. For each specific application, an optimum exists

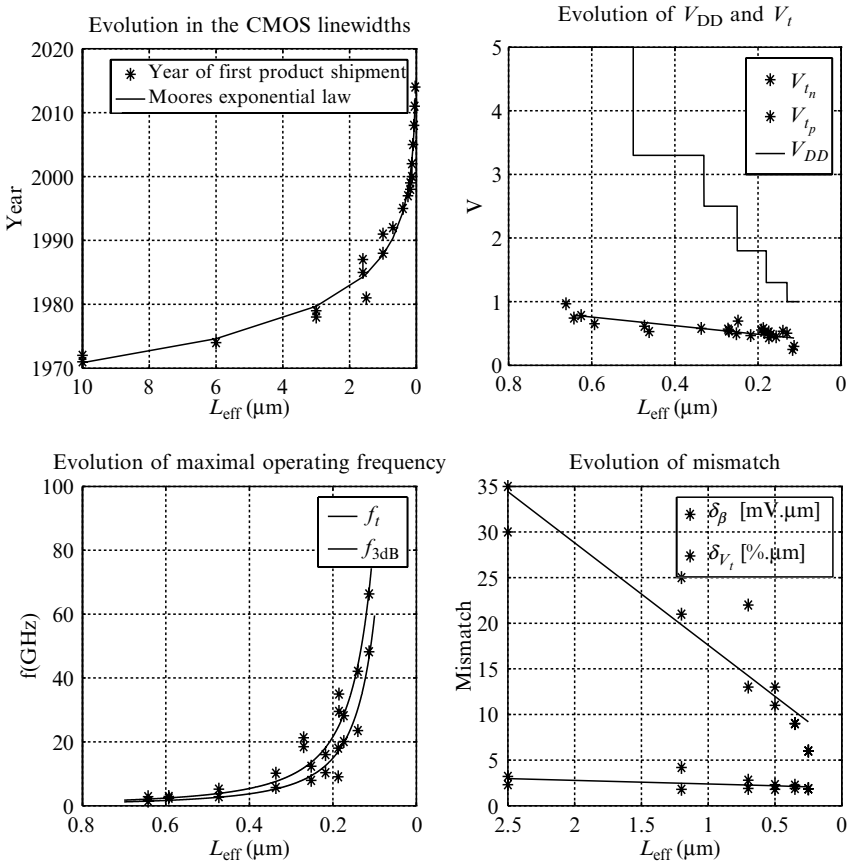


Figure 4.2-3. Consequences of technology scaling.

between thermal noise and distortion. Comparison between LNA's can thus be done by their crucial parameters: gain (G[dB]), the noise figure (NF[dB]) and the input referred third-order intermodulation distortion power (IIP_3[dB]).

The reverse gain of the LNA (also called the reverse isolation or S_{12}) defines the gain from output to the input of the LNA. The signal coming from the local oscillator may couple to the input of the mixer and reach the output of the LNA. The reverse isolation of the LNA defines the amount of this signal that is transmitted to the antenna of the receiver. The basic driving force to limit the S_{12} is based on the spurious emission specifications of the receiver.

The most popular LNA topology, known as the inductively degenerated common source LNA, is shown in Figure 4.2–4, together with its small signal equivalent (figure 4.2–5).

The base of this amplifier is transistor M_1. To decrease the Miller effect, cascode transistor M_2 is added. This also increases the reverse isolation of the LNA. To obtain a large load resistance in combination with a low DC voltage drop, an inductor L_d is used. If the LNA is integrated on a single chip, the resonant frequency of the gate-source capacitance in combination with the load inductance can be chosen so that the voltage gain is maximized for the frequency band of interest.

In order to maximize the power transfer between the antenna and the LNA, the input impedance must be matched to the impedance of the antenna (e.g., 50 Ω). This can be done by adding a grounded resistor at the gate of transistor M_1. However, this type of impedance matching adds too much noise to the amplifier stage. Instead of a resistor, an inductor L_s

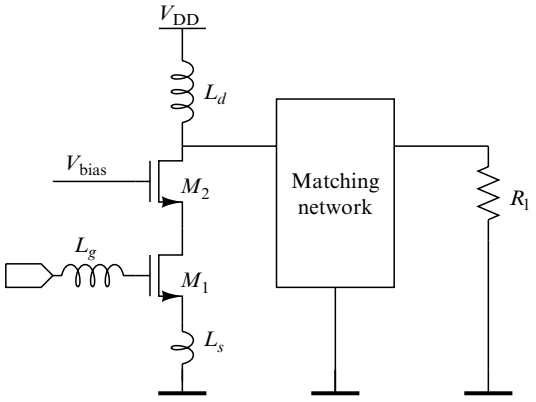


Figure 4.2-4. Inductively degenerated common source LNA.

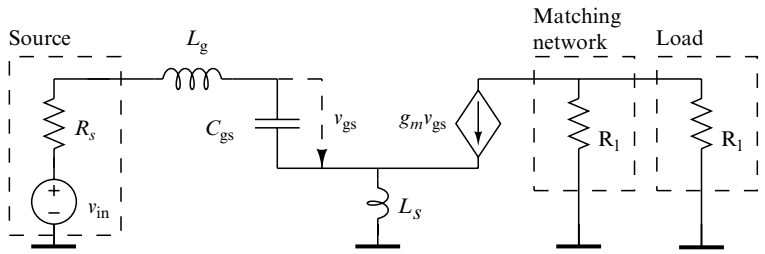


Figure 4.2-5. Small signal equivalent of the CS-LNA.

can be added between the source of M_1 and the ground. The input impedance can be calculated as:

$$Z_{\text{in}} = \frac{1}{j\omega C_{\text{gs}}} + j\omega L_s + \omega T L_s,$$

which reduces to the resistive value $Z_{\text{in}} = \omega_T L_s$ for the operating frequency ω_0. The gate inductance L_g is chosen such that the imaginary part of the input impedance is zero for the frequency of interest.

If the LNA is connected to a resistive load R_L, the maximum power gain can be calculated as:

$$G_T = \frac{R_L}{4R_S}\left(\frac{\omega_T}{\omega_0}\right)^2.$$

Notice that the power gain goes up with increasing ω_T. This implicates that deep submicron technologies improve the gain of the LNA. However, it has to be noticed that G_T and ω_T are a function of $V_{\text{gs}} - V_T$.

Using deep submicron imposes to reduce $V_{\text{gs}} - V_T$, which could result in reduced performances towards low-voltage applications, through ω_T, which can be a problem in low-voltage applications.

The main noise sources of the previously described LNA are the channel noise of M_1 and the thermal noise of the load resistor R_L. It can be shown that the noise factor of the amplifier reduces to the following expression:

$$F \approx 1 + \frac{\gamma}{\alpha} g_{d0} R_S \left(\frac{\omega_0}{\omega_T}\right)^2 + 4\left(\frac{\omega_0}{\omega_T}\right)^2 \frac{R_S}{R_L},$$

where g_{d0} is the drain-source conductance at zero V_{ds}. The excess noise factor γ is one at $V_{\text{DS}} = 0$, and decreases to 2/3 in saturation. Due to high

electric fields in short-channel devices, γ can even be larger than 3. One can see that smaller technologies improve the noise figure through ω_T, despite the counteracting effect of the increasing excess noise factor γ. The table below shows several ways to decrease the noise figure:

	Fixed parameter	Variable parameter
NF	bias current I	V_{GST}
NF	device width W/L	V_{GST}
NF	V_{GST}	I_{bias}

For an NMOS in the saturation region, the input referred third-order intermodulated interception power IIP_3 can be approximated by:

$$IIP_3 = 5.25 + 20\log_{10}(2\omega_0 R_S C_{gs}) + 10\log 10 \times \left(\frac{V_{GST}\cdot(2+\Theta V_{GST})\cdot(1+\Theta V_{GST})^2}{\Theta}\right)$$

in which $V_{GST} = V_{GS} - V_T$. Factor Θ models the mobility degradation due to velocity saturation. For a fixed current, increasing the V_{GST} results in decreasing the device width and thus also C_{gs}. These effects almost cancel each other, which result in the fact that the only way to improve the linearity of the inductively degenerated common source LNA is to increase the current consumption. On the other hand, deep submicron devices offer the ability to bias the transistors in the velocity saturation region. In this way, a relatively constant transconductance and thus a higher linearity can be achieved if sufficient gate overdrive voltage is applied.

Another important issue in CMOS LNA-design is ESD (Electrostatic Discharge). As technologies scale further towards nanometer dimensions, the decreasing oxide thickness reduces the breakdown voltage for a given technology. The parasitic capacitance of the ESD-protection structures has a serious impact on the performance of the LNA, including ESD-protection in the LNA-design, which is thus inevitable.

7. DOWNCONVERSION MIXER

In a direct conversion architecture, considerations about mixer noise, gain, and linearity are similar to LNA linearity, since the signals involved are the same. In a heterodyne architecture, linearity specs are relaxed proportional to the amount of filtering. Also the amount of harmonics in the spectrum of the LO is important, since higher harmonics also

downconvert noise and spurs within 2–3, ... times the center frequency into the signal band. These components can not be filtered away afterwards. For example, in a square wave, the power of the third harmonic is only −9.54 dB below the carrier. As a result, the mixer should show a linear behavior in both LO and RF-path, and the LO-signal should not contain a large amount of harmonics.

When mixing with a quadrature signal, another important topic is the quality of the quadrature signal. Imperfect quadrature will result in imperfect image rejection for a low-IF architecture and distortion of the signal in a zero-IF receiver. Several mismatches, amongst others in the quadrature generation circuitry, or quadrature mismatch between the mixers will contribute to nonideal image-suppression. Figure 4.2-6 shows the image suppression as a function of the total mismatch (mismatch between quadrature LO-signals, mismatch in the mixer, and mismatch in the quadrature baseband processing). For a typical reciever requirement of −35 dB, a phase and amplitude matching of 6 deg and 0.25 dB, respectively, is required. This is in the same order of magnitude as the matching performance of advanced technologies.

8. AD CONVERTER

In today's submicron CMOS technologies, the analog front-end and digital circuits are integrated on the same die. The AD converter [6] forms the interface between the analog continuous-time continuous-amplitude circuits, and the discrete-time and amplitude of the digital processing part. The design of AD converters faces an increasing challenge because of mismatch, reduced supply voltages, and relative high-threshold voltages in a cheap all-digital CMOS technology.

Reducing the analog supply, while preserving bandwidth and dynamic range, has no fundamental effect on the minimum power consumption. However, this absolute limit is usually obtained by neglecting the possible bandwidth B limitation due to the limited transconductance of the active devices. If the maximum bandwidth is proportional to $\frac{g_m}{C}$, one can show that [7]

$$\mathrm{SNR} \cdot B = V_{\mathrm{pp}}^2 \frac{g_m}{8kT}$$

In most cases, decreasing the supply voltage with a factor α results in a proportional reduction α of the signal swing V_{pp}. Preserving SNR and bandwidth is only possible by increasing g_m by α^2.

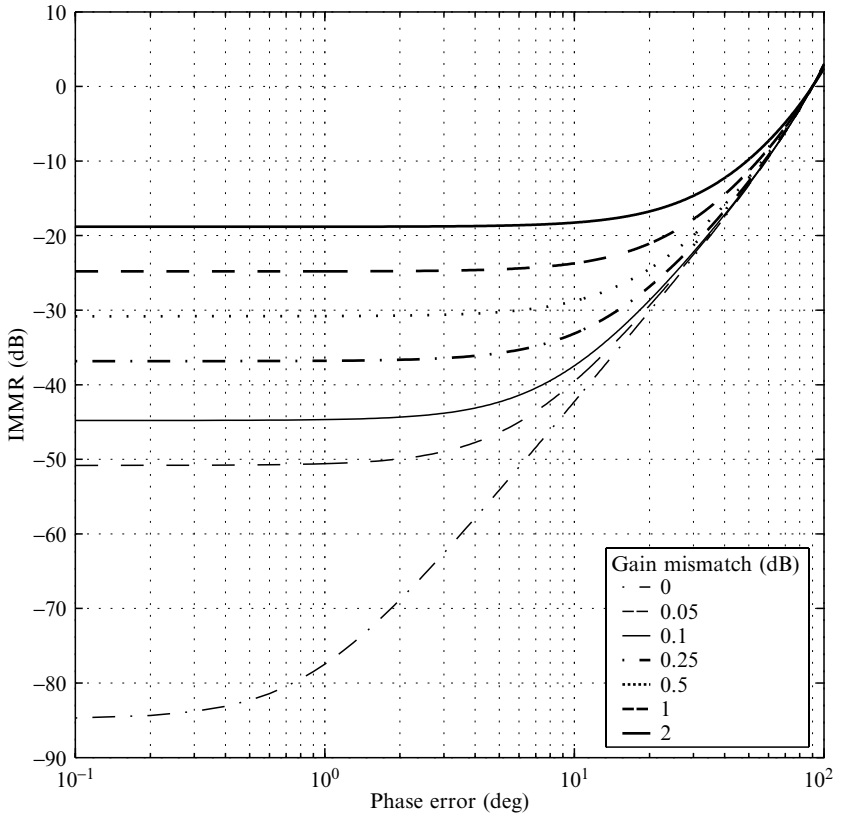


Figure 4.2-6. Image rejection due to mismatch.

In the case of a MOS device, where $g_m \approx \frac{I}{V_{gs}-V_T}$, the power consumption remains constant if the overdrive voltage $V_{\mathrm{gs}}-V_T$ scales down with the same factor α.

This effect can be noticed in high-speed (flash) AD converters [7]. For the same speed and accuracy, the power drain for a high-speed ADC can be compared:

$$\frac{P_1}{P_2} = \frac{V_{dd2}}{V_{dd1}} \cdot \frac{t_{ox1}}{t_{ox2}} \cdot \frac{(1+(C_{db1}/50\varepsilon_{ox}))}{(1+(C_{db2}/50\varepsilon_{ox}))}$$

When going to submicron technologies, oxide thickness is downscaled and the channel doping is increased. As a result the bulk capacitance increases. It is clear that due to a stringent mismatch demand and the increasing bulk capacitance, the expected power decrease is counteracted when technology scales down, as shown in Figure 4.2-7.

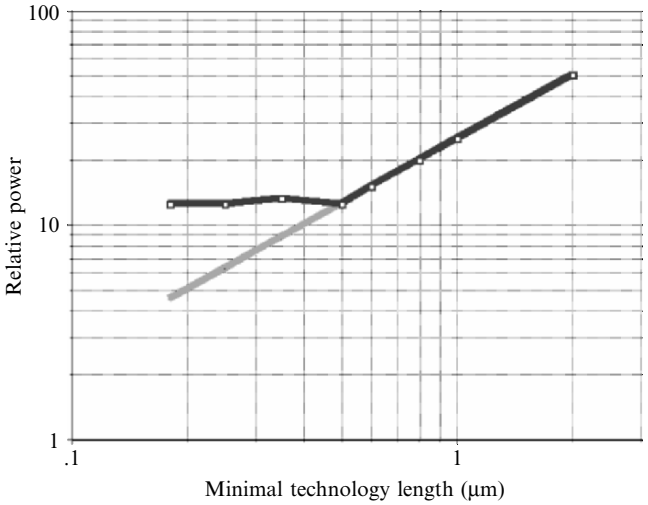


Figure 4.2-7. Power consumption versus technology

However, when technology scales further, the current factor β of the MOS transistor becomes dominant, leading to the following equation:

$$\frac{P_1}{P_2} = \frac{V_{dd2}}{V_{dd1}} \cdot \frac{t_{ox2}}{t_{ox1}} \cdot \frac{(1 + (C_{db1}/50\varepsilon_{ox}))}{(1 + (C_{bd2}/50\varepsilon_{ox}))},$$

which makes the power increase since the power supply scaling is not compensated by an increase of the matching properties of the technology. In these analyses, substrate noise, power supply noise, and ground noise are not considered, but undoubtedly they become more important when technology scales down.

On the other side, an important advantage of moving into deep-submicron technologies is that, the power consumption of digital circuits is decreased while speed is increased. Every AD converter system includes a large amount of digital circuitry. For digital building blocks, it is always beneficial to move into smaller technologies.

For this reason, single loop $\Sigma\Delta$ converters are very attractive since they are always combined with digital circuits on the same die to form a system. As described before, reduction of the supply voltage reduces the maximum signal swing. For the same dynamic range, noise floor and mismatch should be reduced as well. The dynamic range of a $\Sigma\Delta$ converter can be expressed as:

$$\mathrm{DR} = 10 \cdot \log \frac{V_{in,\max}^2 \cdot \mathrm{OSR} \cdot C_s}{2kT}$$

Table 4.2-1. Figure of merit FOM $= \frac{4kT \cdot f_B \cdot \mathrm{DR}}{P}$.

Name	Supply (V)	DR (dB)	Signal BW(kHz)	Power (μW)	FOM
Yao, 2004	1	88	20	140	1493e-6
Peluso, 1998	0.9	77	16	40	300e-6
Gerfers, 2003	1.5	80	25	135	306e-6
Dessousky, 2001	1	88	25	950	275e-6
Sauerrey, 2002	0.7	75	8	80	52e-6
Peluso, 1997	1.5	74	3.4	100	14e-6
Keskin, 2002	1	80	20	5600	5.9e-6

The OSR is the oversampling ratio and C_s is the sampling capacitance in the first integrator. For a certain dynamic range, if the signal swing is reduced, the oversampling ratio or the sampling capacitance must be increased. Mismatch error, on the other hand, can only be reduced by increasing the size of the device. To maintain speed, this will always result in a higher power consumption. The most interesting property of the ΣΔ converter is the noise suppression inside the loop. For a single loop modulator, the noise suppression can be calculated as:

$$F_{\mathrm{sup,k}} = \frac{\mathrm{OSR}^{2k}}{\pi^{2k}} \cdot (2k+1) \cdot \left(\prod_{i=0}^{k} a_i \right)^2,$$

where F is the noise suppression factor and a_i denotes the loop coefficient of the ith stage. This property allows the sampling capacitance of the following stage to be scaled down proportionally. This results in a decrease of the load capacitance of the used OTAs and thus a reduction in power consumption. The advantage of this strategy can be seen in the results of the first 90 nm ΣΔ designs [8]. If a comparison is made based on speed and power, it can be noticed that both for low-voltage designs (Table 4.2-1), and in state of the art reports (Table 4.2-2), 90 nm designs can achieve better performance.

9. LOCAL OSCILLATOR

The core component and the most critical one of the LO-synthesizer is the voltage controlled oscillator. The next section will deal with the specifications and possible implementations.

In an ideal world, the LO-signal is assumed to be one single tone. However, due to several noise contributions, the frequency of the tone of

Table 4.2-2. ISSCC 2004.

Name	Topology	Tech. (μm)	BW (MHz)	DR (dB)	*V*dd (v)	*P* (mW)	FOM
Yao	Switch-cap, 1b	0.09	0.02	88	1	0.14	1487
Putter	Continuous-time, 1b	0.18	1.1	81	1.8	6	381
Gaggi	switch-cap, 3b	0.13	1.1	82	1.5	8	360
Balmeli	switch-cap, 4b	0.18	12.5	84	1.8	200	259
Ying	Continuous-time, Ib	0.18	2.5	86	1.8	150	109
Breems	continuous-time, 4b	0.18	10	67	1.8	122	7
Ueno	continueous-time, Ib	0.13	1.92	55	0.9	1.5	7
Philips	continuous-time, Ib	0.18	1	49	1.8	2.07	1

an oscillator is not stable in time. The spectrum of a noisy oscillator is shown in Figure 4.2-8(b). As a consequence, part of the nearby blocking signal is folded into the signal band when mixing with such a signal (Figure 4.2-8(c)). Phase noise is defined as:

$$L\{\Delta\omega\} = \frac{P_{\Delta\omega}}{P_{\omega 0}}$$

A lot of attempts have been made to model the amount of phase noise. [9] A simple and useful model for the phase noise density at an offset $\Delta\omega$ from the center frequency $\Delta\omega$ in the $1/f^2$-region is:

$$k\mathrm{TR}_{\mathrm{eff}}[1 + A]\left(\frac{\omega 0}{\Delta\omega}\right)^2,$$

in which *Req* is the equivalent series dissipative part of the oscillator. [10]

It is shown that, for GHz applications, the bottleneck for the phase-noise spec is the coil in the resonator. [11, 12] Full integration of entire wireless systems requires this coil to be integrated on the CMOS die too. Losses in integrated coils are mainly due to series resistance of the wiring metal, including skin effect, and Eddy currents generated in the underlying lossy substrate. For these reasons, even in small technologies, high-quality coils do occupy a large part of the chip area. The power lost into the substrate also couples with nearby circuitry, so a careful layout is very important.

For relatively low frequencies, in the order of magnitude of a few GHz, the dominant noise source is the inductor. The contributions of the tuning varactor and the active cell could almost be neglected. Since they are not

contributing to the phase noise, they could be optimized for tuning range, power, ... [12].

For applications in the order of magnitude of tens of GHz, the dominant noise sources become situated in the varactors and gain-cell, since the quality of an inductor becomes better with increasing frequency. A codesign of all the parts of the entire VCO becomes necessary. Full optimization of the tank has led to tank-quality factors of 40 to 50 [12].

10. QUADRATURE GENERATION

In the case the local oscillator is not delivering a differential signal, a quadrature signal must be derived from the differential outputs of the local oscillator. Several techniques are suited for this purpose:

- *Polyphase filter:* The operation is based on the phase shift of RC-and CR-filters near their cutoff frequency. Again matching of integrated resistors and capacitors is a limit to the obtainable bandwidth and accuracy.
- *Digital divider:* A signal of twice the operating frequency is divided in a digital way to 4 quadrature signals. This division is independent of the

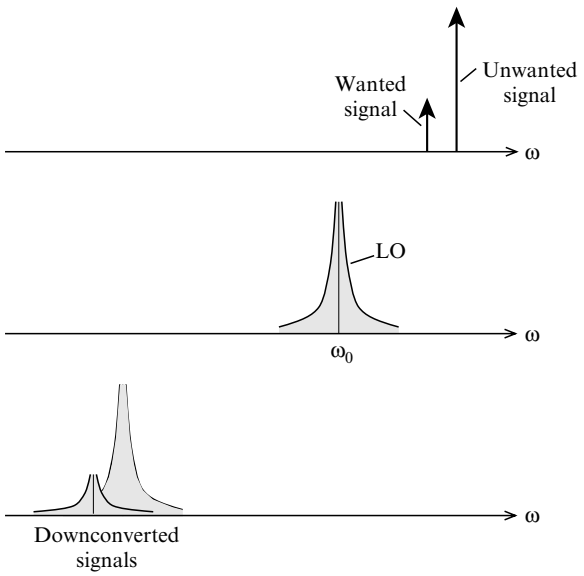


Figure 4.2-8. Effect of a noisy LO-signal.

frequency, yielding a very high bandwidth. Disadvantages with this approach are the higher power consumption due to the high-operating frequency, and the dependency of the duty-cycle, and in this way also the output DC-level, on matching.

- *Injection locked quadrature oscillator:* A quadrature ring oscillator is forced to oscillate at the incoming frequency. The quadrature oscillator locks to the incoming signal if the frequency is within a certain range around the free-running frequency. The same matching sensitivity as above applies to this type of quadrature generator. The power is usually quite low, and the systems have a considerable bandwidth. [13]

11. UPCONVERSION MIXER

Since the only signal involved in the upconversion path is the signal being transmitted, a lot of the specifications for the transmitter depend on the properties of this signal. An important distinction that must be made is the difference between ''constant envelope'' signals and signals carrying information in the amplitude as well. Figure 4.2-9 show the effects of nonlinearity on the Error vector magnitude (=averaged shift of constellation points) for several modulation types. When comparing these results with the Crest-factors given in Table 4.2-3, it is clear that this is a dominant factor in the sensitivity to nonlinearities.

On top of the requirements for a good *EVM,* which is a measure for the quality of the transmitted signal, there are also limitations to the maximal amount of power being emitted in frequency bands away from the channel. Unwanted emissions in these bands are also related to linearity requirements in the transmitter. As the power supply keeps scaling down in advanced CMOS technologies, while the required output power does not, the available headroom disappears, posing severe demands on design of a linear upconvertor.

Table 4.2-3. Crest factors for several modulation types.

Modulation	CF	CF_I
8-PSK	0.0	3.6
16-QAM	2.3	2.6
64-QAM	3.2	4.0
UMTS	3.9	6.5
OFDM	4.7	7.6

12. POWER AMPLIFIER

Usually, the RF power amplifier [14–18] is implemented as a separated building block together with discrete components, like inductors, capacitors, and striplines. Low cost and portable transceiver systems, however, benefit from fully integrated PA's, which limits the freedom of architectures and components that can be used.

RF power amplifiers are designed to amplify the input signal and deliver relatively high power to a low impedance load. DC- to RF-power conversion efficiency is thus important due to their high power consumption in the system. In PA design, several definitions of efficiency are used:

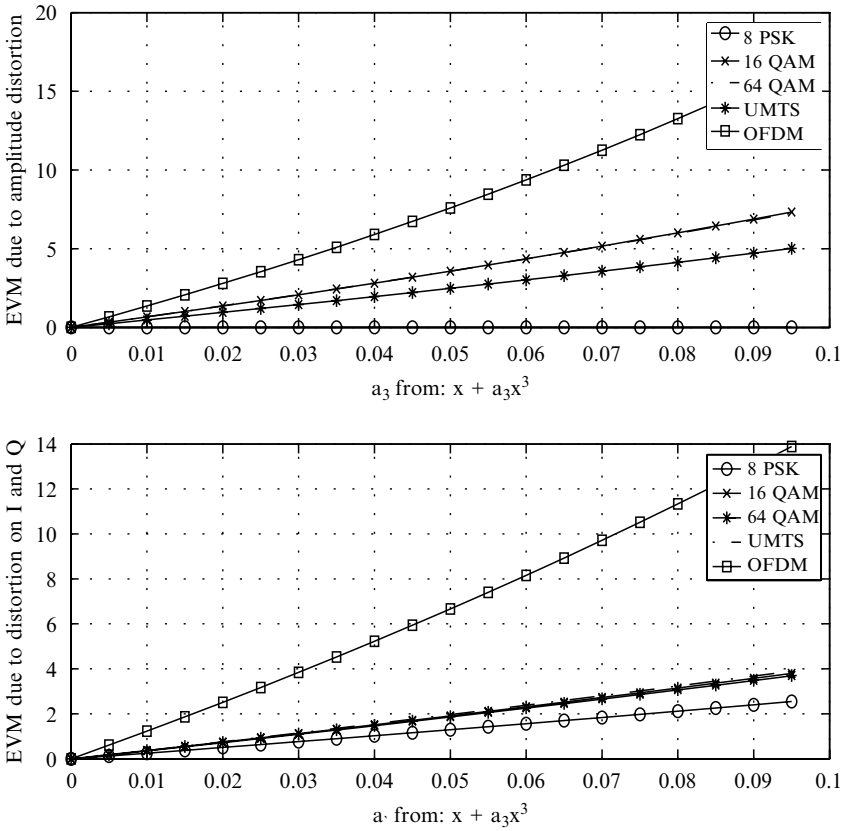


Figure 4.2-9. Deviation from the ideal constellation points due to distortion: (a) Distortion of the amplitude and (b) Distortion of both I and Q signals.

- The drain efficiency *(DE)* is the ratio of RF output power to DC input power:

$$\mathrm{DE} = \frac{P_{\mathrm{out,\ rf}}}{P_{\mathrm{in,\ dc}}};$$

- The power added efficiency *(PAE)* takes the loading of the input source by the amplifying transistor into account:

$$\mathrm{PAE} = \frac{P_{\mathrm{out,\ rf}} - P_{\mathrm{in,\ rf}}}{P_{\mathrm{in,\ dc}}}; \text{ and}$$

- The overall efficiency compares the RF output power with the total input power (RF and DC):

$$\eta_{\mathrm{overall}} = \frac{P_{\mathrm{out,\ rf}}}{P_{\mathrm{in,\ rf}} + P_{\mathrm{in,\ dc}}}.$$

The previous definitions only give us the efficiency at a specific output level. Amplitude modulation affects the RF output power and results in a time-dependent efficiency. In this case, an average efficiency must be calculated.

The linearity of a PA is a measure of how the input signal is reproduced by the PA. A bad reproduction of the input signal results not only in a distorted output signal, but also out-of-band signals. Several types of linearity can be distinguished; amplitude nonlinearity is caused by variable gain or saturation of the amplifier. Amplitude to phase conversion creates unwanted phase modulation of the output. Finally, the frequency response of the amplifier reshapes the spectrum of the output signal. Several measures for linearity exist:

(1) The intermodulation distortion (IMD) compares the fundamental frequency to the third order component, as explained in the LNA section.
(2) The adjacent channel power ratio (ACPR) compares the power in a specified band out of the band of interest to the RMS power of the output signal:

$$\mathrm{ACPR} = \frac{\int_{f_c - of\ f\ \mathrm{set} - \frac{BW}{2}}^{f_c - of\ f\ \mathrm{set} + \frac{BW}{2}} S(f) \cdot df}{\int_{f\ \mathrm{low}}^{f\ \mathrm{high}} S(f) \cdot df}$$

(3) In case of QAM (quadrature amplitude modulated) signal, the distance between the ideal signal vector and the received vector is a measure for distortion. This distance is called the error vector magnitude (EVM).

In case of constant envelope signals (CW, FM, and PM), the power amplifier always operates at *PEP* (peak envelope power), and thus at maximum efficiency. In case of amplitude modulated signals, the output power must be decreased by the *PAPR* (peak to average power ratio). This is called back-off and has a serious impact on the efficiency since the power drained from the supply generally does not scale linear with RF output power. For example, in case of the 128-subcarrier MB-OFDM (Ultra Wideband), the *PAPR* can shown to be as high as 12 dB, which reduces the maximum achievable efficiency compared to a constant envelope signal with a factor of 15. Figure 4.2-10 shows the back-off between the IdB compression point P_{1dB} and the average input power P_{AVG}.

The class A amplifier is biased in such a way that the transistor acts as a current source for all parts of the input waveform. The drawback is that the quiescent current is large and DE is ideally 50% at PEP.

In a class B and C amplifier, the bias voltage is lowered so that the quiescent current is reduced and the transistor is active for half (class B) or less (class C) of the input waveform. The drawback is higher distortion.

A class D amplifier uses two transistors (switches) to generate a square waveform. A series output filter selects the required frequency component. By avoiding a large voltage and large current at the same time, power consumption in the transistors is reduced to a minimum.

Class E amplifiers use one transistor (switch) in combination with a tuned load network. Power dissipation is avoided by enabling the transistor

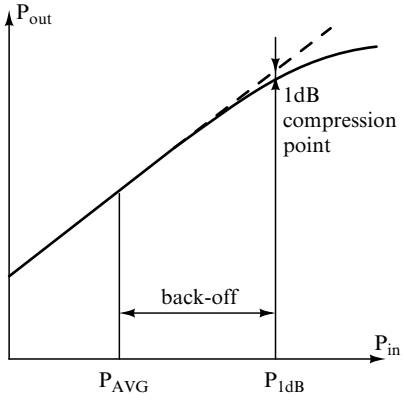


Figure 4.2-10. Back-off.

only when the voltage across it drops to zero. Since the parasitic capacitance of the transistor can be seen as a part the load network, it can be tuned out, which avoids performance degradation. Theoretically, an efficiency of 100% can be achieved.

A class F amplifier makes use of resonators (lumped elements or transmission lines) in the loading network to create a square wave at the transistor drain side, while only passing the wanted frequency to the output. Due to the high-passive component count and low integrate-ability of inductors, this type of amplifier is less suitable for on-chip implementation.

In digital and low power CMOS technologies, switching mode power amplifiers are good candidates due to their high efficiency. A simplified schematic of a class E power amplifier is shown in Figure 4.2-11. The power is supplied through an inductor and can even act as a resonator in combination with the shunt capacitance (which includes the parasitic capacitance of the switching transistor). At high frequencies, the bonding wire inductance and ESD protection circuitry capacitances must be taken into account as part of the loading network.

Due to the switching nature of the structure, the maximum voltage at the drain of the transistor will be higher than the supply voltage, which makes this architecture sensible to failures like oxide breakdown and hot-electron degradation. However, the maximum voltage can be reduced by switching the transistor on before the voltage across it reaches zero.

13. CONCLUSIONS

In this study, the problems related with technology downscaling—limiting the performance of the analog transmit/receive circuitry were

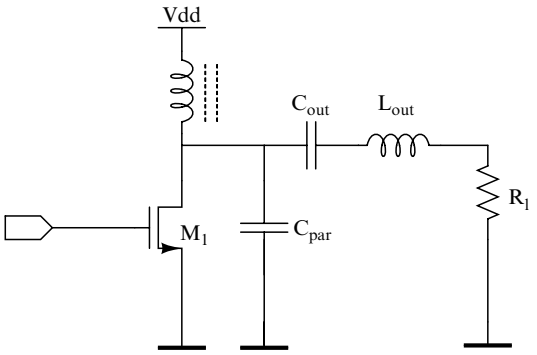


Figure 4.2-11. Simplified class E amplifier.

discussed in detail. It was shown that—although there is a benefit in speed (C decreased) and a benefit in power consumed in the *digital* part, problems arise in the analog part. Most of these problems are due to the decreasing power supply voltage. Due to these problems, analog design becomes the bottleneck in advanced wireless system design.

REFERENCES

[1] Lee, T. H., 1998, The design of cmos radio-frequency integrated circuits.
[2] Crols, J. and Steyaert, M., 1998, Low-if topologies for high-performance analog front-ends for fully integrated receivers, *IEEE Trans. Circiut Syst.* Vol. II. Cas-45, pp.269–282, March 1998.
[3] Moore, G., 1965, Cramming more components onto integrated circuits, *Electronics* **38** (8), April 19.
[4] ITRS. http://public.itrs.net/
[5] Leroux, P. and Steyaert, M., 2005, LNA-ESD Co-Design for Fully Integrated CMOS Wireless Receivers, *International Series in Engineering and Computer Science*.
[6] Johns, D. and Martin, K., 1997, *Analog integrated circuit design*.
[7] Uyttenhove, K. and Steyaert, M., 2002, Speed-power-accuracy tradeoff in high-speed cmos adcs, *Transactions on Circuits and Systems II,* 247–257.
[8] Yao, L. and Steyaert, M., 2004, A Iv 88 dB 20 khz sd modulator in 90 nm cmos, *Proceedings of the ISSCC,* 80–81.
[9] Hajimiri, A. and Lee, T.H., 1998, A general theory of phase noise in electrical oscillators, *IEEE Journal of Solid-state Circuits,* 179–194.
[10] Craninckx, J. and Steyaert, M., 1995, A 1.8 GHz low phase-noise voltage controlled oscillator with prescaler, *IEEE Journal of Solid-state Circuits,* 1474–1482.
[11] Decock, W. and Steyaert, M., 2001, A cmos 10 GHz voltage controlled Ic-oscillator with integrated hing-q inductor, *Proceedings of the European Solid State Circuits conference,* 496–499.
[12] Maget, J., Tiebout, M. and Kraus, R., 2001, Comparison fo cmos vcos for units tuned by standard and novel varactors in standard 0.25/im technology, *Proceedings of the European Solid State Circuits conference,* 500–503.
[13] Kinget, P., et al., 2002, An injection locking scheme for precision quadrature generation, *IEEE Journal of Solid-state Circuits,* 845–851.
[14] Albulet, M., 2001, RF Power Amplifiers, Noble Publishing Associates, 10 May, 2001.
[15] Cripps, S. C., 1999, RF Power Amplifiers for Wireless Communications, Artech House Publishing, April, 1999.
[16] Cripps, S. C., 2002, Advanced Techniques in RF Power Amplifier Design, Artech House Publishers, 15 June, 2002.
[17] Kenington, P. B., 2000, High Linearity RF Amplifier Design. Artech House Publishers, 1 Otober, 2000
[18] Raab, H., Asbeck, P., Cripps, S., Kenington, P. B., Popovich, Z. B., Pothecary, N., Sevic, J. F. and Sokal, N. O., 2002, Rf and microwave power amplifier and transmitter technologies, *IEEE Transactions on Microwave Theory and Techniques*, 22–36.

Chapter 4.3

VECTOR PROCESSING AS AN ENABLER FOR AMBIENT INTELLIGENCE

Kees van Berkel, Anteneh Abbo, Srinivasan Balakrishnan, Richard Kleihorst, Patrick P.E. Meuwissen, and Rick Nas
Philips Research Eindhoven, Eindhoven University of Technology
{kees.van.berkel, anteneh.a.abbo, srinivasan.balakrishnan, richard.kleihorst, patrick.meuwissen, rick.nas}@philips.com

Abstract Ambient intelligence (AmI) applications can be very computationally demanding and at the same time require a highly flexible hardware–software implementation. Examples include fourth generation (4G) mobile communication and smart cameras. Programmable vector processing offers the required combination of flexibility and efficiency. Three Philips vector processors will be introduced: OnDSP, EVP, and Xetal, together with implementation details, benchmarking, and application.

Key words ambient intelligence; SIMD; smart cameras; software-defined radio; vector DSP; vector processing

1. INTRODUCTION

Aarts et al [1] define ambient intelligence (AmI) as "electronic environments that are aware of and responsive to the presence of people," or *embedded* environments that are *context aware*, *personalized*, *adaptive*, and *anticipatory* (cf. Table 4.3-1).

Such an environment can for example be a living room, a car, an office, an airport, or the world, depending on the user services (or user functions) at hand.

Using the examples of *4G wireless communication* and *smart cameras*, we will argue that ambient intelligence:

S. Mukherjee et al. (eds.), AmIware, 223–243.

Table 4.3-1. Ambient Intelligence [1].

Embedded	Many invisible distributed (networked) devices throughout the environment;
Context-aware	That know about your situational context;
Personalized	That can be tailored toward your needs and can recognize you;
Adaptive	That can change in response to you and your environment; and
Anticipatory	That can anticipate your desires without conscious mediation.

(1) May require a flexible architecture to support the context awareness, personalization, adaptiveness, and the ability to anticipate;
(2) Can be very computationally intensive, often requiring many tens of giga operations per second (GOPS). Furthermore, for consumer applications, these many GOPSes must be available at low costs, and—for practical use—at very low levels of power consumption.

The combination of the required computational efficiency (low costs and low power) and the required flexibility is a challenging one, involving complex trade-offs between dedicated hardware and software on a variety of programmable architectures.

In Section 2, vector processing is proposed as an architecture that offers a high computational efficiency for a large class of relevant algorithms. Specific architectures for vector processors developed at Philips Electronics are introduced in Section 3, including benchmarking with other processors. In Section 4, we return to system-level considerations and applications, including 4G wireless communication and smart cameras.

1.1. 4G Mobile Communication

Future mobile handsets will need to support multiple wireless communication links, potentially including 2G cellular, 3G cellular, wireless local-area network (WLAN), personal-area network (PAN), broadcast, and positioning. A layered structure of such a future network, adapted from Becher et al [2], is shown in Figure 4.3-1 and Table 4.3-2. These layers are to be integrated in a common, flexible, and seamless IP core network, supporting global roaming and a single access number per user. This requires both horizontal (intrasystem) and vertical (intersystem) handover, as indicated by the arrows.

A handset that supports one or more standards in each of the layers of Figure 4.3-1 allows its user to be always connected, anytime and anywhere. Moreover, with some additional "ambient intelligence" the user

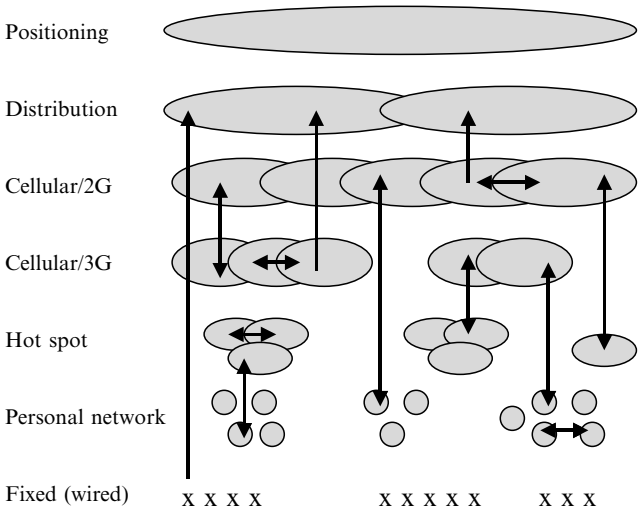


Figure 4.3-1. Layered structure of a 4G network.

can be *Always Best Connected* (ABC) [3] i.e., always connected by the best (combination of) radio links and link parameters, such as bit rates and Quality of Service (QoS) parameters. Here "best" is taken from the user's perspective, based on:

- The user's profile (e.g., known to the operator);
- The network characteristics and load;
- The conditions of the various radio channels involved;
- The terminal capabilities (e.g., screen size, available battery energy);
- The location of the user and the time of the day; and
- The requirements of the active applications.

Table 4.3-2. Layers of a future seamless wireless communication network.

Layer	Link range [$\log_{10}$ m]	Down/up link	Mobility	Standards (examples)
positioning	6–7	down	full	GPS, Galileo
distribution	5–6	down	full	DVB-T/H, DAB
cellular/2G	4–5	down, up	full	GSM, IS95, PHS
cellular/3G	3–4	down, up	full	UMTS, CDMA2000, TD-SCDMA
hot-spot	2–3	down, up	local	802.11 a/b/g/n, WiFi
personal	1–2	down, up	local	Bluetooth, DECT, UWB
fixed	0–1	down, up	none	POTS

The 4G network of Figure 4.3-1 gives rise to a set of formidable requirements on the flexibility of the architectures of future handsets:

(1) For each layer there exists a multitude of, often regional, standards. Designing or optimizing an architecture for each standard is very costly; a generic, multistandard architecture is preferred.
(2) Some handheld devices may even have to support multiple standards per layer (e.g., in a world phone). This implies that the architecture must support reconfiguration in the field;
(3) Individual standards typically evolve over the years toward higher bit rates, more features, and more services. For example, 3G cellular standards will need to support high-speed downlink packet access (HSDPA), and for WLAN multiple-antenna schemes are being studied (MIMO, IEEE 802.11n);
(4) For a given standard, new algorithms are continuously developed to improve performance (e.g., a lower bit-error rate or a more efficient spectrum usage). Ideally, it should be possible to adopt these improvements without redesign of the hardware; and
(5) Some applications may require multiple radios to be active simultaneously. Then, the architecture has to support a form of multitasking.

The combination of these requirements calls for a highly flexible architecture; one that can be programmed (or configured) late in the design or manufacturing process, or even in the field, possibly by downloading new software versions over the air interface.

In addition, the digital baseband processing required for these standards can be as high as 10 giga instructions per second (GIPS), measured on a conventional digital signal processor ([4], Figure 4.3-2), where 1 GHz (=GIPS) is roughly equivalent to 5 GOPS. Furthermore, the power budget is restricted to only a few hundred mW.

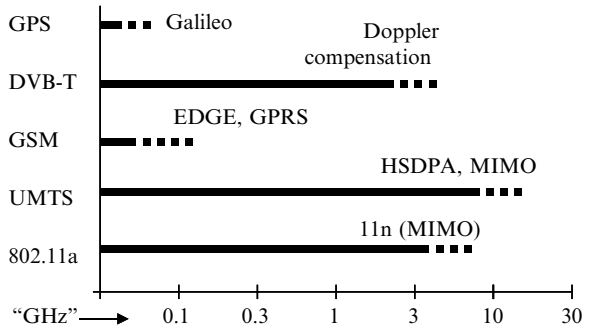


Figure 4.3-2. Load estimates for the baseband processing for various wireless standards, measured in MHz of a conventional digital signal processor.

1.2. Smart Cameras

Smart cameras are among the emerging new fields of electronics. They will play an important role in AmI, because they translate scenes into decisions. Smart cameras are devices that capture live video and process it autonomously inside the camera. Smart camera applications typically reduce the high-data-rate-input images or live video from the observed scene to low-data-rate output in the form of decisions or identification results. Among the examples that are worked on currently are person and object identification, gesture control, event recognition, and data measurement. More challenging applications are expected to come. It is important that these platforms are programmable since new applications emerge every month. The complexity (and possible error-proneness) of the algorithms and the fickleness of real-life scenes also strongly motivate complete programmability. The points of scientific interest for smart cameras are in the application areas, software, and IC development.

It seems like a daunting task to create programmable hardware that is able to process millions of pixels per second for complex decision tasks. However, this is possible by exploiting the inherent parallelism present in the various levels of image processing. A desirable ''silicon'' property of the resulting parallel architectures is that they are easily scaled up or down in performance and/or power consumption whenever the application changes. This not only lowers the need to develop new architectures from scratch for new vision application areas, it also allows the design team to use the same software suite with only some settings changed to include files. This significantly reduces the overall cost of vision solutions as they can be reused among the portfolio of the producer.

Two types of processors that play an important role in smart camera architectures are the single instruction multiple data (SIMD) massively parallel processor, and (one or more) general purpose DSPs [5, 6]. Enough has been written about general purpose DSPs, so we will mainly focus in this section on the merits of SIMD processors for the computationally demanding image processing tasks. After reading this section, it will be clear that they have unique and very clear merits from a silicon and algorithmic point of view for realistic IC implementations of programmable smart cameras for ambient intelligence.

2. VECTOR PROCESSING

Conventional microprocessors as well as digital signal processors execute instructions one at a time, and each instruction typically deals with a single scalar operation (e.g., loading an integer value from memory or

adding two integer numbers). In order to meet the required compute power, we need to apply parallelism on some scale, say, P. The two basic options are:

- *SIMD parallelism* (single instruction stream, multiple data stream): a single instruction can specify P identical operations on P (pairs of) operands. For example, load P values from memory, or add two rows of P integers each, elementwise, resulting in P sums. Making this SIMD parallelism explicit in an algorithm is commonly called *vectorization*.
- *MIMD parallelism* (multiple instruction stream, multiple data stream): P (identical) processors can execute P independent scalar instructions in parallel.

The SIMD has a number of advantages over MIMD. First, the silicon area and power consumption of storing instructions, fetching instructions from memory, caching instructions, decoding instructions, etc., is shared by P operations. Second, sequential execution of a single instruction stream is simple and well-understood. This results in highly efficient instruction scheduling, and in a low overhead of task switching.

Potential drawbacks of SIMD are the following. First, it is not clear whether the algorithms needed for signal processing and video/image processing can be vectorized. This is discussed in Section 2.1. Second, not all parts of an algorithm can be vectorized equally well, potentially resulting in an overall speed-up well below P. This phenomenon is described by Amdahl's Law, and is covered in Section 2.2. A number of SIMD architecture features that maximize generality, flexibility, and performance are discussed in Section 2.3.

Some vector processors are capable of operating on vectors of arbitrary size. The processor must then chop very long vectors in segments of length at most P. In the sequel we assume fixed length vectors.

2.1. Many Algorithms Can Be Vectorized

Despite the apparent restrictions of SIMD, many algorithms can be vectorized efficiently. Efficiently here means that during most instructions, P resources operate in parallel on P data items controlled by a single instruction. Table 4.3-3 contains a list of algorithms that the authors have vectorized for a variety of different applications.

We have also observed that the vectorization of these algorithms typically scales well with P (i.e., doubling P nearly doubles the throughput of the algorithm). There are, however, both limits and exceptions to this form of scalability. For example, an N-point complex FFT scales well up to $2N = P$, and not beyond.

Table 4.3-3. Algorithms that can be vectorized efficiently.

Communication algorithms	(Multi) media processing algorithms
rake reception	RGB2YUV conversion and v.v.
UMTS acquisition	(I)DCT
cordic	SAD (including bilinear interpolation)
(I)FFT	motion estimation
Fast Hadamard transform	video scaling
OFDM symbol (de)mapping	vertical peaking
16 QAM demapping	disparity matching
equalization	RGB rendering
symbol-timing estimation	color segmentation
interference cancellation	noise filtering (morphology)
joint detection (TD-SCDMA)	object filtering (i.e. aspect ratio)
Viterbi decoding [9]	color interpolation,
etc.	*etc.*

2.2. Amdahl's Law

The speed-up that can be achieved by SIMD parallelism is limited by the fraction of the program code that allows vectorization. This limitation is known as Amdahl's Law [10]. For example, when 90% of an algorithm can be sped up by a factor $P = 32$, the remaining 10% will dominate the execution time and the overall speed-up is less than a factor 8! This dependency of speed-up on the fraction of the code that can be vectorized is depicted in Figure 4.3-3.

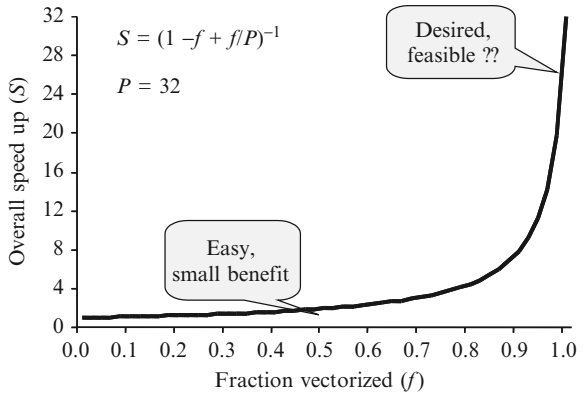


Figure 4.3-3. Amdahl's law.

Analysis of the algorithms of Table 4.3-3 revealed a large fraction of the nonvectorizable parts of those algorithms comprising pointer arithmetic (address calculations), regular scalar operations, and simple loop control.

2.3. More than SIMD

To further increase the amount of parallelism, we propose to pack multiple SIMD operations in a single instruction. For example, a single instruction could comprise both a load of a vector of P values *and* the addition of two previously loaded vectors. Packing multiple operations in a single instruction is known as very long instruction word (VLIW) [10]).

Furthermore, to counter the implications of Amdahl's law, such a VLIW instruction must also contain (multiple) address calculations [11], support for zero-overhead looping [11], as well as multiple other scalar operations. All parallelism combined can be visualized as in Figure 4.3-4 In this example, each little square box denotes the hardware resource for a simple scalar operation (e.g., add, multiply, load, etc.). Here we have assumed $P = 16$, and as many as 6 vector and 6 scalar operations. The boxes labeled with a $\oplus$ denote so-called address-computation units, capable of operations, such as post increment. Note that the VLIW parallelism is orthogonal to the scalar + vector parallelism.

3. ONDSP, EVP, AND XETAL

Philips Semiconductors and Philips research have over the recent years developed three different vector processors, for different application domains:

- OnDSP targeting WLAN;
- EVP for 3G wireless communication and beyond; and
- Xetal for video and image processing.

Table 4.3-4. Cycle counts for a 64-point complex FFT.

Processor	Ref	Code	Clock cycles	SIMD
EVP_{16}	[4]	opt	64	16×16
OnDSP	[4]	opt	160	8×16
Tigersharc	[7]	opt	174	$2 \times 8 \times 16$
VIRAM		opt	357	16×32
TMS320C6203	[14]	opt	646	N.A.
Altivec MPC7447	[14]	opt	956	8×16
Carmel 10xx	[14]	otb	5568	N.A.
AMD K6-2E+/ACR	[14]	otb	10,751	N.A.

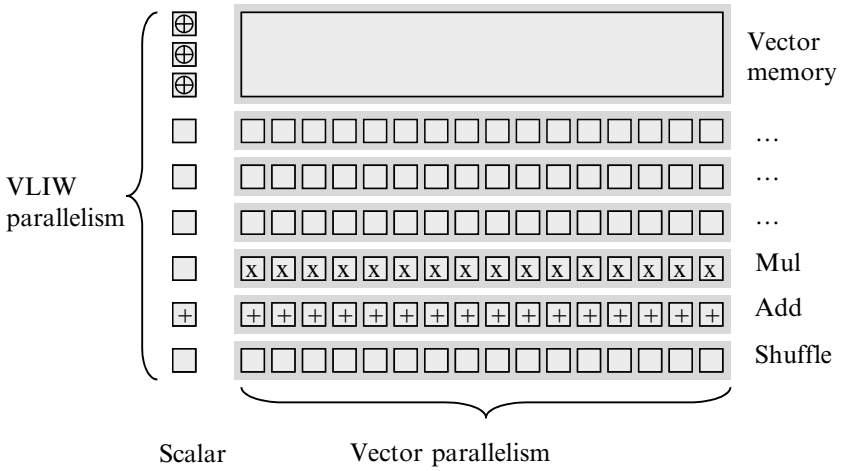


Figure 4.3-4. Overall parallelism = (scalar parallelism + vector parallelism) × VLIW *parallelism.*

3.1. OnDSP

The OnDSP vector processor is a key component of several multi-standard programmable Wireless LAN baseband product ICs [13]. The OnDSP architecture, with a vector size of $P = 8$ (128 bits), is depicted in Figure 4.3-5. A single VLIW instruction can specify a number of vector operations (e.g., load/store, ALU, MAC, address calculations, and loop-control). OnDSP supports a couple of specific vector instructions, including

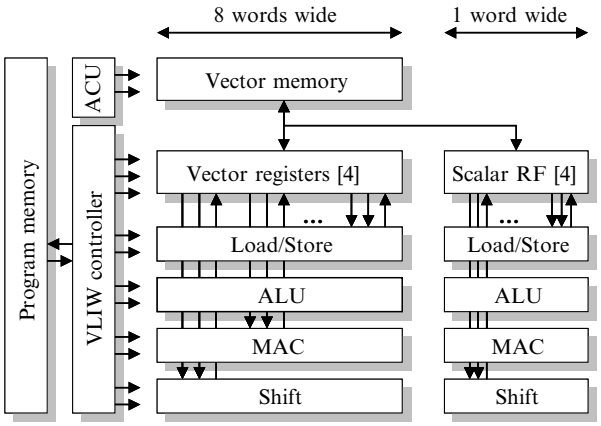


Figure 4.3-5. The OnDSP architecture.

word insertion/deletion, sliding, and gray coding/decoding. Addresses to the vector memory must be a multiple of *P*.

Table 4.3-4 illustrates the performance of OnDSP in comparison to several well-known processors for a 64-point FFT, one of the basic kernels of WLAN 802.11a. Note that TigerSharc [7], VIRAM, and AltiVec [12] are also vector processors. Only the Carmel and AMD processors show "out-of-the-box" (otb) performance. For all other processors, the FFT algorithm and code has been optimized for the processor (opt).

In a 0.12 μm CMOS process, OnDSP measures about 1.5 mm^2 (250 kgates), runs 160 MHz (worst-case commercial), and dissipates about 0.8 mW/MHz including a typical memory configuration. A macroassembler is used for VLIW scheduling, although optimization by hand is used for critical code.

3.2. EVP

The embedded vector processor (EVP)[4] is a productized version of the CVP [8]. Although originally developed to support 3G wireless communication standards, the current architecture proves to be highly versatile. Care has been taken to cover the OnDSP capabilities for OFDM standards. The EVP architecture is depicted in Figure 4.3-6. The main word width is 16 bits, with support for 8-bit and 32-bit data. The EVP supports multiple data types, including complex numbers.

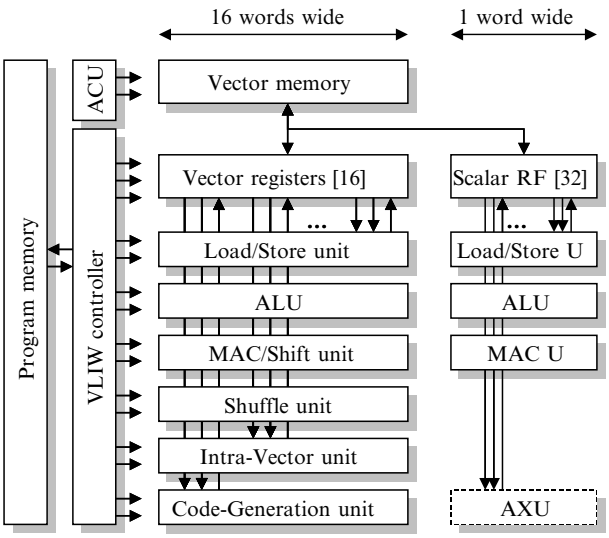


Figure 4.3-6. The EVP architecture.

The SIMD width is scalable, and has been set to $P = 16$ (256 bits) for the first product instance called EVP_{16}. The maximum VLIW-parallelism available equals 5 vector operations *plus* 4 scalar operations *plus* 3 address updates *plus* loop-control. Specific vector operations include the following:

- Shuffle operation can be used to rearrange the P elements of a single vector according to an arbitrary pattern, also a vector of length P;
- Intravector operations can be used to add (or take the maximum of) the P elements of a single vector, possibly split in M segments of P/M elements each, with M a power of 2; and
- CDMA-code generation can be used to generate P successive complex code chips in a single clock cycle.

Programs are written in EVP-C, a superset of ANSI-C. Programs written in plain C will be mapped on the scalar part of the EVP, and hence will not utilize the vector operations. The EVP-C extensions include vector data types and so-called *function intrinsics* for vector operations, all in a C syntax. The EVP-C compiler takes care of register allocation (scalar and vector registers) as well as VLIW instruction scheduling (scalar and vector operations combined). The EVP tool flow further comprises an EVP-C host-emulation library, a linker, a bit-true/cycle-true simulator, a profiler, and an integrated debugger.

Table 4.3-5 illustrates the performance of EVP in comparison to several well-known processors when de-spreading a rake finger, one of the basic kernels of CDMA standards such as UMTS.

In a 90 nm CMOS process, the EVP_{16} core measures about $2\,mm^2$ (450 k gates), runs at 300 MHz (worst-case commercial), and dissipates about 0.5 mW/MHz (core only) and 1 mW/MHz, including a typical memory configuration. These numbers are based on gate-level simulations of annotated netlists.

3.3. Xetal

Xetal is a high-performance DSP originally developed for early-vision processing, handling the computationally intensive pixel manipulations [6].

Table 4.3-5. Load [MHz] of a UMTS-FDD rake finger.

Processor	Ref	Load [MHz]	Arithmetic resources
EVP_{16}	[8]	0.5	$16 \times$ (MAC + ALU + PN gen.)
Tigersharc	[7]	1	$2 \times 8 \times$ (MAC + ALU)
4 UMTS DP	[15]	6	$4 \times$ (MAC + ALU + PN gen.)
UMTS DP	[15]	25	$1 \times$ (MAC + ALU + PN gen.)
TI C62	[16]	40	
Carmel	[15]	125	2MAC + ALU
TI C54x	[15]	300	1 MAC/ALU

Currently, Xetal is applied in different parts of the video processing chain: from front-end image enhancement to the final display rendering. The Xetal block diagram is shown in Figure. 4.3-7. The linear processor array (LPA) is the major workhorse with 320 simple RISC-like processing elements (PEs). Each PE has its own accumulator and flag storage units. There are 32 line-memories each with 320 data elements used as temporary storage area. The conversion from sequential video stream to parallel vector data and vice-versa takes place in the sequential I/O memories. Xetal is programmable in a C-like language with extensions to accommodate, among other things, the vector data type.

The current Xetal chip is realized in a 0.18 μm CMOS process with a die size of 20 mm^2. The chip runs at 25 MHz and provides up to 8 GOPS of performance with a maximum power dissipation of 2.5 W. In fact, most algorithms operate on data that shows a high degree of correlation, and thus the actual power dissipation is usually quite less. Depending on the application demand, the operating clock frequency and supply voltage can be varied to optimize performance efficiency. In addition to this, the architecture can easily be scaled up or down with respect to the number of processing elements [17]. In Figure 4.3-8, the performance scaling is shown for a maximum operating frequency of 50 MHz. The highlighted curve corresponds to a design with 320 PEs that can deliver up to 16 GOPs.

The impact of varying the operating voltage and frequency of a given design is shown in Figure 4.3-9. The improved GOPS/Watt figure at lower

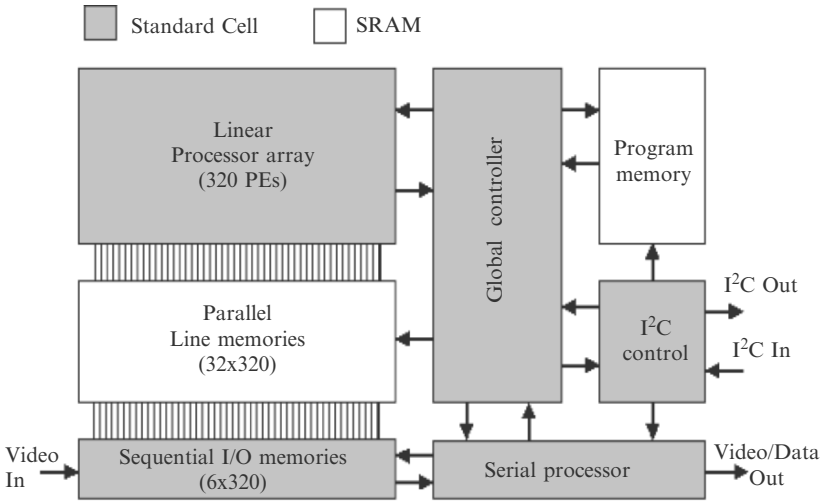


Figure 4.3-7. Xetal architecture.

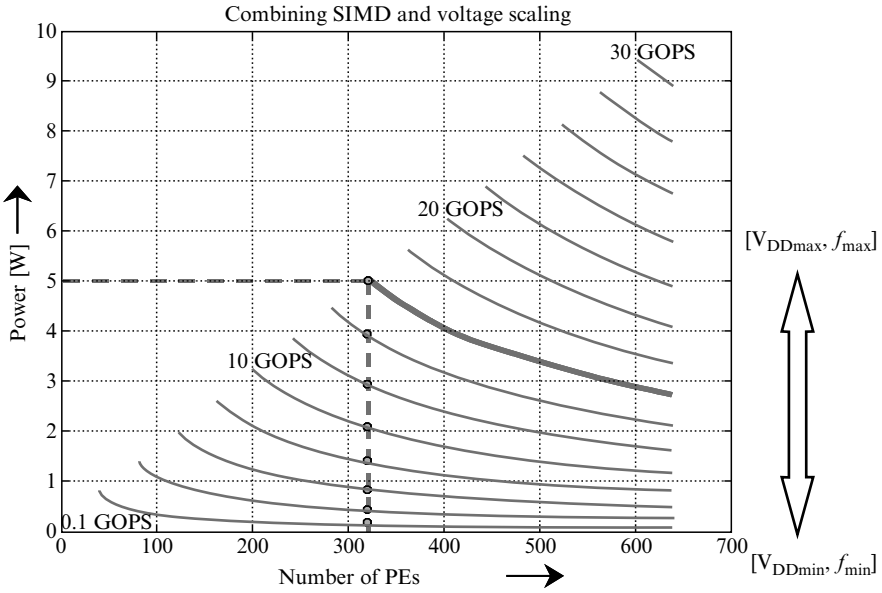


Figure 4.3-8. Xetal power consumption as a function of the number of processing elements, supply voltage (*V*dd), and clock frequency.

operating points is attributed to the nonlinear relationship between supply voltage and operating speed of logic cells.

In order to see the benefits of SIMD-based processing platforms, a number of performance metrics are given in Table 4.3-6. Although, Pentium belongs to a different class of processors that handle sequential

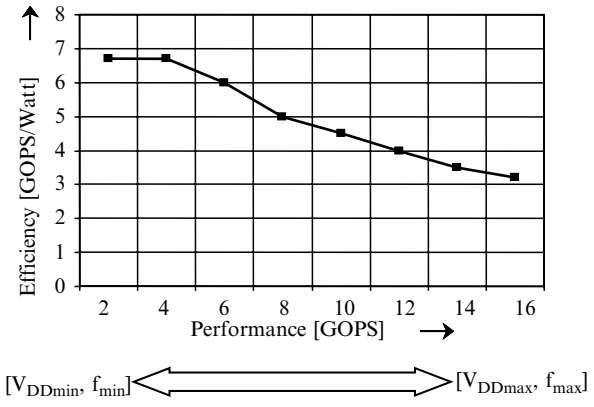


Figure 4.3-9. Xetal efficiency as a function of performance when supply voltage and frequency are varied.

Table 4.3-6. Raw computational performance of various processor types to highlight the merits of SIMD parallel processing.

		Pentium4	IMAP-CE 128 PEs	Xetal 320 PEs
clock frequency	Hz	2.4G	100 M	25 M
peak performance	GOPS	6	12.8	8
size	mm^2	131	121	20
bandwidth	Gb/sec	58	204	160
peak power cons.	W	59	4	2.5
power efficiency	GOPS/W	0.1	3.2	3.2
area efficiency	$GOPS/mm^2$	0.045	0.2	0.4

computations, it is clear that there is a lot of overhead in terms of area and power dissipation per unit of operation. The GOPS figures should be read with care since in IMAP-CE an operation is a single action performed on a single bit, whereas in Xetal a multiply-accumulate action on a 10-bit data is considered an operation.

4. SYSTEMS AND APPLICATIONS

Vector processing combines a very high efficiency with a moderately high flexibility. Nevertheless, in a complex system an optimal architecture will generally also include other components (Figure 4.3-10).

- Dedicated hardware consumes less dynamic power than any programmable architecture. When flexibility is not required for certain

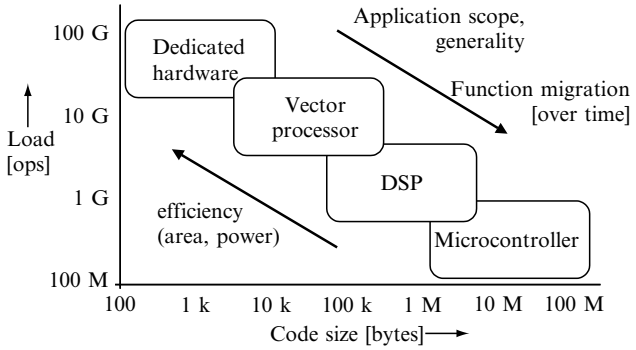


Figure 4.3-10. A HW architecture for a complex system that is both flexible and efficient may comprise dedicated hardware, DSP(s), microcontroller(s), and vector processor(s).

parts of the system that are active most of the time, dedicated hardware is the preferred option.

- Conventional DSPs can also efficiently support algorithms that are intrinsically nonvectorizable (e.g., control code, Huffmann decoding). Also, porting of legacy DSP code to a vector DSP may not always be attractive.
- Microcontrollers, such as ARM and MIPS, are highly versatile and flexible, with abundant tool support and available middle ware. For all computing, except number crunching, they are to be preferred, given their maturity and excellent tool support.

In Figure 4.3-10, dedicated hardware, vector processors, DSPs, and microcontrollers are ordered along the axes of code size (bytes) and computational load [ops] for a complex product, such as a high-end mobile phone. The more flexible category will cover a substantially larger fraction of the system's complexity (code size), but a comparably smaller fraction of the load (operations per second). Note that even dedicated hardware can be weakly programmable, by writing a number of configuration registers. The general trend is that functions migrate over time toward a more flexible category, usually from hardware to software.

4.1. 4G Mobile Communication

As we have seen in Section 1.1, a 4G mobile handset may have to perform 10+ GHz of conventional DSP compute power in order to support the increasing number of wireless communication standards. This compute power is required for the so-called baseband processing. When the radio functionality of the different standards can be expressed in software on a flexible compute platform, it is increasingly common to refer to such an architecture as a *software defined radio* [18, 19].

In Figure 4.3-11,the baseband section is split into three stages: filters, modem, and codec. As is argued in [4], these three sections have quite different characteristics and have quite different demands on flexibility. More specifically, the modem stage involves very different algorithms for the different standards, offers most benefits for algorithmic improvements, and is most often affected by revisions of these standards. It is this stage where the flexibility of vector processors offers most benefits.

Figure 4.3-12 gives the load on the EVP_{16} for the modem stage of a number of standards. Note that with a maximum EVP_{16} clock frequency of 300 MHz, most standards leave quite a lot of headroom on the EVP. This available headroom can be used:

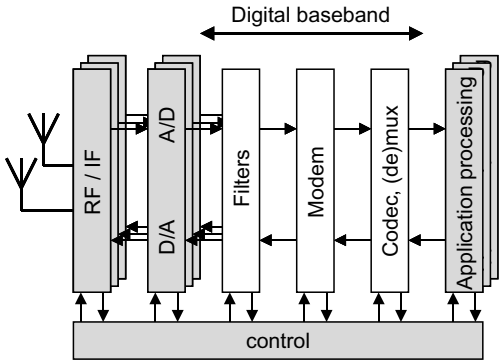


Figure 4.3-11. Stages in a software-defined radio.

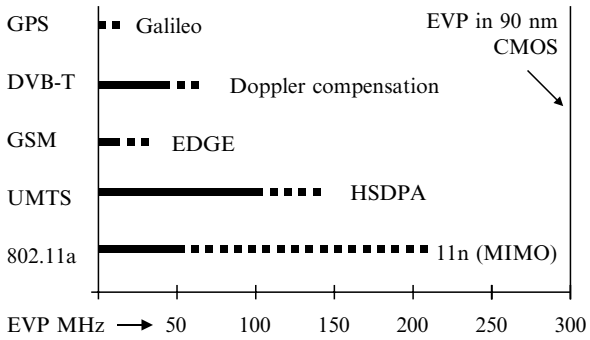


Figure 4.3-12. Estimated EVP_{16} load numbers for the modem stage of various receivers.

- To introduce improved but more demanding algorithms;
- To scale the supply voltage to reduce power consumption (cf Section 3.3); and
- To run multiple standards simultaneously.

Although it may be possible in principle to map also other functionality on the EVP, it is more practical and more convenient to introduce other components:

- Dedicated, weakly configurable hardware can be used for the filter stage and some parts of the codec stage (e.g., a Viterbi decoder);
- A conventional DSP is used for irregular, less-demanding parts of the codec stage, and also to support legacy 2G standards; and
- A conventional microcontroller can be used for protocol processing and the user interface.

4.2. Mapping Vision Applications on Smart Cameras

The algorithms in the application domain of smart cameras can be classified into three levels: *low*, *intermediate,* and *high*. (See Table 4.3-7.)

The *low- or early-image processing level* is associated with typical kernels like convolution and data-dependent operations using a limited neighborhood of the current pixel. In this part, often (the initial steps towards) pixel classification is performed. Because every pixel can be classified in the end as "interesting," the algorithms per pixel are essentially the same. We exploit this inherent data parallelism by operating on more pixels per clock cycle, using an SIMD architecture [20, 21]. As explained in Section 2, SIMD architectures issue the same instruction on all data items in parallel, which lowers the overhead of instruction fetch, decoding, and data access. This leads to economical solutions to meet the high performance and throughput demanded at this task level.

In the *intermediate level,* measurements are performed on the objects found to analyze their quality or properties in order to make decisions on the image contents. It appears that SIMD type of architectures can do these tasks, but they are not very efficient because only part of the image (or line) contains objects and the SIMD processors are always processing the entire image (or line). A general purpose DSP is often more appropriate, assuming that the performance demands are met. If the performance needs to be increased, a viable way is to use the property that similar algorithms are performed on multiple objects, leading to task-parallel object processing on different processors.

Finally, in the *high-level* part of image processing, decisions are made and forwarded to the user. General-purpose processors are ideal for these tasks because they offer the flexibility to implement complex software

Table 4.3-7. A classification of image processing algorithms, examples, and characteristics.

Level	Classification	Examples	Characteristics
high	decision-making real-time OS networking		- some decision tasks - complex processing
intermediate	object processing	- shape analysis - shape coding - segmentation	- lots of objects: 1 ... 300 k objects/second - similar processing per object
low	(early) pixel ... processing	- image improvement - edge enhancement - convolution kernels	- lots of pixels: 1 ... 1000 M/second - similar processing per pixel

tasks and are often capable of running an operating system and doing networking applications.

A complete smart-camera architecture has all these three components as shown in Figure 4.3-13. Here, Xetal handles the part of SIMD processing doing the low-level tasks, while TriMedia functions as the general-purpose DSP that takes care of the intermediate and high-level tasks.

An illustrative application mapping, where we can easily indicate the three processing levels, is face detection and recognition as shown in Figure 4.3-13. Here, the low-level part of the algorithm classifies each pixel as belonging to a face or not (face detection). This is mapped on the SIMD processor. The intermediate-level determines the features and identifies each detected face object. Finally, in the high-level part of the algorithm, the decision is taken to open a door, to sound an alarm, or to start a new-guest program. The latter two levels of tasks are performed by the general-purpose DSP. More information on this application can be found in [6].

5. CONCLUSION AND CHALLENGES AHEAD

Vector processing can be very silicon-area and power efficient, because the energy and area involved in instruction storage, fetching, decoding, etc., can be shared by many identical operations. For algorithms that can be fully vectorized, this can give a speed-up of P, the SIMD width of the

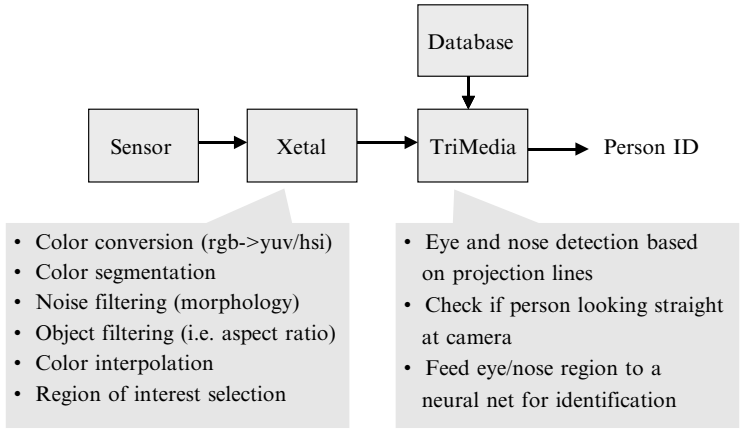


Figure 4.3-13. Face recognition actually has two parts: detection and recognition. The complete detection part is perfectly mapped to the SIMD processor doing the low-level operations. The recognition part is dealt with by the DSP.

processor. For algorithms that are not fully vectorizable, overall efficiency can be made close to P by means of additional parallelism: scalar operations, address computations, and control operations, such as looping. A further increase of efficiency can be obtained by adding VLIW parallelism, on top of the scalar *plus* vector parallelism. By carefully combining these forms of parallelism, and by including a rich mix of vector instructions, vector processing can be both remarkably versatile and very efficient.

The vector processors OnDSP, EVP, and Xetal illustrate this claim on a wide range of algorithms. Nevertheless, for a complex system, such as software-defined radio or a smart camera, it is generally advantageous to include dedicated hardware, a general-purpose microcontroller, and sometimes also a conventional DSP.

In order to further increase the application scope and in order to aim at yet higher levels of computational performance, we intend to address the following challenges:

(1) *Vectorizing compilers*. Making vector parallelism explicit (in terms of e.g., EVP-C function intrinsics) is quite doable, but it would be much more attractive if a compiler could extract vector parallelism from a plain C or MatLab text. Although there is a long tradition in vectorizing compilers (Wolfe [22]), their efficiency is still a concern. Also, it turns out to be difficult to take full advantage of features, such as a shuffle operation.

(2) *Seemingly nonvectorizable algorithms*. For some algorithms, it is doubtful whether a vectorizing compiler will ever be able to extract the vector parallelism. Quite ingenious program transformations may be required to parallelize algorithms, like IIR, FFT, and variable-length coding.

(3) *Scalable vector shuffle*. The shuffle operation, of which vector permutation is a special case, is a remarkably useful operation. It is critical to many algorithms, like FFT, motion estimation, as well as memory-access operations, such as scatter-gather [10]. Unfortunately, the size, delay, and energy of a general shuffle scales poorly with the SIMD width. Further research is required into shuffle operations that do the most common shuffle operations in a single clock cycle, and the less frequent ones in multiple cycles still using a reasonable silicon area.

(4) *Architecture scalability*. In order to be able to scale the processor architecture from, say, $P = 16$ to $P = 1024$, other scalability challenges must also be addressed, including wire-dense layouts, and various circuit issues.

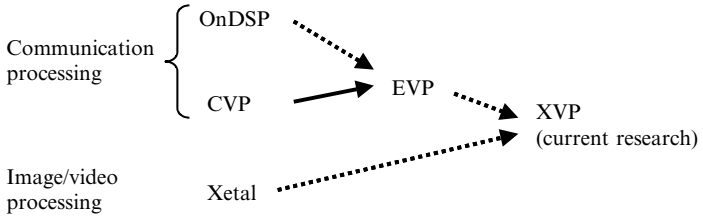


Figure 4.3-14. Evolution of vector processors at Philips Electronics.

(5) *Domain-specific vector operations.* For critical algorithms in media or communication applications, it may pay off to introduce application-specific instructions (e.g., the CDMA-code generation in the EVP.)

The above challenges are being addressed in a study towards the XVP processor (Figure 4.3-14). In parallel, we aim at further extending the application scope of vector processing to mobile communication, future television, and ambient intelligence.

REFERENCES

[1] Aarts, E., Harwig, H. and Schuurmans, M., 2001, Ambient intelligence, *in The Invisible Future*, McGraw-Hill.
[2] Becher, R., et al., 2001, Broad-band wireless access and future communication networks, Proc. of the IEEE, **89**(1), 58–75.
[3] Gustafsson, E. and Jonsson, A., 2003, *Always best connected*, *IEEE Wireless Communications*, **10**(1), 49–55.
[4] Kees van Berkel, C. H., et al., 2004, Vector processing as an enabler for software-defined radio in handsets from 3G + WLAN onwards, *Proc. of the 2004 Software Defined Radio Technical Conference*, **Volume B**, 125–130, November 16–18, 2004, Scottsdale, Arizona.
[5] Abbo, A. A. and Kleihorst, R. P., A programmable smart-camera architecture, *ACIVS2002*, September 2002, Gent, Belgium.
[6] Fatemi, H., Kleihorst, R. P. and Corporaal, H., Real-time face recognition on a smart camera, *ACIVS2003*, September 2003, Gent, Belgium.
[7] Friedman, J. and Greenfield, Z., 2000, The TigerSharc DSP architecture, *IEEE Micro*, **20**(1), 66–76.
[8] Kees van Berkel, C. H., et al., CVP: A programmable co vector processor for 3G mobile base-band processing, *Proc. World Wireless Congress*, May 2003.
[9] Nur Engin, et al., Viterbi decoding on a co-processor architecture with vector parallelism, *IEEE Workshop on Signal Processing Systems* (*SIPS'03*), Seoul, August 2003.

[10] J.L Hennessy, D. Patterson, *Computer Architecture*, 3rd edition, Morgan Kaufmann Publishers, 2003.
[11] Lapsley, P., et al., 1994–1996, *DSP Processor Fundamentals*, Berkeley Design Technology, Inc.
[12] Gwennap, L., 1998, G4 is first PowerPC with Altivec, *in Microprocessor Rep*., November 16, 1998, 17–19.
[13] Kneip, J., et al., 2002, Single chip programmable base-band ASSP for 5 GHz wireless LAN applications, *IEICE Trans. Electron,*.**E85-C**(2), 359–367.
[14] Embedded Microprocessor Benchmark Consortium. http://www.eembc. hotdesk.com/
[15] Walther, U., et al., 1999, New DSPs for next generation mobile communications, *Global Telecommunications Conf. – Globecom '99*, 2615–2619.
[16] Dent, P. R., 2000, W-CDMA reception with a DSP-based software radio, *3G Mobile Communication Technologies, Conference Publication No. 471*, 311–315.
[17] Abbo, A. A., Kleihorst, R. P., Choudhary, V. and Sevat, L., *Power consumption of performance-scaled SIMD processors, PATMOS-2004*, Santorini, Greece, September 2004.
[18] Software Defined Radio Forum www.sdrforum.org
[19] Glossner, J., et al., 2003, A software-defined communications baseband design, *IEEE Communications Magazine*, 120–128.
[20] Jonker, P., 1994, Why linear arrays are better image processors, *Proc. 12th IAPR Conf. on Pattern Recognition*, 334–338, Jerusalem, Israel.
[21] Hammerstrom, D. W. and Lulich, D. P., 1996, Image processing using one-dimensional processor arrays, *IEEE Proceedings*, **84**(7), 1005–1018.
[22] Wolfe, M., 1996, *High Performance Compilers for Parallel Computing*, Addison-Wesley Publishing Company.

Chapter 4.4

XTREME LOW POWER TECHNOLOGY DEVELOPMENT USING A VIRTUAL DESIGN FLOW
Enabling Technologies for Ambient Intelligence Applications

P. Christie, R. K. M. Ng, G. Doornbos, A. Heringa, A. Kumar, and V. H. Nguyen
Philips Research Leuven
{*p.christie*, *ranick.ng*, *gerben.doornbos*, *anco.heringa*, *aatish.kumar*, *viet.nguyenhoang*}@*philips.com*

Abstract We outline the issues relating to the development of an xtreme low power (XLP) process option, and its associated logic and memory design styles, targeted for ambient intelligence (AmI) and medical implant applications. The most obvious route to low switching energies is by reducing the supply voltage. For an economically viable process, however, we must also ensure that the resulting clock rates are high enough to entice designers to use the process for real commercial applications. This requires that the impact of XLP technology choices for both the front-end devices and back-end interconnect be assessed at the level of critical path delays and memory access times. Moreover, these technology choices must be selected to be optimum for a range of typical applications designed using the XLP process, and not just for a single benchmark test vehicle. This has led to the development of a virtual design flow for technology assessment, which integrates device, interconnect, logic, and memory design options at the system level. Examples, of how this new design flows, have been used to assess new XLP front-end and back-end technology options will be presented. Data on low voltage system-level operation of standard bulk, metal gate, high-k, and multi-gate CMOS will be discussed.

Keywords CMOS; interconnect; simulation

S. Mukherjee et al. (eds.), AmIware, 245–261.

1. INTRODUCTION

For many ambient intelligence (AmI) applications, the limiting factor is the circuit power dissipation. Specifically, the dynamic switching energy is largely determined by the voltage to which the distributed capacitance of the interconnect must be charged. This voltage is, in turn, determined by the trade-offs selected when designing the device for a target static leakage current and drive strength. The dynamic power dissipation associated with this switching energy is then defined by the delay of data signals through critical paths that may well contain both logic and memory. There is, therefore, an intimate coupling between front-end and back-end process design, which is very difficult to evaluate without access to a complete layout flow (place/route and timing analysis) and compact model support (SPICE). At Philips Research, we have been developing a virtual design flow (VDF), which attempts to overcome many of these difficulties in order to allow novel CMOS architectures to be rapidly analyzed within a realistic interconnect environment. The analysis method is based on the same timing and power analysis algorithms employed by designers of ASIC chips. Such a capability is crucial for developing a new generation of XLP fabrication processes, since they cannot be based on a simple shrink of any existing process.

Our approach has been to design a representative range of device architectures with an emphasis on low operating voltage (V_{dd}). Data from the following devices will be presented: standard CMOS, standard CMOS with high-k gate dielectric, standard CMOS with metal gate, with the more exotic architectures being represented by fully depleted silicon-on-insulator (FDSOI) devices, and double gate (DG) devices. Since they are quite different from standard CMOS architectures, the geometries of the FDSOI and DG devices are sketched in Figure 4.4-1.

Due to an emphasis on low power, device design begins by specifying three target-static leakage levels per micron of gate width: standard leakage (SL) 1000pA/μm, low leakage (LL) 100pA/μm, and ultralow leakage (ULL) 10pA/μm. For each of these leakage specifications, an associated target for the device drive current is defined by using standard CMOS as the reference device. For devices other than the reference V_{dd} of 1.2 V for the standard CMOS device, V_{dd} is treated as a variable to ensure that all devices have an identical I_{on}. This procedure allows the performance (critical path delay) differences between each of the device architectures to be more easily interpreted. This, in turn, permits a much simpler analysis of the differences in

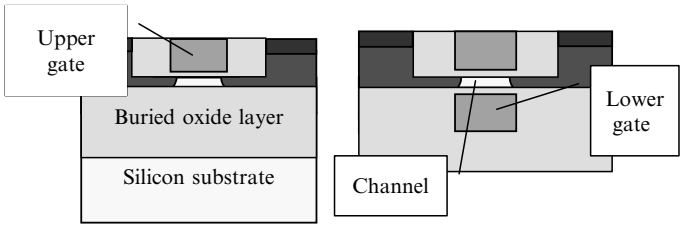


Figure 4.4-1. Simplified architectures of the fully depleted silicon-on-insulator and double gate architectures used in this study. The FDSOI device on the left is fabricated on top of a buried oxide layer, which provides excellent electrical isolation. The double gate device on the right possesses two gate electrodes above and below the channel.

switching energies, since they are largely determined by the value of V_{dd}^2 for a given circuit instance.

The VDF begins by embedding the devices within simple standard cell and 6T SRAM layouts, using mixed-mode technology computer-aided design (TCAD). This enables timing information to be extracted using standard timing analysis formats, based on actual timing algorithms used within the design community. Of course, actual timing analysis requires knowledge of the layout of a particular instance of a circuit. To overcome this we have developed a statistical virtual layout procedure based on simplified models of place and route algorithms. Due to its statistical nature, virtual layout does not attempt to analyze the timing of an individual circuit instance, but rather yields more pertinent information concerning the timing of an ensemble of circuits that could be fabricated using the XLP process under development. The essential point of pseudo-layout is that it can provide statistical distributions of wire lengths so that the properties of different back-end interconnect architectures can be coupled to the timing matrices extracted from the embedded device architectures.

In the next section, we describe the device design procedure in more detail and this is followed in Section 3 by a discussion of how timing matrices are extracted from devices embedded within logic cells. Section 4 provides an overview of the VDF method with an emphasis on how the delay metrics of other commonly used methods (intrinsic gate delay, ring oscillator delay, etc.) compare with our VDF method. The analysis of 6T SRAM is covered in Section 5 and data are presented which compare 6T SRAM noise margins, write margins, read currents, and read-out delays for the same device architectures considered in the logic analysis. The conclusions of our investigations are summarized in Section 6.

2. DEVICE DESIGN PROCEDURE

A major factor in determining the level of off-state transistor leakage is pocket or halo doping (see [1] or [2]). Halo doping is used to remedy the observed increase in subthreshold leakage with decreasing channel length. Proper tuning of this halo doping results in a threshold voltage roll-up for decreasing channel length (with low drain biases). By adapting the halo doping, the threshold voltage and the static leakage are determined. This method has been used in our design procedure to set subthreshold leakage at a specified value. For FDSOI and DG devices, the parameter determining the subthreshold leakage is the work function of the metal gate and not the channel doping. Therefore, for these two devices, the gate work function is varied instead of doping profile. For all devices, the corresponding drive current is calculated by applying the given supply voltage to the gate ($V_{\text{gate}} = V_{dd}$).

The large set of device characteristics needed for characterization were generated with the device modeling package Medici, using the modified local density approximation, which takes into account quantum mechanical effects at the oxide–silicon interface. Gate leakage and junction leakage have been considered separately and were not directly included in the device simulation. For devices other than standard CMOS, the supply voltage has been varied from its default value of 1.2 V, to get common settings for leakage (I_{off}) and drive strength (I_{on}) across all device architectures. For the alternative standard CMOS structures (hi-k, metal gate), the same doping architecture as for the standard CMOS was chosen; again only the halo doping value was tuned. This leads to the following tuning procedure:

(1) Start with a reasonable initial device structure with supply voltage of 1.2 V;
(2) Tune the doping until I_{off} is at the specified value;
(3) Adapt the supply voltage until I_{on} is at the specified value; and

Check whether I_{off} is still close enough to the specification; if not, go to step 2.

In the comparison of different architectures, a reference value for the current of both NMOS and PMOS has been set with reference to standard CMOS. Gate dielectric leakage current has been calculated based on calibrated parameters extracted from the best experimental results obtained with nitrided oxides. However, for standard CMOS, the gate leakage was above 10% of the transistor leakage, and thus these could not be considered as viable LL or ULL devices. However, the drive currents obtained for these standard CMOS devices have been used as the target

currents for the other devices. These are given in Table 4.4-1. For all devices, gate length is 45 nm and, for the NAND gates, we selected 405 nm for the NMOS gate width and 485 nm for the PMOS gate width. For the inverter, the widths were chosen to be 405 and 710 nm, for the NMOS and PMOS, respectively. Other relevant lengths for the NAND cell are shown in Figure 4.4-2.

With these leakage and I_{on} specs the calculated supply voltages for each device are shown in Table 4.4-2. The value is an average for the I_{on} of the PMOS and NMOS devices. The table entries for Bulk are only as a reference; their gate leakage was above specification. Design solutions for the ULL versions of the metal gate and high-k were also out of specification (OS) due to the high channel doping needed, inducing excessive junction leakage.

3. LOGIC CELL TIMING EXTRACTION

From a design perspective, the detailed physical characteristics of the different device architectures may be characterized by two simple timing

Table 4.4-1. Target drive current values.

Leakage	NMOS I_{on} current	PMOS I_{on} current
Standard	660 μA/μm	325 μA/μm
Low	549 μA/μm	288 μA/μm
Ultralow	452 μA/μm	234 μA/μm

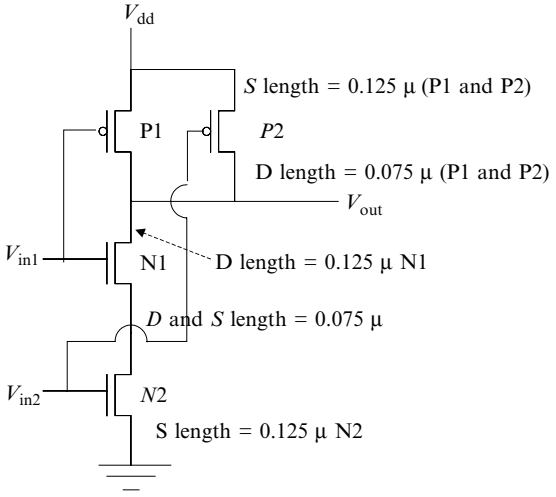


Figure 4.4-2. Layout of NAND cell used in this study.

Table 4.4-2. Supply voltages for SL, LL, and ULL for different architectures.

Leakage	Bulk	MG	Hi-K	FDSOI	DG
Standard	1.2	1.14	0.93	1.13	0.76
Low	1.2	1.14	0.96	1.13	0.76
Ultralow	1.2	OS	OS	1.14	0.78

matrices. These matrices define the delay and the output transition time of a given logic cell as a function of input transition time and the capacitive load at the output of a cell. The generation of these timing matrices is done via mixed mode TCAD simulation. Mixed mode simulation allows TCAD devices to be combined with external circuit elements, such as load capacitors and voltage sources. For an inverter, two pairs of timing matrices must be generated: one pair for a falling input and one pair for a rising input. For a NAND cell, four input logic transitions are defined: rising and falling signals on input A with input B constant, and rising and falling inputs on input B, with input A constant. Thus, four pairs of timing matrices must be generated. The delay and output transition time matrices for each logic transition were derived for the following transition times and load capacitances, respectively: 10, 100, 200, 500, and 1000 ps, and 1, 50, 100, 200, and 500 fF. This resulted in 5×5 matrices. In addition, to the delay and output transition time, the input capacitance of the library cell and the power consumption of the cell during switching, were also extracted.

4. SYSTEM-LEVEL LOGIC SIMULATION

When large cell arrays are laid out within an automated design flow, nearest-neighbor placement of connected cells is extremely difficult to achieve due to the level of cross-linking between chain cells and other cells in the array. To model the statistical distribution of connecting lengths, we employ a virtual placement method, based on the procedure of Davis et al. [3], and modified to include region II of Rent's rule [4]. This was used to generate an estimate of the wire length distribution for a cell array consisting of 230,400 cells. Wires from the distribution were then virtually routed [5] into routing layers, with shortest wires allocated to the lowest layers until full, and then moving up through the interconnect layers, until all wires are allocated to layers (assuming a routing channel utilization of 50%). The resulting virtual wire length distribution is shown in Figure 4.4-3.

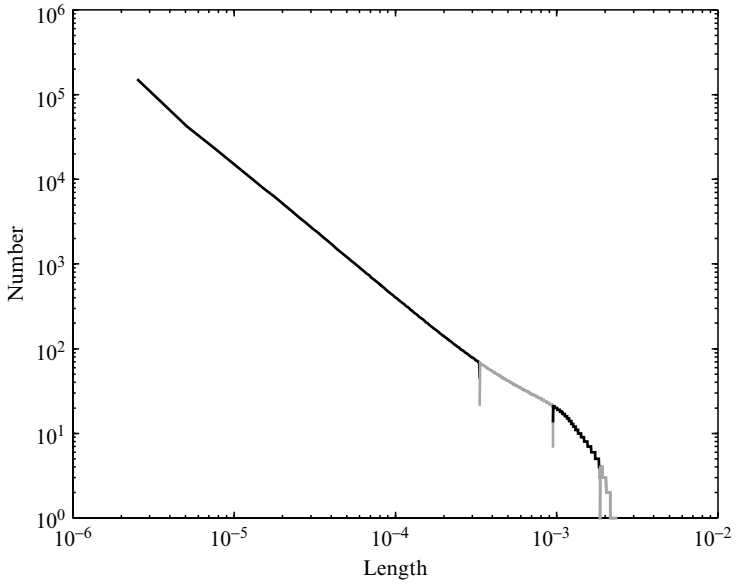


Figure 4.4-3. Wire length distribution.

In this analysis, we have defined a critical path of 24 NAND gates, embedded within the array of 230,400 cells. The NAND critical path is wired by sampling wires lengths from the virtual wire length distribution and their lengths converted to capacitive loads by multiplying their length by the capacitance per unit length, shown in Table 4.4-3, extracted from the 45 nm back-end architecture of Fig. 4.4-4. The capacitance extraction was performed using the program Raphael, and the effects of dielectric barriers, etch-stop layers, and a side-wall slope of 87 deg are included. The total capacitive load is the sum of the sampled interconnect load and the cell input capacitance.

Timing analysis is performed by driving the first cell by a step input, and then using the output transition time matrix to determine the cell output transition time, and hence, the input transition time of the next cell,

Table 4.4-3. Capacitances per unit length.

Capacitance (pF/m)	M1	M2-M6
Lateral	135	118
Upper	24	27
Lower	21	26
Total	180	171

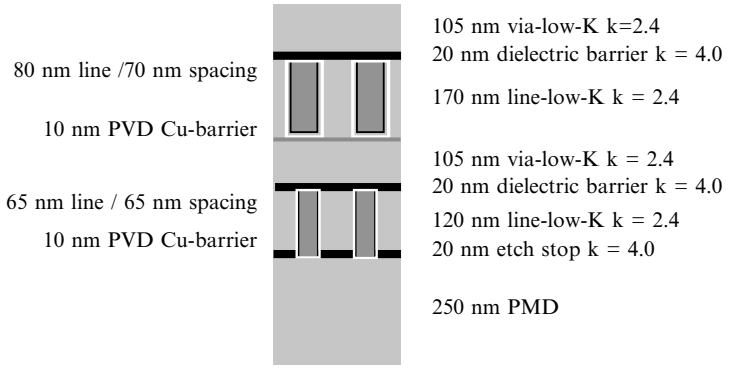


Figure 4.4-4. Back-end architecture for 45 nm node.

and so on, along the chain. Linear interpolation between delay matrix elements was employed. The total path delay is the sum of the cell delays determined from the delay matrices. The sampling process is stochastic, and so a different trial will yield a different range of lengths, and capacitive loads. The final delay is defined to be the most common (modal) delay obtained from a histogram formed by 1000 repeated timing trials [5] (see Figure 4.4-5). This captures the inherently stochastic nature of layout tools.

Figure 4.4-6 shows the delay computed by three commonly used metrics, plus the new VDF metric for the devices that made the mapping onto the SL (1000pA/μm) leakage option. The additional metrics are: the intrinsic gate delay $C_{\text{gate}} V_{dd}/I_{on}$, a nine-stage NAND ring oscillator delay, a hybrid gate/wire delay $C_L V_{dd}/I_{on}$, and the 24 NAND VDF delay, described above. For the hybrid delay, C_L represents the capacitance of a wire length, whose length was calculated by assuming that all the routing channels in six wiring layers (with geometry of Figure 4.4-4) are utilized at 50% occupancy. All delays for each metric are normalized so that the relative performance of the bulk 1000 pA/μm devices calculated by all metrics is 1.00.

We observe: (1) all metrics approximately agree when applied to the bulk metal gate devices, but that there are big discrepancies for more exotic devices and (2) if each metric is used to select the best and worst architecture, we obtain four completely different selections. The same basic trends were also observed for the LL and ULL options, and are not included due to space constraints.

These differences arise because of two main reasons: the first is that the intrinsic gate delay and the hybrid delay are essentially quasi-static and so do not incorporate the effects of advanced architectures on lowering

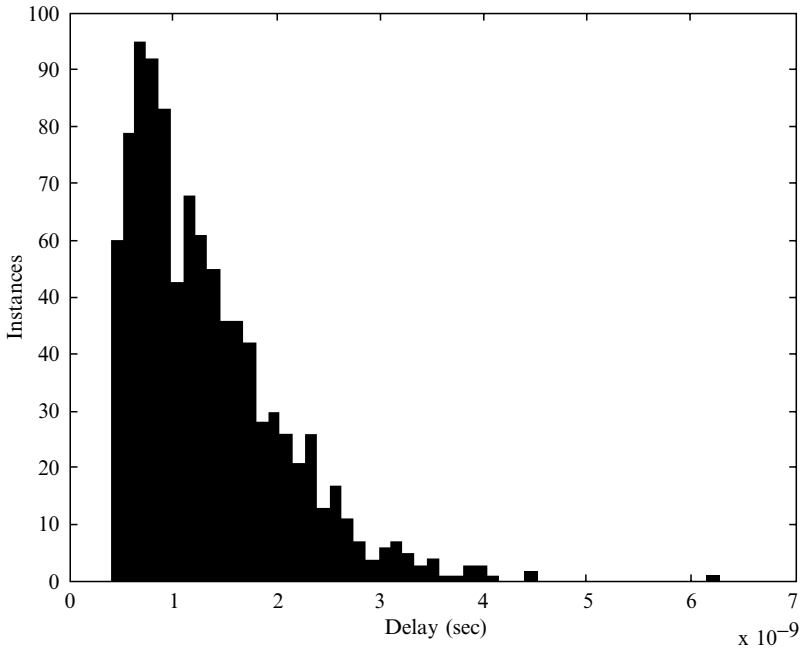


Figure 4.4-5. Histogram of delay times for standard CMOS. One thousand trials sampling wires from the distribution of Figure 4.4-3 were used.

threshold voltages and transition times; the second is that the intrinsic gate delay and the ring oscillator delay do not incorporate wire loading. Also, for the hybrid delay, the load capacitance corresponds to a wire of average length, which is not necessarily the most appropriate statistical measure of such a skewed wire length distribution.

Figure 4.4-7 shows the relative dynamic switching energy of the 230,400 NAND cell array, implemented in the different device options. The darker shading indicates the switching energy of the cell determined by TCAD mixed mode simulation, while the lighter shading indicates the dynamic switching energy of the interconnect. Since the chosen design procedure varies V_{dd} to obtain constant drive currents across all devices, the wire and cell dynamic switching energies are largely determined by the different V_{dd} values (see Table 4.4-2). The double gate and Hi-k bulk devices are the clear winners. Due to lack of space, data for the LL and ULL options are not included, but show similar trends for devices that met the specifications.

Figure 4.4-8 applies the new metric to cross-generational analysis of the 24 NAND gate VDF delay for ULL standard CMOS devices within the

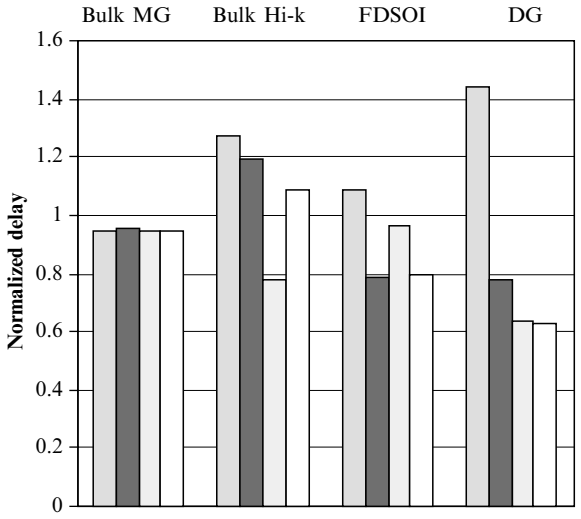


Figure 4.4-6. Delay metrics normalized to standard 1000 pA/μm bulk device delay. Delay are normalized for the four metrics applied to the bulk 1000 pA/μm device are therefore one and are not shown. The bars, from left to right in each bar cluster represent: intrinsic gate delay CgV_{dd}/Ion, nine-stage NAND ring oscillator delay, the hybrid gate/wire delay $C_L V_{dd}/I_{on}$, and 24 NAND path delay.

180, 120, 90, and 65 nm technology nodes. For all these technology options, the average leakage specification is approximately 10 pA/μm. Thus, for a fair comparison, only the 45 nm devices which made the mapping to the ULL specification can be used: the FDSOI and the DG architectures. As might be expected, the two novel device architectures break the historical degradation in performance associated with UL standard CMOS devices. It should be noted that this improvement in speed is not caused by the high intrinsic drive strength of these novel device architectures. During the design process, the high drive strength was traded off for a lower supply voltage in order to match the drive of the standard CMOS architecture. The underlying reasons for the improved performance of these devices is the improved subthreshold

Performance and also that the lower supply voltages mean less charge needs to be supplied to achieve a logic transition. For low power technologies, the lower drive strengths mean that a large component of the delay is caused by the speed with which the device can pass through the subthreshold regime of operation. ULL devices with better subthreshold control are therefore faster than equivalent standard CMOS devices with the same drive strength.

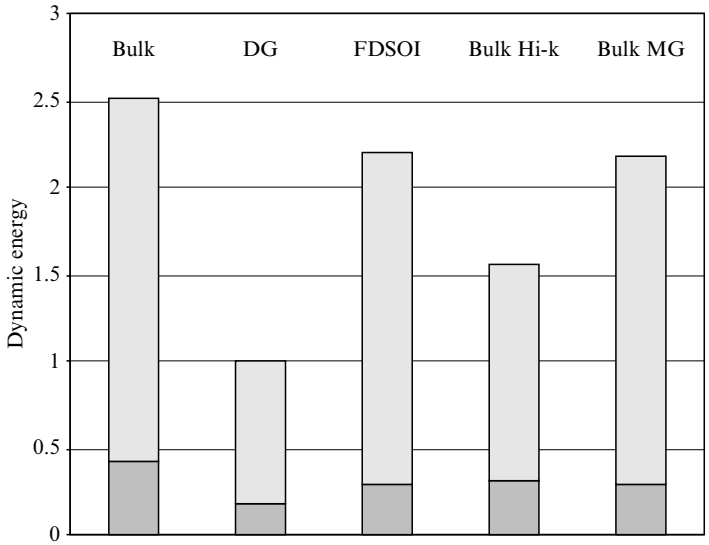


Figure 4.4-7. Dynamic switching energy of embedded NAND cell for SL (1000 pA/μm) device option. The darker shading indicates the switching energy of the cell, while the lighter shading indicates the dynamic switching energy of the interconnect associated with a 230,400 cell array connected using the interconnect of Figure 4.4-4. (assuming a routing efficiency of 50%).

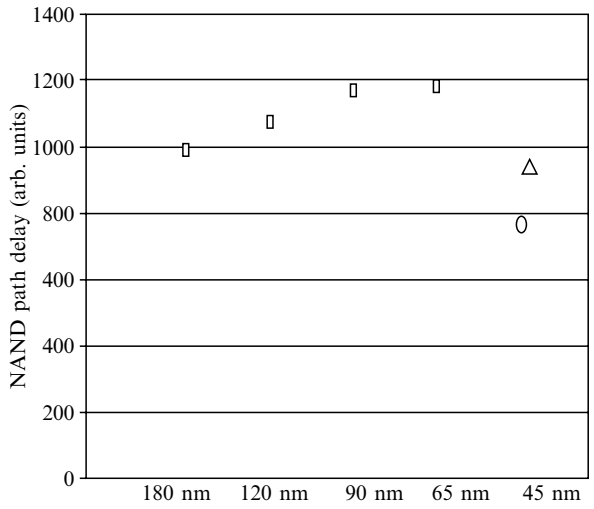


Figure 4.4-8. Virtual design flow (VDF) path delay calculated for ULL technologies from 180 to 45 nm node. For technologies from 180 to 65 nm, the data are for standard CMOS. At the 45 nm node, standard CMOS devices could not be designed to meet the ULL leakage specification, and so the data are for FDSOI (triangle) and double gate (circle).

5. SYSTEM LEVEL MEMORY SIMULATION

Our aim in this exercise is to compare the relative merits of each of the process options (Scaled Bulk, High-K Bulk, Metal Gate Bulk, Fully-depleted SOI, and Dual Gate Architectures) for the core circuit elements within 6T (6 Transistor) SRAM memory cells. Under the assumption that small geometry effects exhibit similar behavior across the different technology platforms, we proceed to extract the electrical characteristics of such basic circuit elements directly from circuit elements constructed with 2D TCAD models of the transistors. While such results may be used for comparing relative merits of the individual technologies in the context of an SRAM cell, care should be exercised in associating them with the exact obtainable values in silicon. Nonetheless, the final values are still valuable as a guide to the expected magnitudes of the electrical parameters for the SRAM cell. We define the following parameters for analysis:

- *Static Noise Margin (SNM)*: the ability of an SRAM cell at retaining data regardless of noise disturbances at the internal nodes and process induced mismatches between the MOS devices in the mirror halves of the memory cell.
- *Write Noise Margin (WNM):* the ease with which the memory cell can be reprogrammed. It is basically defined as the voltage, to which the Bitline or Bitline bar has to change before the voltages stored at the internal nodes (i.e., the stored data in the memory cell), will flip. The write-margin of the 6T-SRAM is highly dependent on the size of the pass gate (PG) MOS devices, but hardly influenced by the size of the pull down (PD) MOS devices.
- *Read current:* reading a stored data in a memory cell involves discharging one of the two precharged interconnects (Bitline and Bitline bar) until a certain amount of voltage difference is established between them. The time taken to establish the required differential signal depends on the maximum amount of current that can be sinked by the memory cell. This current is mainly determined by the size of the PG and PD MOS devices.

The action of reading/accessing a data stored in the SRAM memory bank involves first decoding the address and then asserting the correct word-line. Once the word-line is activated, the accessed SRAM memory cell communicates its stored data by discharging one of the two precharged data-bit-lines, to which it is attached. This generates a differential voltage across the pair of bit-lines that is subsequently detected by the sense amplifier. In this study, the approach taken to assess the read-timing

speed of the SRAM is to consider the time needed to generate the necessary voltage difference between the bit-lines from the uppermost memory cell.

The speed with which this voltage difference is established depends on the parasitic discharge by gate leakage, the total capacitance on the pair of bit-lines, and the read current of the memory cell itself. Ideally the leakage and the junction capacitances should be minimized, whereas the read-currents be maximized. The impact of the leakage currents of the memory cell on the read timing is most noticeable when all the other memory cells sharing the same pair of bit-lines possess the opposite data state from the cell that is being accessed. This is because, in such a situation, the direction of the leakage currents has the primary effect of producing a parasitic discharge of the bit-line bar and behaving as an additional current source that inhibits the discharge of the Bitline. The outcome is simply to prolong the time to set up a voltage difference.

To compare the 6T-SRAM read-time for different technology options taking into account the characteristics of each of the device architectures, a physics-based approach was used whereby a single bit-slice column for 1024 cells with precharger was simulated using mixed mode TCAD. This approach allowed combinations of different finite element MOS device models with specific dimensions (widths) per device to form an accurate description of the 6T-SRAM cell, with additional circuit components involved in the operation of a 1024 cell bit-slice column.

The layout/arrangement of the memory cells along the Bitline suggest a multisection RC line as a plausible electrical model, where we identify the resistance as the interconnect resistance between the abutted memory cells, and the capacitors as the sum of the interconnect capacitance and the junction capacitance from the individual memory cells. In addition, interconnect simulations based on scaled 65 nm dimensions and ITRS specifications for the interconnect properties at 45 nm node show that interconnect of a 400 um wire can be approximated to 10% accuracy by an *RC* section.

Since we are only interested in the worst-case read time of a Bitline column, where only the topmost cell in the Bitline is accessed, it is much more efficient (from the perspective of limited computer resources and simulation time) to approximate the combined parasitic loading on the topmost memory cell. A suitable simplified representation of the 1024 cell Bitline column would then be a single 6T SRAM cell (formed by six separate TCAD MOS devices), linked by the lumped *RC* model of the interconnect to 1023 6T SRAM cells that can, in turn, be lumped together. (In other words, a 6 TCAD MOS device model enlarged with a numeric factor of 1023.) Furthermore, in state-of-the-art SRAM designs, the cell layout used is rectangular with the MOS devices laid out horizontally. This implies that cell height of the individual memory cells can be fixed, in this

work, to 350 nm. With such simplifications the length of the interconnect, and hence, the associated electrical load for a bit-slice column of 1024 cells can be estimated. The remaining write and precharge circuits are implemented using ideal MOS devices simply because they only serve the purpose of setting up the correct data states in the individual 6T-SRAM cells and also to prebias the Bitline/BitlineB to supply voltage prior to data read-out execution.

For the extraction of the read timing, the SRAM cells for all the technology options were designed to have a read current of 30 uA. The results are tabulated in Table 4.4-4, for the 1000 pA/μm option. On the whole, the read timing data extracted for all the device options are fairly similar. This is a consequence of setting the I_{on} of all the individual MOS devices and also the setting of the cell read currents to be equal. However, in spite of this the read speeds are not exactly the same because, for each of the technology options the implemented MOS devices have different physical characteristics. These are reflected in the value of their junction capacitances, sub-threshold control, and mobility parameters. Based on the results in Table 4.4-4, the SRAM cells realized in FDSOI technology are the fastest with read times about 21.1% lower than the equivalent conventional bulk memory cells. Nevertheless, it should also be noted that the dual gate option is only marginally slower than the SOI option by at most 3%.

The small delay is due to the fact that dual-gate technology has more gate capacitance than in the case of the FDSOI devices. More interestingly, the results between the different bulk devices show very little speed difference. In fact, the speed varies within a range of 2.5% faster to approximately 5% slower than the conventional bulk architecture. Of all the technology options, it appears that bulk metal gates on average gives the slowest read speed for the SRAM cell.

Note also that the interconnect loading at 45 nm contribute a significant amount to the delay in the read timing of memory bit slice. The size of

Table 4.4-4. Extracted electrical parameters for all options at 1000 pA/μm.

	PUP/PDN/PGN	$I_{cell\text{-}read(\mathrm{uA})}$	SNM (mV)	WNM (mV)	$t_{read(ps)}$	$V_{DD}(v)$
Bulk	1/2/1.63	30	180	480	373	1.2
Bulk Hi-K	1/2/2	30	110	445	393	0.93
Bulk Metal Gate	1/2/1.95	30	175	445	365	1.12
Fully Depleted SOI	1/2/1.25	30	225	450	294	1.15
Dual Gate	1/2/1.3	30	167	270	331	0.76

Transistor sizes are normalized so that a size of the pull up (PUP) transistor is one.

the contribution can be seen from Figure 4.4-9. For the bulk memory cell, the timing delay increases from 140 to 350 ps, once interconnect loading is included.

In the absence of a full memory cell layout, a comparison of the memory cell area can be obtained by looking at the sum of the sizes of the transistors PUP, PDN, and PGN. We see that in all comparable instances the conventional bulk devices offer the smallest sized memory among the bulk architectures. Not surprisingly, the fully depleted and dual gate structures offers the smallest projected memory cell area, assuming that the physical layout of the individual transistors (i.e., number/size of terminal contacts) are identical to the bulk. Notice too, that the transistor sizes for these two technologies are found to be nearly identical, due in part to their very good sub-threshold control, and the fact that the ions are equal. As for the SNM, this parameter emerges as a consequence of the choices made for the read currents. Although there is quite some variation between the values as a result of the different supply voltages, all of the values are comfortably above 100 mV.

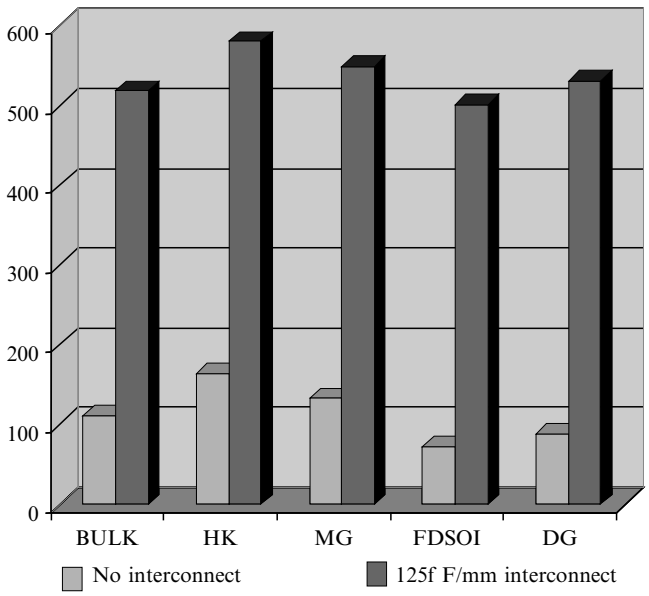


Figure 4.4-9. Read timing delay for 1000 pA/μm leakage spec and 30 uA read current memory cell.

6. CONCLUSIONS

We have developed a virtual design flow for assessing the system level performance of Xtreme low power technology options. The predictions of three other commonly used performance metrics agree with the VDF for bulk/metal gate devices. However, as the devices become more "novel," the predictions diverge rapidly, due to the increasing importance of wire loading and sub-threshold behavior on the performance of more advanced device architectures. In order to make valid power consumption comparisons between radically different architectures, all devices were designed to have the same I_{on} and I_{off} specifications by varying the supply voltage for each device. This enabled a comparison of the dynamic switching energies and path delays among architectures. A cross-generational analysis illustrated the trade-offs between architecture, leakage, and delay, and indicated that FDSOI and DG architectures can meet the 10 pA/μm leakage specification while providing significant gains in performance due to their better sub-threshold control.

The performance of the SRAM cell at 45 nm node was also studied by replicating the usual design procedure for SRAM cells within a TCAD environment. Key characteristics of the SRAM cells, such as static-noise margins, write-noise margins, cell-read currents, and read-timing delays were extracted for a range of pass gate, pull down gate, and pull up gate sizes. The device characteristics and physics as captured by the TCAD models predict that 6T SRAM memory cells made using all the device options have nearly similar transistor dimensions and stability properties. However, the results also clearly show that the bulk metal gate implementation on the whole gives the slowest read timing, largest transistor dimensions to achieve equivalent cell read currents. Also, it is of interest to note that while the 6T SRAM cell in FDSOI technology shows outstanding performance in terms of read delay without interconnect loading, the advantage is much eroded once interconnects are taken into account. In fact the difference in performance reduces from a near 45% improvement from bulk to FDSOI to a smaller gain of approx 15%.

ACKNOWLEDGMENTS

We would greatly like to acknowledge the large contributions to this work made by all the people in the front-end CMOS, advanced CMOS devices, Interconnect and Lithography groups at Philips Research, Leuven, and also the deep submicron cluster at Philips Research in Eindhoven.

REFERENCES

[1] Taur, Y. and Nowak, E. J., 1997, *IEDM Tech.Digest*, **215**.
[2] Thomson, S., Packan, P. and Bohr, M., 1998, *Intel Technical Journal,* Q3, 1–19.
[3] Davis, J. A., De, V. K. and Meindl, J. D., 1998, A stochastic wire-length distribution for gigascale integration (GSI) Part I: Derivation and validation, *IEEE Trans. Electron Devices*, **45**(3), 580–589.
[4] Iqbal, M., Sharkawy, A., Hameed, U. and Christie, P., 2002, Stochastic wire length sampling for cycle time estimation, *Proc. IEEE/ACM System Level Interconnect Prediction Workshop*, 91–96.
[5] Christie, P., 2001, A differential equation for placement analysis, *IEEE Trans. on VLSI*, **9**(6), 913–921.

Section 5

Energy Supply and Management

Chapter 5.1

ENERGY SCAVENGING IN SUPPORT OF AMBIENT INTELLIGENCE
Techniques, Challenges, and Future Directions

Shad Roundy
LV Sensors
sroundy@lvsensors.com

V. Sundararajan
University of California at Riverside
vsundar@engr.ucr.edu

Jessy Baker, Eric Carleton, and Elizabeth Reilly
University of California at Berkeley

Brian Otis
Electrical Engineering Department, University of Washington, Seattle
botis@ee.washington.edu

Jan Rabaey, and Paul Wright
Berkeley Wireless Research Center, University of California, Berkeley
jan@eecs.berkeley.edu

Abstract Ambient intelligence (AmI) depends on the existence of vast quantities of wireless sensors distributed throughout the environment. While advances in IC fabrication technologies, circuit designs, and networking techniques have greatly reduced the cost, size, and power consumption of potential wireless sensor platforms, the development of suitable power sources for many applications lags. The purpose of this chapter is both to review technologies for scavenging energy from the environment to power wireless sensors, and to discuss challenges and future research directions for vibration-based energy scavenging. Many potential energy scavenging technologies are presented along with current state of research and theoretical maximum power densities. Special focus is given to scavenging energy from mechanical vibrations extant in many environments. Results from vibration-based energy scavengers using piezoelectric structures developed by the authors are presented demonstrating power production of approximately $375\,\mu W/cm^3$. While wireless sensor nodes have successfully been powered by vibration-based energy scavengers, several improvements are possible, and indeed necessary, for widespread deployment. Three areas

S. Mukherjee et al. (eds.), AmIware, 265–284.

for improvement are discussed: adaptive frequency tuning, alternative geometry investigation, and microfabrication and integration of generators.

Keywords energy scavenging; piezoelectric; vibrations

1. INTRODUCTION

Pervasive networks of wireless sensor and communication nodes have the potential to significantly impact society and correspondingly create large market opportunities. The last several years have witnessed a large research effort based around the vision of ubiquitous networks of wireless sensor and communication nodes [1–3]. As the size and cost of such wireless sensor nodes continue to decrease, the likelihood of their use becoming widespread in buildings, industrial environments, automobiles, aircraft, etc., increases. However, as their size and cost decrease, and as their prevalence increases, effective power supplies become a larger problem. The scaling down in size and cost of CMOS electronics has far outpaced the scaling of energy density in batteries, which are by far the most prevalent power sources currently used. Therefore, the power supply is quickly becoming largest and most expensive component of the emerging wireless sensor nodes being proposed and designed. The cost of batteries is compounded by the fact that batteries must be either replaced or recharged on a regular basis. This regular maintenance could easily become the single greatest cost of installing a wireless sensor network for many applications. If wireless sensor networks are to truly become ubiquitous, replacing batteries in every device every year or two is simply cost prohibitive.

The purpose of this chapter is twofold. First, existing and potential power sources for wireless sensor networks will be briefly reviewed. Second, challenges and future research directions for vibration-based energy scavengers will be discussed. While many scavenger devices have been demonstrated, a number of issues limit their widespread application. This chapter will discuss some of these issues including resonance tuning, optimal design geometries, and microfabrication with a goal to monolithic integration with sensors and wireless electronics. It should be noted that the design of optimal power electronics to maximize the power transferred to the load is also an important research area, but outside the scope of this chapter.

2. POWER CONSUMPTION

Before considering power sources, it is useful to consider the power demand of a typical wireless sensor node. Assuming that the radio transmitter operates at approximately 0 dBm (which would roughly correspond

to an average distance of 10 m between nodes), the peak power consumption of the radio transmitter will be around 2–3 mW depending upon its efficiency. Using low power techniques [4], the receiver should not consume more than 1 mW. Including the dissipation of the sensors and peripheral circuitry, a maximum peak power of 5 mW is quite reasonable. Given a maximum data-rate for the radio of 100 Kbit/s, and an average traffic load per node of 1 Kbit/s (these numbers are based on real-radio prototypes and a realistic smart home scenario), every node operates at a duty cycle of approximately 1%. During the remaining 99%, the only activities taking place in a node are a number of background tasks: low-speed timers, channel monitoring, and node synchronization. The latter actually is the dominant power consuming source of the node if not handled appropriately. Using advanced "wake-up radio techniques" or semiasynchronous beaconing techniques, the average "standby" power of the node can be limited to 50 μW or lower. Combining peak and standby power dissipation leads to an average power dissipation of approximately 100 μW.

Several small low power wireless platforms are currently available commercially. Companies providing wireless sensor platforms include Dust Networks [5], Crossbow [6], Xsilogy [7], Ember [8], and Millenial Net [9]. The power needed to operate these platforms depends on how and where they are used. Based on the authors' investigations, they generally require an average power consumption about 1 order of magnitude higher than the 100 μW proposed above (generally one to several mW). However, research projects have demonstrated that a wide range of applications is possible within a power budget of approximately 100 μW.

3. REVIEW OF POTENTIAL POWER SOURCES

One may classify possible methods of providing power for wireless nodes into three groups: store energy on the node (i.e., a battery), distribute power to the node (i.e., a wire), and scavenge available ambient power at the node (i.e., a solar cell). Of course, combinations of the three methods are also possible. In fact, even in an energy scavenging or power distribution method, some onboard energy storage must be available. Energy storage and energy scavenging technologies will be briefly reviewed here. The reader is referred to Roundy et al. [10] for a discussion of methods to distribute power to a wireless sensor node.

The vast majority of wireless devices, including wireless sensors, are currently designed to run on batteries. Figure 5.1–1(a) shows the average possible power draw from three common primary battery chemistries

versus time. Based on a power budget of 100 μW as specified above, 1 cm^3 of lithium battery will provide approximately 1 year of operation. Figure 5.1–1(b) shows the average power draw from three common secondary (rechargeable) battery chemistries. It should be remembered that rechargeable batteries are a *secondary* power source. Therefore, in the context of wireless sensor networks, another primary power source must be used to charge them. In most cases, it would be cost prohibitive to manually recharge each device. More likely, an energy scavenging source on the node itself, such as a solar cell, would be used to recharge the battery.

Alternative energy storage technologies include microfuel cells [11] and microradioactive generators [12]. Hydrocarbon-based fuels have very high energy densities compared to batteries. For example, methanol, a common fuel for microfuel cells, has an energy density of 17.6 kJ/cm^3, which is about six times that of a lithium battery. Efficiencies of large-scale fuel cells have reached approximately 45% electrical conversion efficiency and nearly 90% if cogeneration is employed [13]. Efficiencies for microscale fuel cells will certainly be lower. The maximum obtainable efficiency for a microfuel cell is still uncertain. Demonstrated efficiencies are generally below 1% [14]. Given the energy density of fuels, such as methanol, small-scale fuel cells need to reach efficiencies of around 20% in order to be more attractive than primary batteries. Some radioactive materials contain extremely high energy densities. Li and Lal [12] have developed the beginnings of a microradioactive generator using the ^{63}Ni isotope. The power level is equivalent to 0.52 $\mu W/cm^3$ at an efficiency of $4X10^{-6}$. With significant engineering improvements, this could become a significant energy source for wireless devices.

Energy scavenging offers another alternative to solve the energy supply problem, and a number of approaches have been studied over the past years. Several researchers have pursued naturally occurring temperature

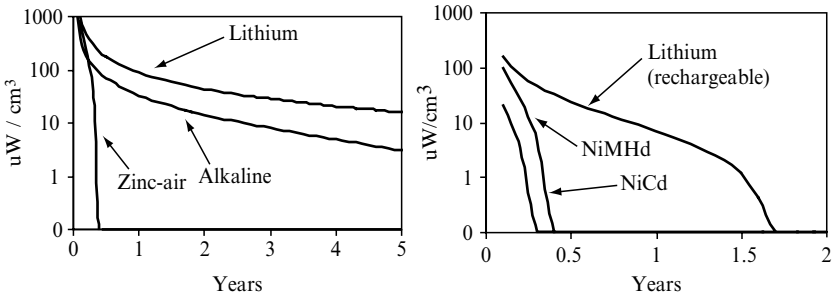


Figure 5.1-1. (a) Average power draw versus lifetime for primary battery chemistries. (b) Average power draw versus lifetime for secondary (rechargeable) battery chemistries.

variations as one potential source [15–17]. Shenck and Paradiso [18] have built shoe inserts capable of generating 8.4 mW of power under normal walking conditions. Much research has focused on solar (photovoltaic) power [19]. Federspiel and Chen [20] used a small (about 10 cm in diameter) airflow turbine to generate power intended for use by a Mica Mote [6]. They produced approximately 350 $\mu W/cm^2$ from air velocity at 5 m/s. Vibration energy scavengers based on electromagnetic [21–24], electrostatic [25–27], and piezoelectric [28–30] conversion have been suggested in the literature. The opinion expressed in this chapter is that, piezoelectric vibration-to-electricity converters offer the most potential at the meso-scale level. Reported power densities are on the order of 100–300 $\mu W/cm^3$. Table 5.1–1, which has been updated in this chapter from previous research [10] shows comparisons among many potential sources. It is the authors' opinion that no single energy-scavenging solution will fit all environments and application spaces. Rather, unique solutions need to be considered on an application-by-application basis. However, as clearly indicated in the table, vibrations as a power source compare well to other potential energy scavenging sources.

Table 5.1-1. Updated comparisons among potential energy and power sources.

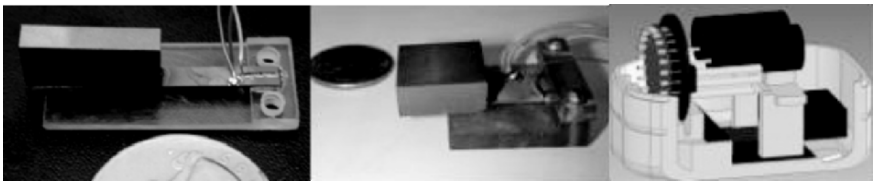
Power Source	P/cm^3 ($\mu W/cm^3$)	E/cm^3 (J/cm^3)	$P/cm^3/yr$ ($\mu W/cm^3/yr$)	Secondary storage needed	Voltage regulation
Primary battery	—	2880	90	No	No
Secondary battery	—	1080	34	—	No
Microfuel cell	—	3500	110	Maybe	Maybe
Ultracapacitor	—	50–100	1.6–3.2	No	Yes
Heat engine	10^6	3346	106	Yes	Yes
Radioactive (^{63}Ni)	0.52	1640	0.52	Yes	Yes
Solar (direct sunlight)	15,000 *	—	—	Usually	Maybe
Solar (inside)	10*	—	—	Usually	Maybe
Temperature	40*†	—	—	Usually	Maybe
Human power	330	-	—	Yes	Yes
Air flow	380††	—	—	Yes	Yes
Pressure variation	17†††	—	—	Yes	Yes
Vibrations	375	—	—	Yes	Yes

*Sources whose metric is power per cm rather than power per cm^3.
†Demonstrated from a 5°C temperature differential.
††Assumes air velocity of 5 m/s and 5% conversion efficiency.
†††Based on 1 cm^3 closed volume of helium undergoing a 10°C change once a day.

4. VIBRATION-BASED ENERGY SCAVENGING

While most of the power sources listed in Table 5.1–1 have an important role to play within the world of energy scavenging for wireless sensor networks, the focus here is on vibrations as a power source. In the last 3 years, as many as 50 publications have emerged on the potential of vibrations, at least three new companies now have been formed based on the development of vibration-oriented scavenging products, and several established companies have begun development on such products. These recent articles and small businesses add additional weight to the proposition that vibrations are abundant, but as yet untapped, for power sources in microelectronics. While many scavenger devices have been demonstrated, a number of issues limit their widespread application. The remainder of this chapter will explore those issues.

The opinion expressed in this chapter is that, piezoelectric vibration-to-electricity converters offer the most potential at the meso-scale ($\sim 1\text{cm}^3$) level. The last row of Table 5.1–1 indicates $\sim 375\,\mu\text{W/cm}^3$ for vibration sources. This value is based on simulation and experimental results from the device shown in Figure 5.1–2(a) driven by vibrations of $2.25\,\text{m/s}^2$ at 60 Hz[10]. The device in Figure 5.1–2(a) has been used to power a 1.9 GHz radio transmitter [31]. In a second test, using the larger generator in Figure 5.1–2(b), and integrated into a complete package in Figure 5.1–2(c), temperature readings were sent from one of the Mica2Dot "motes" fabricated by Crossbow Inc. [6]. These were powered at about a 1% duty cycle. The "on" power of the larger node was 40–60 mW, and the average power was 400–600 μW. This "TempNode unit"—run entirely from scavenging—was part of a "smart" building project for regulating temperatures in a residence.



(a) (b) (c)

Figure 5.1-2. Meso-scale piezoelectric generators.

5. BRIEF REVIEW OF MODELING FOR PIEZOELECTRIC ENERGY SCAVENGERS

To frame the challenges and constraints placed on the energy generation, it is useful to review the models developed thus far; but only at a level of depth needed to indicate the major design parameters that contribute to the power scavenged. Figure 5.1–3 shows a schematic of a two-layer bender mounted as a cantilever.

When this device is driven by vibrations, the generator provides an AC voltage. If a resistor is connected across the piezoelectric electrodes, a simple RC circuit results. Although simply driving a resistor is not that useful in practice, this simple circuit gives a reasonably good approximation for the amount of power that can be generated. Furthermore, it leads to the development of an analytical model that can give insight into the design—performance relationships. The development of an analytical model for this generator, and the validation of this model, has been published elsewhere [10], so only the results will be given here. The power delivered to the load resistor is given by:

$$P = \frac{1}{2R} \frac{\left(\frac{2k_{31}t_c}{k_2}\right)^2 \frac{c_p}{\varepsilon} A_{in}^2}{\left[\frac{\omega_n^2}{\omega \mathrm{RC}_b} - \omega\left(\frac{1}{\mathrm{RC}_b} + 2\zeta\omega_n\right)\right]^2 + \left[\omega_n^2(1+k_{31}^2) + \frac{2\zeta\omega_n}{\mathrm{RC}_b} - \omega^2\right]^2}, \quad (1)$$

where ω is the frequency of the driving vibrations, ω_n the resonance frequency of the generator, c_p the elastic constant of the piezoelectric ceramic, k_{31} the piezoelectric coupling coefficient, t_c the thickness of one

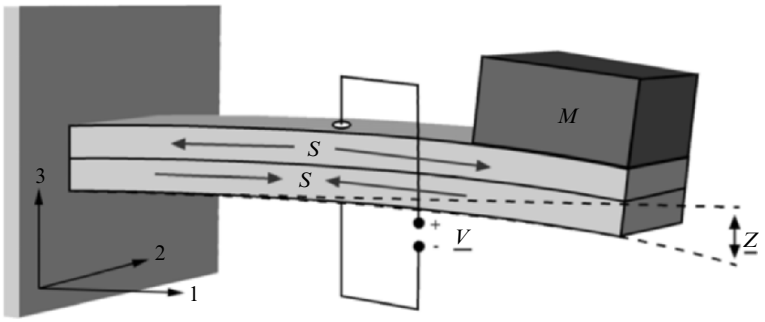


Figure 5.1-3. Two-layer bender mounted as a cantilever. S, is strain; V, voltage; M, mass; and z, vertical displacement.

layer of the piezoelectric ceramic, k_2 is a geometric constant that relates average strain in the piezoelectric material to the tip deflection, ε the dielectric constant of the piezoelectric material, R the load resistance, C_b the capacitance of the piezoelectric bender, and ζ the dimensionless damping ratio.

If it is assumed that the device is operating at resonance (i.e., the driving frequency, ω, matches the natural frequency ω_n), then Equation (1) could be rewritten in the form of Equation (2), where $\omega_n^2 = \mathrm{km}^{-1}$, where k, the equivalent spring stiffness, has been substituted.

$$P = \frac{m^2}{2k} \frac{\mathrm{RC}_b^2 \left(\frac{2k_{31}t_c}{k_2}\right)^2 \frac{c_p}{\varepsilon} A_{in}^2}{(4\zeta^2 + k_{31}^4)(\mathrm{RC}_b)^2 k + 4\zeta k_{31}^2 \mathrm{RC}_b \sqrt{\mathrm{km}} + 4\zeta^2 m}. \tag{2}$$

It directly follows from Equation (1) that power is maximized when the natural (or resonance) frequency (ω_n) is equal to the driving frequency (ω). In fact, power output drops off dramatically as ω_n deviates from ω, as is illustrated in Figure 5.1–4(a). Equation (2) indicates that power is dependent on the proof mass (m). While the relationship is not exactly linear, in the region of interest, it is close to linear as demonstrated in Figure 5.1–4(b). It should also be noted that increasing the mass generally has the effect of increasing the amount of piezoelectric material in order to maintain the natural frequency. Finally, the relationship between power and the coupling coefficient of the material (k_{31}^2) is not obvious from the form of Equation (2). Figure 5.1–4(c) indicates this relationship in graphical form. The figure indicates that power increases quickly with increasing coupling

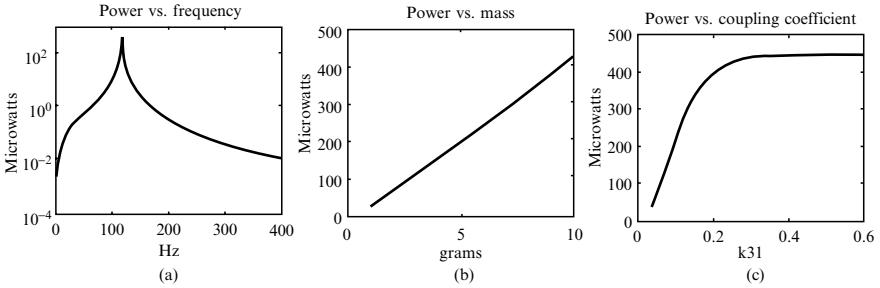


Figure 5.1-4. (a) Power output versus driving frequency. The natural frequency of the design is 120 Hz. (b) Power output versus proof mass. All parameters except piezoelectric beam thickness have been kept constant. (c) Power output versus coupling coefficient. Simulations were performed for a $1\,\mathrm{cm}^3$ piezoelectric scavenger driven by vibrations of $2.5\,\mathrm{m/s}^2$.

up to a point, beyond which the increase is very modest. Nevertheless, as system level coupling (especially for microfabricated devices) is usually below the "knee" (Figure 5.1–4(c)), improvement of the material coupling coefficient is an important area of development.

The design relationships highlighted by the model presented in this section are summarized in Table 5.1–2. Table 5.1–2 also serves as the basis for design improvements and potential directions for current research on piezoelectric energy scavengers.

6. IMPROVED ENERGY HARVESTING

6.1. Self and Adaptive Tuning Energy Sources

As discussed in the previous section, the power drops off dramatically as the resonance frequency deviates from that of the driving vibrations. Under many circumstances, the driving frequency will be known before the device is designed and fabricated, and the appropriate resonance characteristics can thus be "built-in." In other situations, however, this frequency will not be known a priori, or it may change over time. It is also relevant to consider the mass-fabrication of such devices for use by other investigators. It would clearly be advantageous to create a single design that operates effectively over a range of vibration frequencies.

Table 5.1-2. Design relationships and potential improvements for piezoelectric vibration based scavengers.

Design relationship from Equations (1) and (2)	Current designs	Design strategy/improvement
Power versus natural frequency	Operation limited to a very narrow frequency band	Design for adaptive self-tuning of the natural frequency
Power versus mass	Power limited by the proof mass	Explore designs for which the strain from a given mass can be improved
Power vs. piezoelectric coupling coefficient	System coupling coefficient is below "knee" in power versus coupling coefficient curve	Improve designs for better "system" coupling, and improve material properties of thin film piezoelectrics
System integration (constraints implied by equations)	Limited by hand assembly	MEMS designs desirable in order to integrate with sensors and CMOS

Two possibilities exist: first, actuators could be developed that can alter the scavenger's resonance frequency, and second, designs for scavengers with wider bandwidths could be developed. We will first consider resonance frequency tuning actuators.

We have classified potential actuators to accomplish the resonance tuning into two categories: "active tuning" actuators and "passive tuning" actuators. The distinction is that "active" actuators must remain continuously "on" as long as the resonance frequency is not at its nominal frequency. A good example of an "active" actuator would be electrostatic springs [32]. By contrast, "passive tuning" actuators turn "on" to tune the generator, and then are able to turn "off" while maintaining the new resonance frequency. An example would be a moveable clamp that would change the length (and thus the stiffness) of a flexure mechanism. "Active tuning" actuators will be considered first, followed by a design for a "passive tuning" actuator.

We will use a nontechnology specific vibration-based scavenger model to calculate the power required for tuning [21]. The primary idea behind this model is that the generator behaves like a second-order mechanical system. The electrical energy removed from the oscillating mechanical system behaves like viscous damping from the point of view of the mechanical system.

The power output for this simplified model is given by:

$$P = \frac{m\zeta_e \left(\frac{\omega}{\omega_n}\right)^3 A_{in}^2}{\omega\left[\left(2\zeta\frac{\omega}{\omega_n}\right)^2 + \left(1 - \left(\frac{\omega}{\omega_n}\right)^2\right)^2\right]}, \tag{3}$$

where ζ_e is the damping ratio associated with the electrically induced damping, ζ the sum of ζ_e and ζ_m (the damping ratio representing mechanical loss), and other symbols are as defined for Equations (1) and (2).

Equation (3) is a more general power expression than Equations (1) and (2). As such, it applies to a wider range of scavengers (piezoelectric, electromagnetic, etc.), but provides less accurate estimates and less design insight for piezoelectric generators than Equations (1) and (2). However, this simplified model is useful to initiate the exploration into frequency tuning because of its wide applicability.

Three tuning methods might be employed: (1) the tuning actuation force, $F_a(t)$, is proportional to displacement (alters the effective stiffness), (2) $F_a(t)$ is proportional to acceleration (alters the effective mass), and (3) $F_a(t)$ is proportional to velocity (alters the effective damping). It can be

shown mathematically [33], that as long as the scavenger system is reasonably well-modeled by Equation (3), an "active" actuator will never result in improved power output. In other words, the increase in power output derived from applying the actuator will always be less than the amount of power required by the actuator. A "passive" tuning actuator must be able to alter the resonance frequency and then cut power to the actuator while maintaining the new resonance frequency. One possible method to alter the stiffness of a beam structure is by the application of destabilizing axial loads [34]. Figure 5.1-5(a) shows a schematic concept of the structure. In reality, there would be another proof mass on the top of the beam, but this has been removed for illustration purposes. The beam is modeled as a simply supported beam with a proof mass in the middle. The apparent stiffness of the beam is a function of the axial compressive preload and theoretically reduces to zero as the axial preload approaches the critical buckling load. The preload could easily be applied by set-screws or other devices that push on the clamps at either end of the beam. While this would require a significant amount of power during the tuning phase, once the proper preload is applied, all power to the tuning operation could then be turned off. Figure 5.1-5(b) shows how the resonance frequency of the beam in Figure 5.1-5(a) decreases as a function of preload. The resonance frequency can be reduced by approximately 40% with a preload of one-half the critical buckling load. Furthermore, the response of the actuator is fairly linear in the region up to one-half the critical buckling load.

The other approach is to design a structure with a wider inherent bandwidth. Again, using the model represented by Equation (3), the

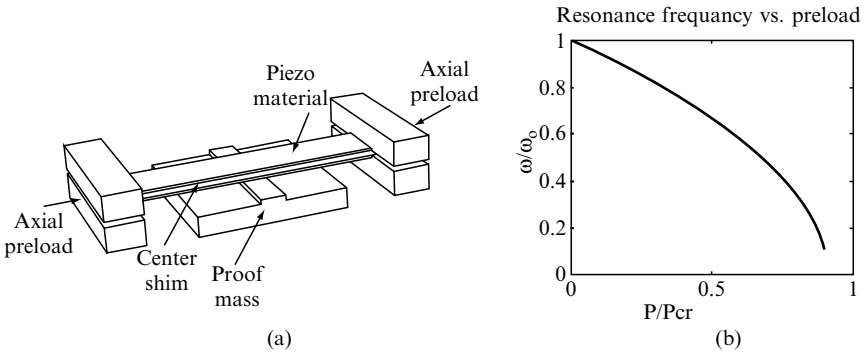
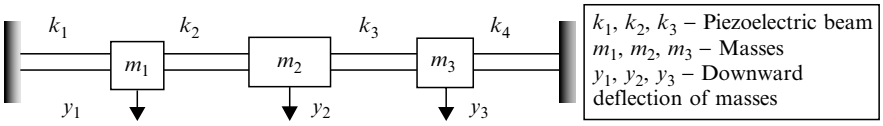


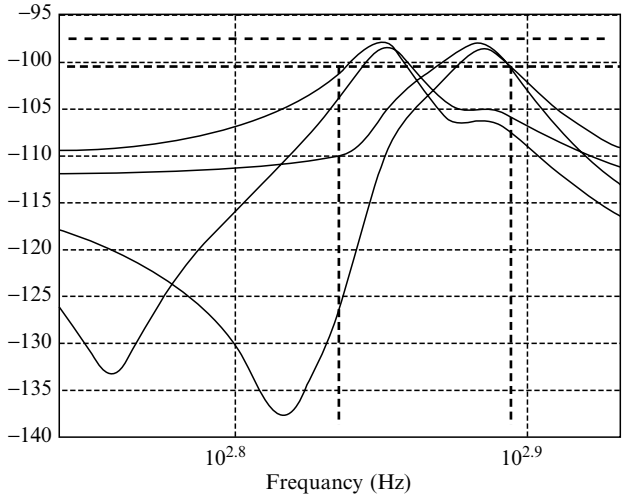
Figure 5.1-5. (a) Schematic of a simply supported piezoelectric beam scavenger with an axial preload. (Top half of proof mass is removed for illustration purposes.) (b) Ratio of tuned to original natural frequency as a function of the axial preload.

bandwidth is approximately $2\zeta\omega_n$. If the bender resonates at 120 Hz and the damping ratio is 0.025 (which corresponds well with measurements), the bandwidth is 6 Hz, or ± 3 Hz.

This implies that to get a high coupling between the source and the piezoelectric transducer, the scavenger has to be tuned to within ± 3 Hz. This can be quite difficult in practice since the scavenger geometry and material properties show enough variation to push the device off resonance. One possible method to design scavengers with a wider bandwidth is to connect N spring-mass-damper systems (Figure 5.1-6(a) shows a system with three masses and four springs), such that the resonance frequency of the $i+1$th system is located one bandwidth away from that of the ith system. Thus the problem is one of determining the springs and the masses given the eigenvalues of the individual systems. The overall system is given by the Equation (4), where the objective is to find $k_1 \ldots k_n$ and $m_1 \ldots m_n$.



(a) Vibration stucture made of 3 masses and springs



(b) Frequency response of the 3 mass-spring structure

Figure 5.1-6. Multidegree of freedom bender designed for higher bandwidth.

$$\begin{bmatrix} \ddot{y}_1 \\ \ddot{y}_2 \\ \cdots \\ \ddot{y}_n \end{bmatrix} = \begin{bmatrix} \frac{-(k_1+k_2)}{m_1} & \frac{k_2}{m_1} & 0 & \cdots & \cdots & 0 \\ \frac{k_2}{m_2} & \frac{-(k_2+k_3)}{m_2} & \frac{k_3}{m_2} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & \frac{k_{n-1}}{m_n} & \frac{-(k_n+k_{n-1})}{m_n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} - \begin{bmatrix} \ddot{y}_f \\ \ddot{y}_f \\ \ddot{y}_f \\ \ddot{y}_f \end{bmatrix} \quad (4)$$

Figure 5.1-6(b) shows the frequency response for the spring deflection amplitudes of Figure 5.1-6(a), where $k_1 = 568\,N/m$, $k_2 = 776\,N/m$ *and* $k_4 = 488\,N/m$. The masses are $m_1 = 2.8125\,g$, $m_3 = 2.268\,g$, while $m_2 = 30\,g$. A damping constant of 0.025 has been added. As can be seen from the figure, the overall bandwidth is 17.5 Hz, with two of the four springs contributing to the output at any given time.

6.2. Alternative Mechanical Structures Using Piezoelectric Materials

Most researchers have focused research efforts on piezoelectric vibration-based generators on a single geometry. Indeed, most researchers have focused on a cantilever beam, or some slight variation of a cantilever beam. A cantilever beam has a number of advantages: it produces relatively low resonance frequencies and a relatively high level of average strain for a given force input, and it is easily realizable in a microfabrication process. However, a broader consideration of potential design geometries can increase the performance of scavengers. Furthermore, different geometries may have characteristics that make them more attractive for a given application. Any design geometry should try to address at least one, if not all, of the following goals:

- Maximize the piezoelectric response from a given input. This could be accomplished by either maximizing the average strain in the material for a given input, or changing the design to take advantage of the 3–3 coupling mode;
- Improve the robustness of the scavenger by reducing stress concentration;
- Minimize losses (damping) associated with the mechanical structure; and
- Improve the manufacturability of the scavenger.

A major area for improvement of the scavenger design is in the geometry of the piezoelectric bender itself. Bulk material properties impose a strain limit typically around 500 microstrain. If this strain limit is exceeded, the brittle ceramic piezoelectric bimorph fractures, compromising its

generative capacity by upwards of 60%. Multiple cycles of overstrain completely destroy the bimorph's generative capacity.

One method of preventing overstrain is to optimize the profile of the bender. Benders produce electric current through deformation or strain of the piezoelectric material. It follows that maximizing average strain should maximize energy output. In order to maximize average strain, strain must be both as uniform and as large as possible without exceeding the material limit. A cantilever with a rectangular profile creates a strain concentration at the clamped end, where the bending moment is at a maximum. With the same volume of PZT and an increasingly triangular trapezoidal profile, the strain can be distributed more evenly, such that maximum strain is reached at every point in the bimorph. A trapezoidal geometry can supply more than *twice* the energy (per unit volume) that the rectangular geometry can supply, reducing both size and cost of the bimorph. The bending energy of a beam is given by:

$$U_b = \int \frac{M^2}{2EI} dx, \tag{5}$$

where M is the bending moment, E the modulus of elasticity, I the moment of inertia, and x the distance from the base (or root) of the beam.

- The value of the bending energy relative to a cantilever of uniform width for two alternative beam geometries is shown in Figure 5.1-7.
- In addition to modifying the profile of the cantilever beam, we are exploring the following designs in detail:
- The double-cantilever beam (see Figure 5.1-8(a)). This design exploits the trapezoidal shape developed earlier in this section, while

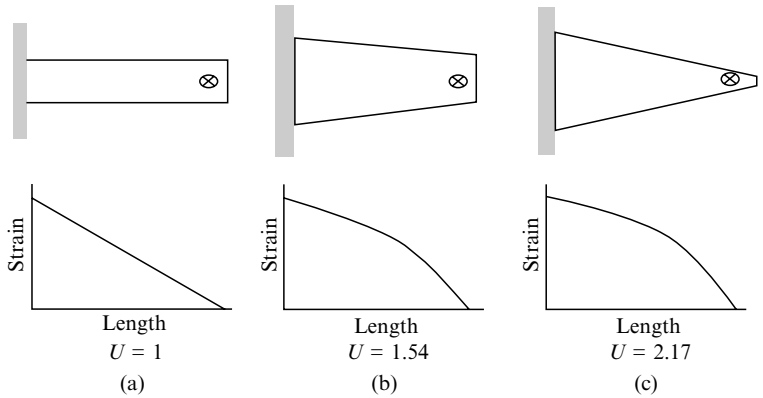


Figure 5.1-7. Relative bending energies and strain profiles for three alternative beam geometries.

allowing for an axial load to be applied at the clamped ends, which would provide for frequency tuning.

- The curved compressor (see Figure 5.1-8(b)). This design utilizes curvature in the generator to ensure that the acting force induces compressive, rather than tensile stress in the piezoelectric material. This extends the fatigue life while maintaining the same level of power output as an equivalent tensile stress. Again, a lateral force can be applied to tune the natural frequency of the generator to its environment.
- The compliant mechanism and stack (see Figure 5.1-8(c)). The third design to be explored is one that translates relatively low force, high displacement vibrations into smaller displacement, higher force motion around a PZT stack. This capitalizes on the higher efficiency of the out-of-plane coupling mode and coefficient (k_{33}), without the usual requirement of high force actuation.

6.3. The Future—Microfabricated Ubiquitous Wireless Sensor Nodes

The ultimate goal of energy scavenging devices for pervasive computing is the realization of a completely integrated, microfabricated sensor node. To create a ubiquitous device of this nature requires careful consideration of a number of materials and processing difficulties, each of which predicate certain constraints on the materials selection and processing. Specifically, the concerns are: (1) the piezoelectric material must have sufficient material properties to produce a useable voltage under strain and have an intimate crystallographic and interfacial contact with the electrodes and growth substrate, (2) the growth, fabrication, and integration of the power source must consist solely of standard microfabrication processes, and (3) the design of the energy scavenging device (with optimized piezoelectric layers) must be able to produce sufficient voltage under ambient vibrational excitation.

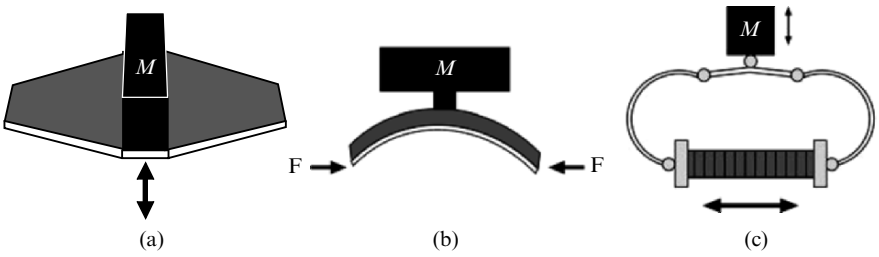


Figure 5.1-8. Schematic representations for three alternative designs: (a) the double cantilever, (b) the curved compressor, and (c) the compliant mechanism and stack.

Although many piezoelectric materials exist and are commercially available as thin films (thickness on the order of 1 micron), few of them exhibit large enough piezoelectric coefficients to generate appreciable voltage while under strain. The $Pb(Zr,Ti)O_3$ (PZT) family of piezoelectrics in general have large piezoelectric coefficients and can be grown epitaxially with careful selection of growth substrate and electrodes (typically oxide electrodes, such as $SrRuO_3$ (SRO) and $LaSr_{0.5}Co_{0.5}O_3$ (LSCO)). Epitaxial, or pseudo-single crystal, thin films are attractive because the piezoelectric coefficients, mechanical constants, and dielectric properties of the films can be an order of magnitude higher than polycrystalline films of the same composition. Also, epitaxial growth processes result in a very intimate contact with the electrode, which can reduce both mechanical fatigue and cyclic depolarization.

The requirement for all growth, fabrication, and integration steps to utilize standard microfabrication techniques necessitates the use of silicon wafer substrates and fabrication processes. Although the literature points to many methods for achieving growth of PZT on Si, most of these processes involve elaborate methods and multiple depositions of exotic buffer layers, most of which are not necessarily compatible with standard postgrowth fabrication processes, such as selective wet etching. An exciting alternative has been recently developed by Motorola [35], and involves the deposition of only one buffer layer, $SrTiO_3$, which is an excellent growth template and allows for the growth of epitaxial PZT thin films. Using SRO as an electrode also allows the use of typical wet and dry etch fabrication processes, thus permitting a design based completely on standard microfabrication processes.

Recent experiments have deposited PZT thin films with SRO top and bottom electrodes onto STO coated Si single crystal substrates (obtained from Motorola). The resultant epitaxial films had c-axis orientation, a remanent polarization of $40\,\mu C/cm^2$, and a d_{33} value of 140 pm/V (d_{31} values are on the order of 60% of d_{33}). Arrays of varying length SRO/PZT/SRO cantilever beams have been patterned and dry etched using standard microfabrication processes as shown in Figure 5.1-9. The inset image is a magnified section of the end of a beam. Selective wet etching processes to *release* the cantilever beams from the Si wafer are currently underway.

7. CONCLUSIONS

Power sources are rapidly becoming a bottleneck limiting the widespread deployment of the wireless sensor devices that will enable pervasive computing. Both lower power designs and development of alternative power sources are desperately needed. While researchers are working on

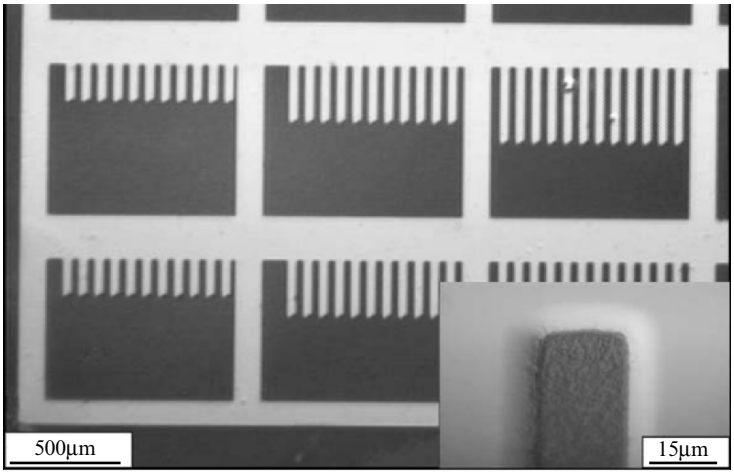


Figure 5.1-9. Dry etched arrays of unreleased SRO/PZT/SRO cantilever beams.

a number of different methods to scavenge ambient energy to power wireless sensor devices, the focus here has been on the design of piezoelectric generators that convert low level mechanical vibrations into electrical power. A number of such generators have been successfully implemented in laboratory settings. Nevertheless, there are significant opportunities to improve the performance of such generators. The authors are pursuing the following approaches to improving the performance of vibration-based energy scavengers:

- **Frequency tuning.** Current designs must resonate at the driving frequency in order to generate a significant amount of power. Designs incorporating multiple proof masses can moderately increase the bandwidth of scavengers from around 6 Hz to 18 Hz. "Active" tuning actuators that must remain "on" and consume power continuously while maintaining a new resonance frequency will not result in improved power output. Therefore, "passive" actuators that tune the frequency and then are able to turn "off" while maintaining the new frequency are needed to improve the power output by frequency tuning.
- **Alternative design geometries.** We have considered a number of potential alternative design geometries. The alternative designs focus on improving the power output for a given input, or improving the robustness of a design. Performance can be improved by perhaps a factor of 2, and reliability can be significantly improved simply by altering the profile of the cantilever beam. More novel geometries are currently under investigation.

- **Development of processes to allow MEMS implementation and integration with sensors and electronics.** Using recently developed processes for growth and microfabrication of thin film PZT structures, *unreleased* vibration-based generators employing microcantilever heterogeneous bimorphs have recently been fabricated. Initial calculations show that an areal power density of $5\,\mu W/cm^2$ and a volume power density of $80\,\mu W/cm^3$ are possible from such structures.

ACKNOWLEDGMENTS

The authors would like to acknowledge a wider group of colleagues contributing to this overall effort. They include Professors Seth Sanders and James Evans and students Eli Leland, Elaine Lai, Daniel Steingart, Chistine Ho, and Michael Montero. Funding from a number of sources include: the Ford Fund, the Luce Foundation, the National Science Foundation, the California Energy Commission, the Department of Energy Integrated Manufacturing fellowship, and the Noyce-Intel fellowship.

REFERENCES

[1] Rabaey, J., Ammer, J., Karalar, T., Li, S., Otis, B., Sheets, M. and Tuan, T., 2002, PicoRadios for wireless sensor networks: The next challenge in ultra-low-power design, *Proceedings of the International Solid-State Circuits Conference*, San Francisco, CA, February 3–7, 2002.

[2] Warneke, B., Atwood, B. and Pister, K. S. J., 2001, Smart dust mote forerunners, *Fourteenth Annual International Conference on Micro-electromechanical Systems (MEMS 2001)*, Interlaken, Switzerland, January 21–25, 2001.

[3] Hill, J. and Culler, D., 2002, Mica: A wireless platform for deeply embedded networks, *IEEE Micro*, **22**(6), 12–24.

[4] Otis, B. and Rabaey, J., 2002, A 300 μW 1.9 GHz oscillator utilizing micromachined resonators. *IEEE Proceedings of the 28th ESSCIRC*, **28**, September 2002.

[5] www.dust-inc.com, 2004.

[6] www.xbow.com, 2004.

[7] www.xsilogy.com, 2004.

[8] www.ember.com, 2004.

[9] www.millenial.net, 2004.

[10] Roundy, S., Wright, P. K. and Rabaey, J., *Energy Scavenging for Wireless Sensor Networks with Special Focus on Vibrations*, Kluwer Academic Press, 2004.

[11] Kang, S., Lee, S.-J.J. and Prinz, F. B., 2001, Size does matter: The pros and cons of miniaturization. *ABB Review,* **2,** 54–62.

[12] Li, H. and Lal, M., 2002, Self-reciprocating radio-isotope powered cantilever, *Journal of Applied Physics*, **92**(2), 1122–1127.
[13] Kordesh, K. and Simader, G., 2001, *Fuel cells and their applications*. VCH Publishers, New York.
[14] Holloday, J. D., Jones, E. O., Phelps, M. and Hu, J., 2002, Microfuel processor for use in a miniature power supply, *Journal of Power Sources*, **108,** 21–27.
[15] Strasser, M., et al., 2002, *Sens. Act. A*, **97–98C**, 528–535.
[16] Pescovitz, 2002, The power of small tech, *Smalltimes*, **2**(1).
[17] Stordeur, M. and Stark, I., 1997, Low power thermoelectric generator—Self-sufficient energy supply for micro systems, *16th Int. Conf. on Therm*, pp. 575–577.
[18] Shenck, N. S. and Paradiso, J. A., 2001, Energy scavenging with shoe-mounted piezoelectrics, *IEEE Micro*, **21,** 30–41.
[19] Randall, J. F., *On ambient energy sources for powering indoor electronic devices* Ph.D. thesis Ecole Polytechnique Federale de Lausanne, Switzerland, May 2003.
[20] Federspiel, C. C. and Chen, J., 2003, Air-powered sensor. *Proceedings of IEEE Sensors 2003*, Toronto, October 22–24, 2003.
[21] Shearwood, C. and Yates, R. B., 1997, Development of an electromagnetic micro-generator, *Electronics Letters*, **33**(22), 1883–1884. (IEEE, 23 October 1997)
[22] Amirtharajah, R. and Chandrakasan, A. P., Self-powered signal processing using vibration-based power generation, *IEEE JSSC*, **33**(5), 687–695.
[23] El-hami, M., Glynne-Jones, P., White, N. M., Hill, M., Beeby, S., James, E., Brown, A. D. and Ross, J. N., 2001, Design and fabrication of a new vibration-based electromechanical power generator, *Sensors and Actuators A (physical)* **92,** 335–342.
[24] Ching, N. N. H., Wong, H.Y., Li, W. J., Leong, P. H. W. and Wen, Z., 2002, A laser-micromachined multi-modal resonating power transducer for wireless sensing systems, *Sensors and Actuators A (physical)* **97–98,** 685–690.
[25] Meninger, S., Mur-Miranda, J. O., Amirtharajah, R., Chandrakasan, A. P. and Lang, J. H., 2001, Vibration-to-electric energy conversion, *IEEE Trans. VLSI Syst*., **9,** 64–76.
[26] Miyazaki, M., Tanaka, H., Ono, G., Nagano, T., Ohkubo, N., Kawahara, T. and Yano, K., 2003, Electric-energy generation using variable-capacitive resonator for power-free LSI: Efficiency analysis and fundamental experiment, *ISLPED'03*, August 25–27, 2003, Seoul Korea.
[27] Mitcheson, P. D., Green, T. C., Yeatman, E. M. and Holmes, A. S., 2004, Architectures for vibration-driven micropower generators, *Journal of Microelectromechanical Systems*, **13**(3), 1–12.
[28] Glynne-Jones, P., Beeby, S.P., James, E.P. and White, N.M., 2001, The modelling of a piezoelectric vibration powered generator for Microsystems, *Transducers '01/Eurosensors XV*, Munich, Germany, June 10–14, 2001.

[29] Ottman, G. K., Hofmann, H. F. and Lesieutre, G. A., 2003, Optimized piezoelectric energy harvesting circuit using step-down converter in discontinuous conduction mode, *IEEE Trans on Power Elect*, **18**(2), 696–703.
[30] Roundy, S. and Wright, P.K., 2004, A piezoelectric vibration based generator for wireless electronics, *Smart Materials and Structures*, **13,** 1131–1142.
[31] Otis, B. and Rabaey, J., 2002, A 300 μW 1.9 GHz oscillator utilizing micromachined resonators, *IEEE Proceedings of the 28th ESSCIRC*, **28**, September 2002.
[32] Adams, S.G., Bertsch, F.M., Shaw, K.A., Hartwell, P.G., Moon, F.C. and MacDonald, N.C., 1998, Capacitance based tunable resonators, J. *Micromech. Microeng,* **8,** 15–23.
[33] Roundy, S., Toward self-tuning adaptive vibration based micro-generators, Smart Structures, Devices, and Systems II, **5649,** 373–384, 2004.
[34] Lesieutre, G. A. and Davis, C. L., Can a coupling coefficient of a piezoelectric device be higher than those of its active material?, *Journal of Intelligent Material Systems and Structures*, **8,** 859–867.
[35] *Structure and method for fabricating semiconductor structures and polarization modulator devices utilizing the formation of a compliant substrate*, US Patent 6,714,768, to Motorola, Inc., Patent and Trademark Office, 2004.

Chapter 5.2

POWER MANAGEMENT OPTIONS FOR AmI APPLICATIONS

Derk Reefman and Eugenio Cantatore
Philips Research, Eindhoven
{derk.reefman, eugenio.cantatore}@philips.com

Abstract An overview of power management (PM) for ambient intelligence (AmI) applications will be presented. The basic principles of one the most important blocks in PM, the DC:DC converter, will be demonstrated, followed by a discussion of how dedicated, seamless integration of PM with specific applications, such as transmitter power amplifiers and digital cores, could lead to a substantial system efficiency improvement. As one of the key driving forces in AmI is presented by the required miniaturization, attention will be paid to the hurdles that need to be overcome in PM to realize this miniaturization. The power source in AmI applications is extremely important, and the relation between various energy scavengers and their interaction with the PM electronics is discussed extensively. Finally, conclusions on power management and energy scavenging for AmI applications are presented.

1. INTRODUCTION

Power management has always been an integral part of the design of electronic equipment. In the early days of electronics, power management was a discipline with primary area of focus on how to provide all electronics with:

- sufficient power; and
- sufficient quality power

where no emphasis was put on either the size of the power supply or the efficiency of the power supply. Whereas, originally this was still a highly nontrivial exercise [1, 2]; the situation changed substantially in the 70s and 80s, when power supply became a commodity still ignoring aspects as

S. Mukherjee et al. (eds.), AmIware, 285–313.

efficiency and size. The main reasoning here was the fact that, even the simplest functions, such as AM/FM radio, were consuming significant amounts of power—also digital functions were highly expensive in terms of energy consumption. As an example, one can consider the quad NAND from the 74 TTL series, the 7400, which consumed no less than 60 mW at 5 V [3] in 1975. The enormous reduction in power consumption of electronic functions—and, in particular for digital functions—have led to a situation where mobile equipment started to be feasible, making power management much more relevant than ever before. This is illustrated in Figure 5.2-1, where a high-level categorization of a modern IC system is presented.

In Figure 5.2-1, it is clear that the analog and digital electronics cannot be viewed independently of each other anymore, and that integral power management needs to be considered on a system level. Apart from the proliferation of mobile equipment necessitating a reconsideration of power management, another mostly consumer electronics-driven phenomenon has appeared for mains-powered equipment. In many applications, the size or other aspects of the device starts to become dominated by cooling of the electronics. This trend is exemplified in, for example,

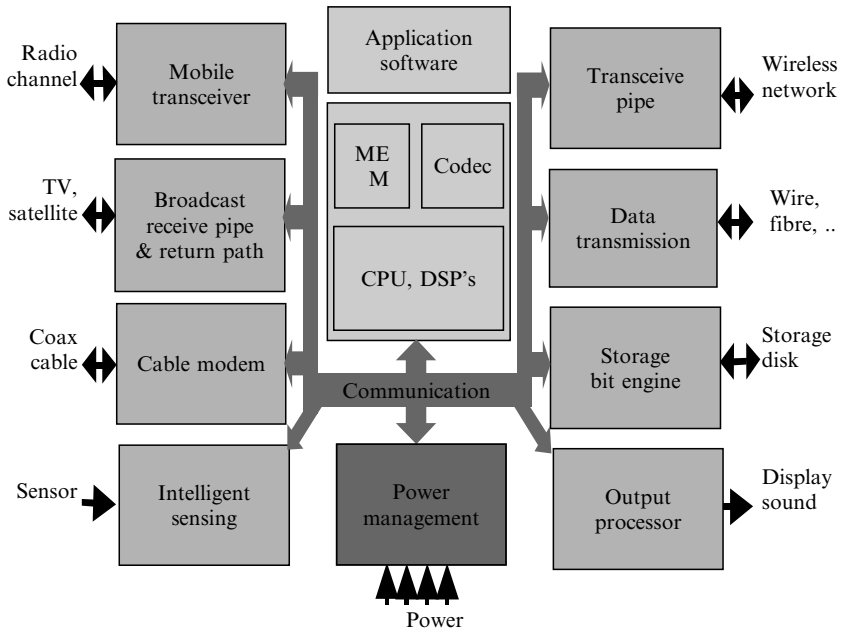


Figure 5.2-1. System overview representative of most modern electronic applications.

multi-channel power amplifiers for DVD and SA-CD players, where typically consumers do not want the size of the 6-channel amplifier to exceed that of the original stereo-box, even though three times as much power is delivered. Another example is the desktop PC, where the cooling fan is responsible for most of the noise of the PC. It is not surprising that in this area emphasis is on power management schemes that reduce the need for fan-cooled power supplies.

At this point in time, the situation is reached, where (if cost were not a sometimes major driving factor) power supplies can have efficiencies of over 90% and are thus no longer a bottleneck in the total power consumption/dissipation. The only route left to make another step in *system* efficiency is to create an intimate link between the power supply and the application. While this presents a clear benefit as optimized power strategies can be developed, at the same time, it conflicts with a general trend observed so far that calls for reuse. Reuse requires a modular approach to the design of various components, thus ruling out any possibility for close interaction between a power supply module and the other parts of the electronics.

The important consequence of this observation is that such an integration leads to a situation, where distributed power management dedicated and optimized for even small parts of the total system, becomes necessary. Such distributed power management in turn calls for small physical size of the power supply, which is a major issue since the power supply size is mainly determined by the size of coils and capacitors, which are determined by laws of physics rather than technology.

A further new development is the upcoming world of energy scavengers, devices that can generate electric energy from energy present in the environment. For these devices, miniaturization is paramount, and in addition, power is supplied not by a *well-behaved* battery, but by a rather badly behaved scavenger. Depending on the type of scavenger, the output voltage can vary between hundreds of volts and millivolts. To complicate things even more, the load line of scavengers is typically a nonlinear function of the energy influx.

To set the scene for the remainder of the chapter, Figure 5.2-2 shows the data rate that can be achieved, as a function of the power consumption for various applications. The clear trend is that power increases linearly with data rate, which basically is based on laws of physics: the only thing that could possibly change is the proportionality factor.

The next section deals with basic principles of power management (with emphasis on the factors determining efficiency) to keep the paper self-contained. In the remainder of this chapter, focus will be primarily on the issues identified above:

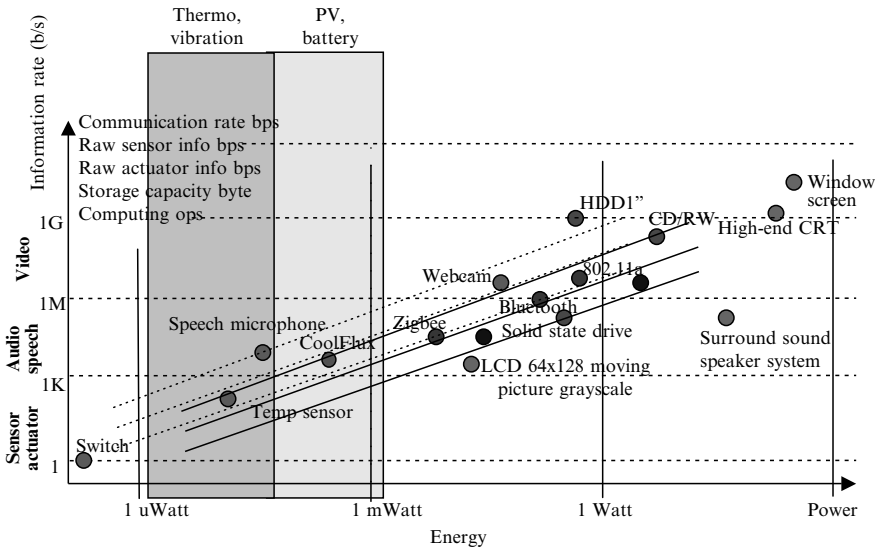


Figure 5.2-2. Data (processing) rate for various devices as a function of the energy consumption. Power is estimated for batteries or scavengers of 1 cm^3 (or 1 cm^2 for a PV). Battery lifetime is assumed 1 year. A guide to the eye for the current situation is provided by the full lines. Dotted is an extrapolation, based on the availability of optimal power management (see Section 3).

- Integration of power supply and application; some case studies to identify potential efficiency gains;
- Size reduction of the power supply unit necessary for integration
- Energy scavenging; what energy amounts can be harvested in what volumes?

The above is not an exhaustive enumeration of topics that are within the scope of power management. For example, battery management is not covered here, but is an increasingly important issue for handheld equipment. For an overview of recent advances in battery management, the reader is referred to [4]. A topical overview of power management is provided in [5].

2. ELECTRICAL CONVERSIONS

In a power management unit, in principle, two types of electrical energy conversion can occur: AC:DC conversion and DC:DC conversion, where the input and output voltages are different. In principle, AC:DC

conversion is straightforward, and can be performed using synchronous rectification [6]. However, this relies on the fact that the available voltages are >0.5 V to overcome the threshold voltages of the active devices. Whenever the scavenger voltage is less than the required threshold voltage, this necessitates the presence of a battery, which supplies the required voltage.

2.1. Principles of DC:DC Conversion

In this section, some basics of power management will be outlined. For a much more extensive overview, the reader is referred to [6]. As high efficiency is paramount, the outline will be limited to inductive converters. DC:DC converters come in different flavors: the upconverter, down-converter, and up/down-converter, which create output voltages higher, lower, or either higher or lower, respectively, than the input voltage. In most handheld applications, both up and down conversion of the battery voltage are necessary. For a battery voltage of 3.6 V, which is a typical voltage for a Li–ion battery, down conversion is required to supply the digital core, which typically runs at 1.8 V or lower. Upconversion is needed for the supply of the display, which often requires voltages as high as 15–20 V for the backlight of the display. In Figure 5.2-3, a standard down- and upconverter topology are depicted.

The attractiveness of this type of converters is that—in principle—the energy conversion is lossless: the only elements in the converter are switches, coils, and capacitors, each of which does not dissipate.

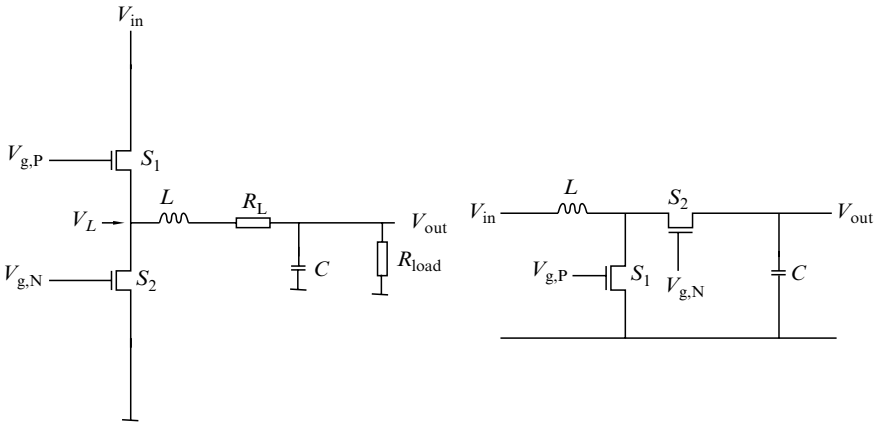


Figure 5.2-3. Standard topologies for a down (left) and up (right) converter.

In the following, the down (or Buck-) converter will be elaborated a bit more in detail. The switches S_1 and S_2 are turned on and off with a certain duty cycle D. When the switch S_1 is closed, and switch S_2 is open, the current through the coil L_1 increases, and energy is stored in the magnetic field around the coil that is built up by the increasing current. Furthermore, the capacitor is charged, and energy is stored in the electric field between the plates of the capacitor. This sequence is illustrated in Figure 5.2-4.

The top figure in Figure 5.2-4 shows how the switch S_1 is closed for a time DT_S, where D is the duty cycle. This is immediately followed by a time $(1 - D)T_S$, where the switch S_2 is closed an d S_1 is open. The voltage V_L is thus subsequently equal to the input voltage V_{in} for a time DT_S, and zero for a time $(1 - D)T_S$. The resulting coil current is indicated in the second part of Figure 5.2-4, where the maximum current is reached

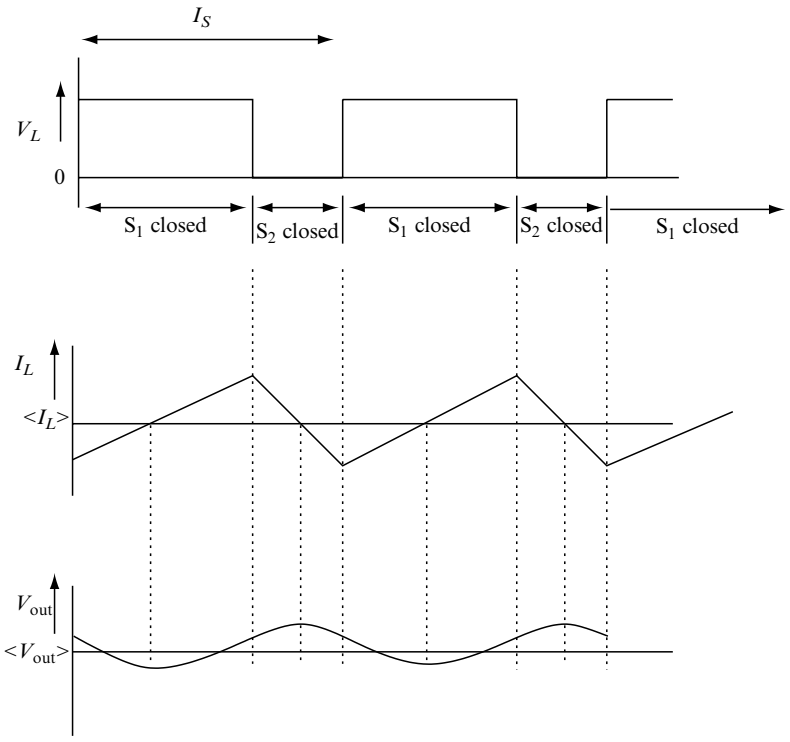


Figure 5.2-4. Waveforms occurring in a Buck converter. Upper graph: voltage at the *VL* node; center: current through the coil; and bottom: output voltage Vout.

at the moment when S_1 is opened (and S_2 closed), and the minimum current when S_2 is opened. The resulting voltage ripple at the output terminal V_{out} is indicated in the lower part of Figure 5.2-4. Typically, the design of a DC:DC converter is such that the output ripple amounts to a few mV.

2.2. Efficiency in DC:DC Conversion

In Figure 5.2-5, several parasitic elements, which cause the efficiency of the Buck converter to be less than the ideal 100%, are shown.

The main elements causing energy dissipation in a Buck converter are parasitic series resistors and capacitors, which are subsequently charged and discharged during a single cycle. Whereas a capacitor in itself is not a dissipating element, the currents flowing through parasitic resistances due to the (dis-)charging process give rise to an energy loss, which is proportional to $f_s C_{\text{para}} V^2$, where C_{para} is the total parasitic capacitance and V is the voltage difference over the capacitor.

Typically, the series resistance R_P and R_N due to the $R_{\text{ds,on}}$ of the switches varies inversely proportional to the area of the switch, whereas the parasitic capacitance is proportional to this area. It can be shown that the energy that is dissipated can be written as [6]:

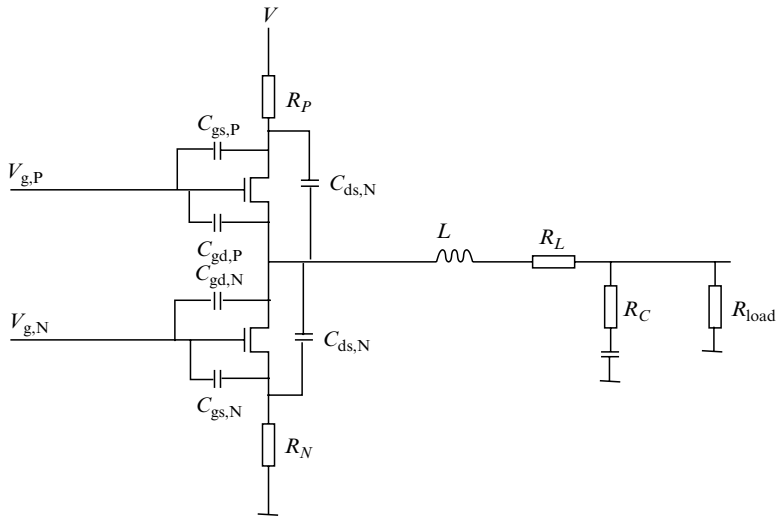


Figure 5.2-5. Some of the parasitics in a Buck converter, which cause the efficiency of the converter to be less than 100%.

$$P_{\text{loss}} = f_s C_{\text{para}} V_{\text{in}}^2 + I_{\text{DC}}^2 (R_{\text{ds,on}} + R_{\text{residual}}) + \frac{1}{192} \left(\frac{V_{in}}{f_s L} \right)^2 (R_{\text{ds, on}} + R_{\text{residual}}), \tag{5.2-1}$$

where $R_{\text{ds,on}} + R_{\text{residual}}$ is the total parasitic resistance from input terminal to output terminal, separated in the channel resistance of the FET ($R_{\text{ds,on}}$) and all other series resistances, such as bond wire resistance and coil resistance R_L. By varying the size of the power transistor, one has some degree of freedom to exchange capacitive losses against resistive losses.

As an illustration, in Figure 5.2-6, the efficiency of a small 40 mW Buck converter (with the lowest values for parasitics practically achieved today in a 20 V process) is depicted as a function of frequency, for three different optimizations: optimized for $f_s = 1\,\text{MHz}$, $L = 10\,\mu\text{H}$, $f_s = 10\,\text{MHz}$, $L = 1.0\,\mu\text{H}$, and $f_s = 60\,\text{MHz}$, $L = 0.1\,\mu\text{H}$. Clearly, the DC:DC converter with the larger coil ($10\,\mu\text{H}$) provides best peak efficiency. The $0.1\,\mu\text{H}$ coil based DC:DC converter displays an optimum efficiency of 78% around 60 MHz. It is important to remark that the values that are achieved in practical situations, are less than 0.9 times

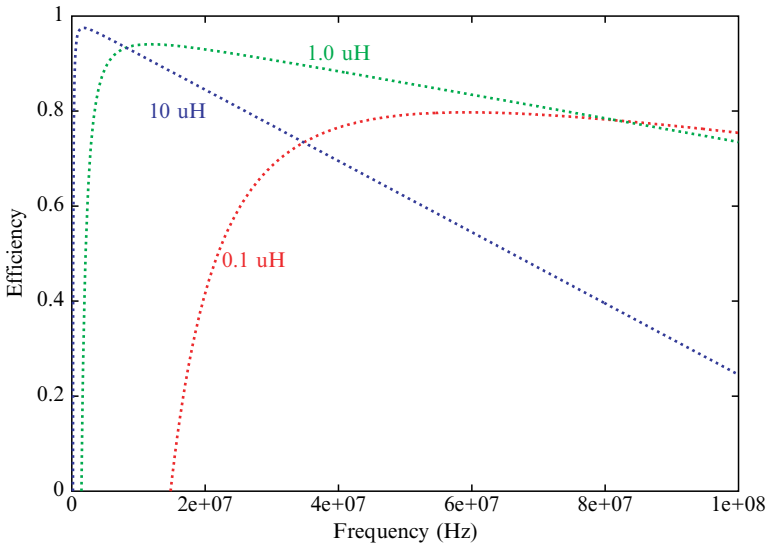


Figure 5.2-6. Peak efficiency of a buck converter for a load of 40 mW, for an inductance value of 10, 1 and $0.1\,\mu\text{H}$.

the values reported in Figure 5.2-6 due to the fact that Equation (5.2-1) ignores some sources of losses that become particularly important while approaching higher frequencies. Hence, from an efficiency point of view, the lower the switching rate, the better.

3. SEAMLESS INTEGRATION OF POWER MANAGEMENT

Seamless integration of power management in the application is close to the optimal (if not the optimal) way to maximize the energy efficiency of an application. In this section, two examples will be discussed: integration of the power supply with an RF power amplifier (PA) (for, e.g., mobile telephony, Bluetooth, etc.) and integration of the power supply in digital cores. Anticipating the discussion in the upcoming two sections, the dotted lines in Figure 5.2-2 already show an estimate of the maximum gain in efficiency that can be obtained with optimal power management.

3.1. Power Supply Optimization of an RF PA

Virtually all RF modulation schemes make use of both phase and amplitude modulation of the RF signal (e.g., GSM-EDGE, UMTS, Bluetooth, ZigBee, etc.). As a result, the RF PA is usually driven in class AB, because the linearity requirements of the PA for the above-mentioned modulation schemes are rather stringent. The typical power required for a UMTS PA is indicated in Figure 5.2-7.

The curve labeled ''Class AB'' shows power consumption as function of the required output power. Clearly, at low output powers, the efficiency is virtually zero: for example, at an output power level of 10 dBm, the efficiency is 5%. At the highest output powers, the efficiency is slightly higher than 35%. Because in typical situations output powers of about 10–15 dBm occur much more often than maximum output powers, the average efficiency is only 10–15%. This situation can be improved by employing a supply voltage to the PA, which is optimally suited for its required output level. In a standard configuration, the PA is used at a supply voltage of about 3.5 V (dictated by the use of a Li–ion battery), and the PA is designed to deliver maximum output at a voltage swing that corresponds to this 3.5 V. However, when only l0 mW is required, the voltage swing is much less, and the remaining large headroom leads to dissipation. However, the PA can deliver the power equally well when

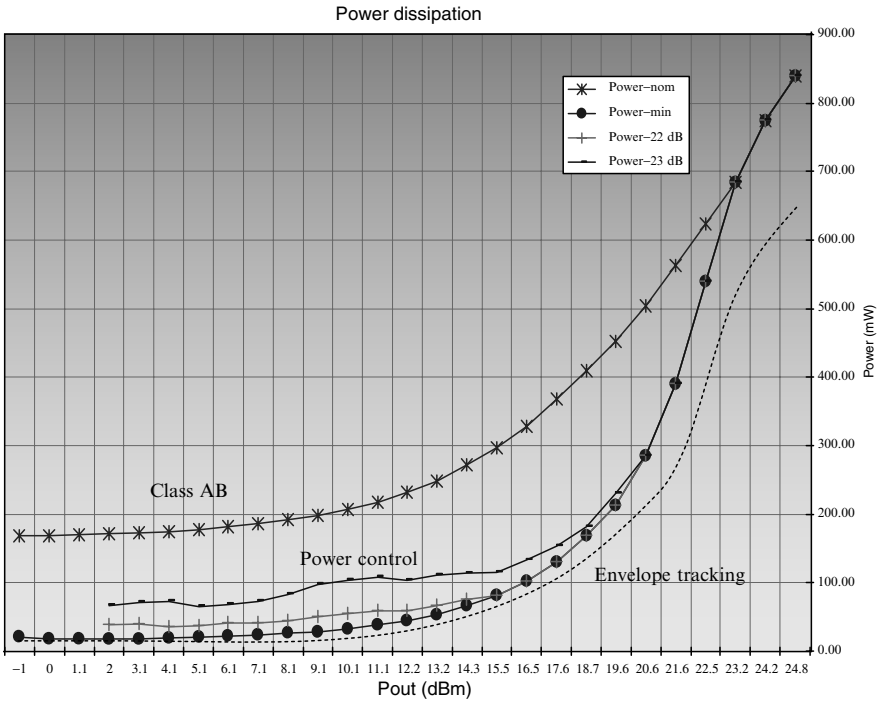


Figure 5.2-7. Power consumption for a typical UMTS PA. The output power is in dBm; 0 dBm is defined as 1 mW; hence, 24 dBm corresponds to about 0.25 W. Data points are measured, except for the dashed curve, which is based on theoretical estimates.

supplied with a much lower supply voltage, hence reducing the headroom due to which the dissipation is significantly less. Such a situation is depicted in Figure 5.2-8.

Clearly, some challenges remain in the control of the DC:DC converter, as it will have to supply the voltage and power, which still results in a sufficiently good signal at the output of the PA, while simultaneously optimizing the efficiency of the PA. To illustrate the degree of performance improvement that can be achieved in this way, the curve labeled ''Power control'' is added in Figure 5.2-7. While at large output powers, there is no or little improvement of the efficiency (which is expected on basis of the fact that the PA has been optimized for high output powers), significant improvement in the lower power areas is achieved. In practical use, the high power output levels are hardly used. Hence, the average efficiency of the system has improved by a factor of about 2 as compared to the situation without a DC:DC converter. However, a further improve-

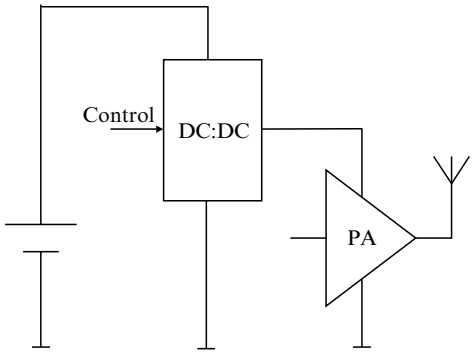


Figure 5.2-8. Simplified diagram of a DC:DC converter that tunes the supply voltage for the PA such as to optimize the overall efficiency.

ment is possible. In Figure 5.2-9, a typical envelope of a UMTS signal is given.

When the supply voltage tracks the required power envelope of the PA, the power efficiency of the PA is optimized. Virtually independent of the required output power, the PA's efficiency is at least 45%. Obviously, this puts significant requirements for the DC:DC converter, such as

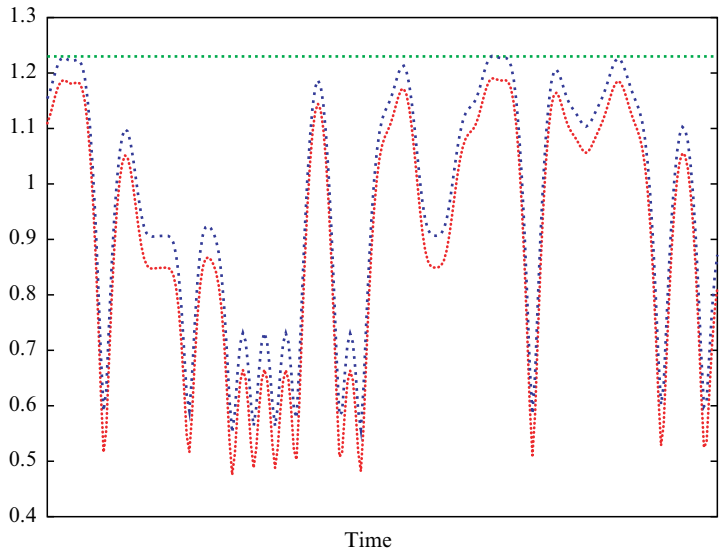


Figure 5.2-9. Example of a UMTS signal. In power control, the PA supply voltage is constant and just above the peak amplitude of the signal (straight line). In envelope tracking, an example power supply voltage is shown as a dotted curve.

bandwidths exceeding 10 MHz and very low ripple voltages (<0.2 mV). As a result of these requirements, such a system has not been proven in practice yet. Whereas for UMTS, efficiency increases by a factor 2 (power control) or even 4 (envelope tracking) are theoretically possible, these increases depend strongly on the power profile of the protocol. For example, in EDGE or Bluetooth, high output powers are occurring most frequently, and theoretical increases in efficiency are reduced to about 1.5 and 3. Further, limited DC:DC converter efficiency will reduce the efficiency gain even more.

3.2. Digital Core Power Supply

Power consumption in active digital cores is determined mostly by switching losses. Hence, the majority of power losses are described by the simple relation:

$$P_{\text{coreloss}} = f_{\text{clock}} C_{\text{tot}} V^2, \tag{5.2-2}$$

where f_{clock} is the clock frequency, C_{tot} the total capacitance that is charged/discharged per clock cycle, and V the supply voltage. As the delay of a logic cell is inversely proportional to the supply voltage [7], the lowest clock frequency that is needed to perform a certain DSP function determines the supply voltage that is necessary. In classical design, the supply voltage of the digital core is fixed to a level so that (in worst case) the core can meet its most tight specification. This means that in less-demanding situations, where the processing rate can be much less than the maximum rate, the supply voltage is significantly higher than required and causing unnecessary losses. Along the same lines as discussed in the previous section (Section 3), this leads to the desire to adjust the power supply voltage according to the load profile of the core.

Figure 5.2-10 shows how classically a DSP is operated when <100% of the maximum processing power is needed by simple duty-cycling: when (for example) 50% is needed, the DSP runs for 50% of the time and is turned off for 50% of the time to save energy. Alternatively, one could lower the supply voltage to such an extent that the resulting maximum processing speed matches with the required needs, due to which the supply voltage can be lowered. If the supply can be lowered by a factor a, this brings a power reduction of a factor a^2. Like the system discussed in the previous section, this calls for a close interaction between the power supply and the core, as is exemplified in Figure 5.2-11. A continuous calibration loop in the digital core determines whether the supply voltage should be increased, or is allowed to decrease a bit. So the power supply must (as in

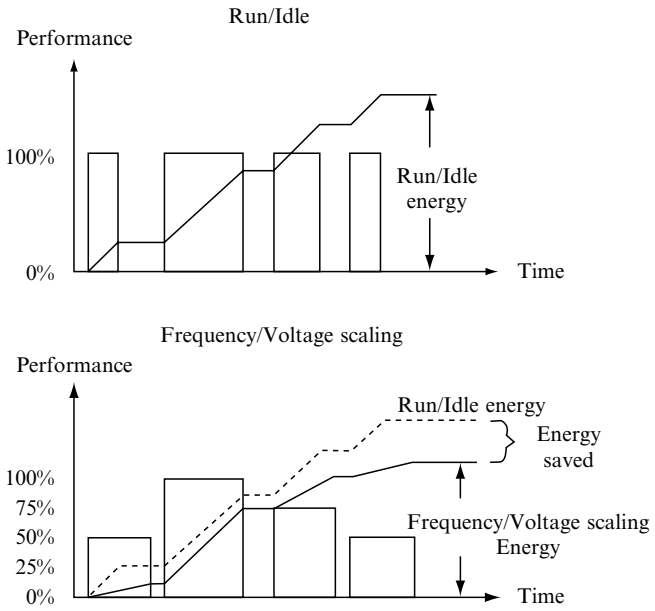


Figure 5.2-10. Comparison between classical situation, where low loads are dealt with by applying duty-cycle modulation and the alternative, where processing rate and supply voltage is scaled down (after [8]).

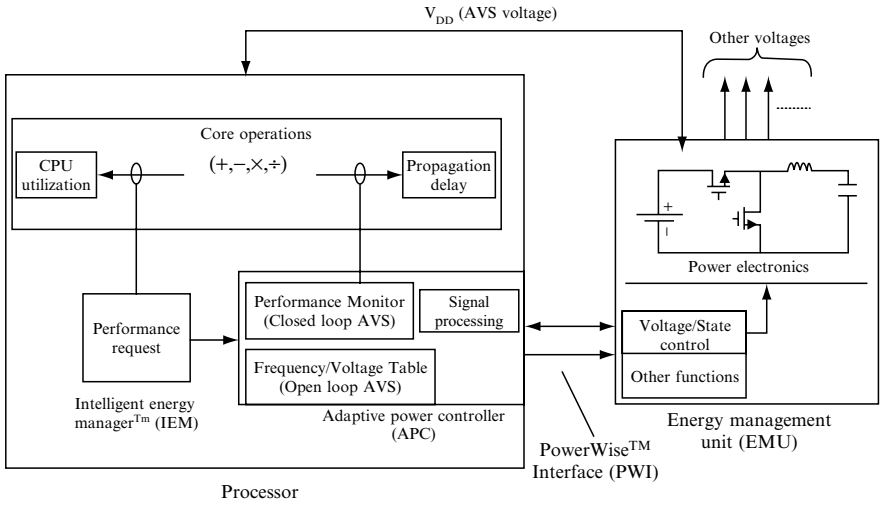


Figure 5.2-11. Intelligent power management of a digital core (ARM7, after [8]).

the previous section) accommodate severe requirements in terms of bandwidth.

In Figure 5.2-12, the power savings that can be achieved employing such a strategy are depicted. In particular when the required processing power is substantially less than the maximum, savings can be as large as a factor of 3. Still, such power savings heavily depend on the applications; typically, applications that already were not too power hungry, can be made even less power hungry whereas in case of applications that require a lot of power, not much can be done.

3.3. Integration Aspects

An issue which arises due to the above signaled developments is that, where previously the application (a PA, digital core, or anything else) was stand-alone, this is not true any more. From a practical point of view, this means that a single component will be replaced by a set of components (original application, DC:DC converter chips, coils, capacitors, etc.). In most cases, this also implies a multiplication of the required volume, which

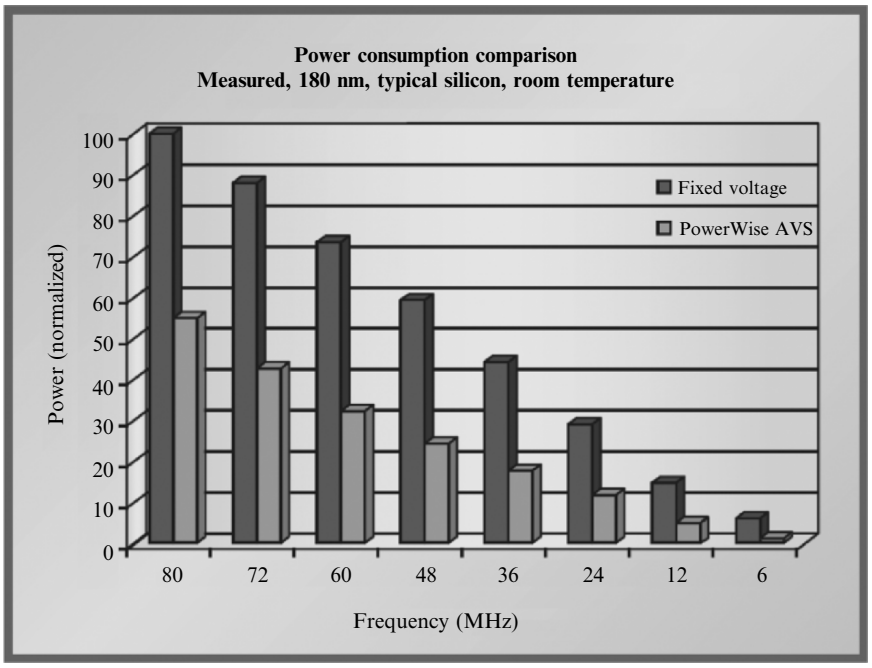


Figure 5.2-12. Efficiency gains due to intelligent power management of a digital core (ARM7, after [8]).

is an undesirable feature in general, and in particular in the field of ambient intelligence: devices need to "disappear" in the environment. As a result, this necessitates not only the integration of power management and application on a system level, but also on a physical level: again, the total system needs to be realized as a single package, which is of the size of the original application.

For the exemplary PA with integrated power management, the scope of this issue is still rather limited: only a single (albeit complex!) DC:DC converter is needed with some sophisticated control mechanism. For the situation where integration in digital cores is considered, the situation becomes potentially more complex. In fact, if it is beneficial to supply the total core with an optimized supply, it is even more energy efficient to split the core in several parts, each with their own optimized energy supply. This leads to a situation where a single package will need to contain a multitude of DC:DC converters.

Clearly, the only solution to this problem is in the miniaturization of the converter, in particular of the bulky external coil and capacitor. The following section will study this issue in more detail.

4. SIZE REDUCTION

The size of the DC:DC converter is typically determined by the size of the capacitance and the coil. As explained in Section Section 2, these elements serve as energy storage. The amount of energy E_C and E_L, respectively, that can be stored by a capacitor and a coil, is given by:

$$\begin{aligned} E_C &= \frac{1}{2}\varepsilon E^2 V_C \\ E_L &= \frac{1}{2}\mu B^2 V_L \end{aligned} \tag{5.2-3}$$

where E, B are the electric and magnetic fields, respectively, ε is the dielectric constant, and μ the magnetic permeability. Clearly, Equation (5.2–3) shows that the total amount of energy stored in a capacitor or coil is proportional to the volume V_C or V_L, respectively, over which the field (either electric or magnetic) extents. In the case of the Buck converter as discussed in Section 1.3, this means that the coil and capacitor need to provide the load with energy during the time $(1 - D)T_S$, when the switch S_1 is switched off. As this time is inversely proportional to the switching frequency f_s, it is clear that the volume of the energy storage elements bears the following relation to the switching frequency f_s:

$$V_C, V_L \propto \frac{1}{f_s}. \tag{5.2-4}$$

Clearly, in order to reduce the size of the L and C components, such that they can be integrated, the switching frequency needs to be increased. In addition to that, coils need to be developed that are capable of carrying the current required by the application without excessive DC or AC losses, which becomes potentially problematic for higher switching frequencies as discussed in Section 2. Integration of capacitors is much less of a problem, even though still far from trivial. In this section, focus will be on the design of coils suitable for DC:DC conversion, and the route toward higher switching frequencies.

4.1. Toward Smaller Capacitors

In Figure 5.2–13, a schematic representation of a Philips trench capacitor in silicon is depicted. While around 2001, trench capacitors were capable of providing about 10–20 nF/mm^2 of capacitance [9], current processes in Philips Semiconductors, Caen already provides almost double this value. It is further anticipated that the capacitor density will increase by more than a factor of 2 in the coming year, thus providing about 100 nF/mm^2. With these capacitances, integration in DC:DC converters is only a matter of time. For coils, the situation is rather different as will be elaborated in the next section.

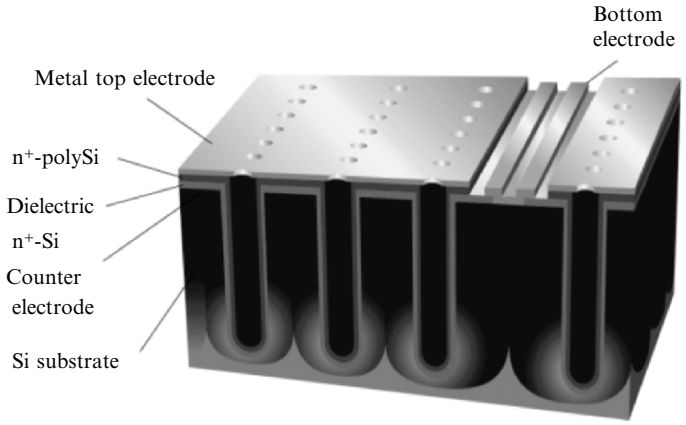


Figure 5.2-13. Trench capacitors in the PICS process.

4.2. Toward Smaller Coils

For the design of coils, several classical categorizations exist:

- Coils with/without magnetically active core material; and
- Planar (2D) and solenoidal (3D) coils.

To identify relative merits of each of these classes, some of the non-idealities of inductors need to be explored, in particular losses.

The following sources of energy losses in inductors can be identified:

(1) **DC copper losses.** The turns of a coil are usually made from copper, which has a nonzero resistance (for example, the specific resistance ρ_{Cu} of Cu equals $\rho_{Cu} = 1.68\,\mu\Omega$cm). Hence, the winding acts as a simple resistor due to this effect.
(2) **AC copper losses.** For high frequencies, the skin-effect becomes important. Due to this effect, the current flows only close to the surface (skin) and the effective resistance is higher. For copper, the skin-depth is 0.2 mm at 100 kHz, 0.02 mm at 10 MHz. In addition, the AC magnetic field induces eddy currents in the copper. The losses due to the eddy currents increase steeply as a function of frequency. Finally, the proximity effect also causes losses.
(3) **Hysteresis losses.** These losses arise from the sweeping of flux from positive to negative values. Hysteresis loss is due to the materials' intrinsic properties due to the energy used to align and realign the magnetic domains. Soft-magnetic materials are therefore preferred as core material.
(4) **Eddy current core losses.** These losses arise from the circulating currents within the magnetic materials due to differential voltage inside the cores itself. This loss is highly dependent upon the specific resistivity of the core material. The higher the switching frequency, the higher the eddy current loss. Due to this effect, the effective inductance will also decrease for increasing frequencies.

The major advantage of the use of a magnetically active material as core of a coil is in the fact that it *increases* μ. From Equation (5.2–3), it is observed that this is a desirable phenomenon, as higher μ implies less volume for identical inductance. However, the price that is paid for this advantage is additional sources of losses, as the last two items in the aforementioned sources of energy loss occur only for magnetically active core material. In practice, it appears extremely difficult to obtain exact closed-form expressions for the losses in a coil. An often used approximation for the total losses due to sinusoidal currents is given by:

$$P_L \approx a \cdot f^c \cdot B^d, \tag{5.2-5}$$

where f is the frequency of the current. Typical values for c, d are:

$$2 \leq c \leq 4;\ 2.5 \leq d \leq 3, \tag{5.2-6}$$

where c typically tends to be larger for larger f. To balance the relative merits of a magnetic core, the increase in μ should be weighed against the increase in core losses. Among the best high frequency ferrites is 4F1, a Philips ferrite currently sold by Philips spin-off Ferroxcube [10]. In Figure 5.2–14, the permeability is depicted as a function of frequency.

In Figure 5.2–14, both the real component μ_s' of the complex series permeability as well as the complex part μ_s'' is indicated. These are related to the ohmic series resistance R_L (that can be thought of as in series with the inductance L, see Figure 5.2–5) and reactance $2\pi fL$ as follows:

$$\frac{\mu_s''}{\mu_s'} = \frac{R_L}{2\pi fL}. \tag{5.2-7}$$

From Figure 5.2–14, we can now estimate the highest switching frequency that can be realized without excessive core losses. The current

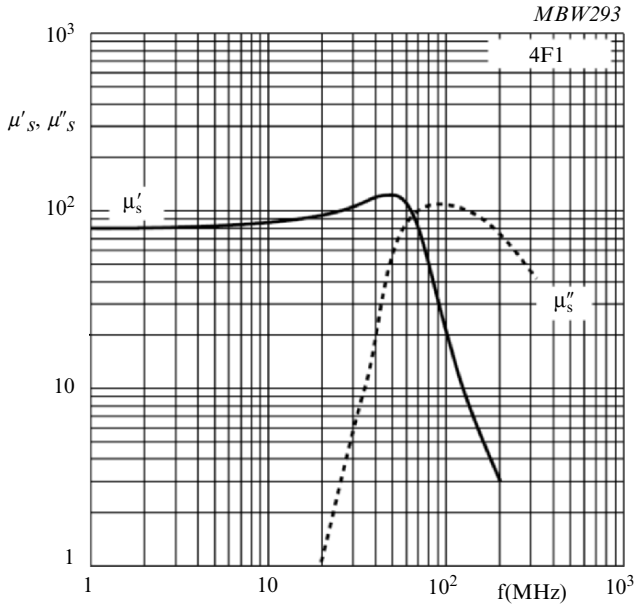


Figure 5.2-14. Real and imaginary part of the permeability of the ferrite 4F1 (from [10]).

waveform in the inductor of a DC:DC converter is typically triangular (see Section 2), meaning that the third harmonic is still very relevant. With Figure 5.2–14, this indicates that core losses are less than 6% only for $f_{sw} < 7\,\text{MHz}$, where losses due to other effects have not been accounted for yet. However, μ_s' is still of the order of 80, creating a potential reduction in coil size by this factor. In absence of such core material, the switching frequency should be increased by a significant amount to obtain the same ripple at the DC:DC converter output. The question is whether the use of these materials, introducing additional technology steps, is economically feasible. In particular when integration is involved, a question arises — whether a technology flow exists that can deal with elements, such as Fe, Ni, Zn, and Mn, which are normally seen as the most dangerous materials in a clean room. Further, the question about mechanical stability arises: thermal expansion coefficients need to be identical for the ferrite and the system in which the ferrite is integrated. A recent development to tackle this latter problem is to disperse ferrite particles in a polymer matrix [11]. However, due to this dilution, the effective μ goes down by a factor, which can be as large as 10, reducing the benefit of the use of such materials.

To keep compatibility with typical fab rules, a large focus has been given to the design of ''coreless'' coils: coils without a magnetically active material. In this case the crucial problem is to get a significant inductance without running into the problem of too much DC resistance. To good approximation, the inductance of coreless solenoids and planar coils is given by [12]:

$$\begin{aligned} L_{\text{solenoid}} &\approx \frac{\mu_0 n^2 \pi a^2}{b} \\ L_{\text{planar}} &\approx 10^{-9}(H/m)\frac{n^2 a^2}{0.2a + 12c} \end{aligned}, \tag{5.2-8}$$

where a is the diameter of the coil, b its length, and c the inner diameter. The equation for planar inductors is phenomenological and not based on fundamental physics. For both solenoidal and planar coils, the inductance varies quadratically with the number of turns. However, the resistance due to the total length of all turns has different dependencies:

$$\begin{aligned} R_{\text{solenoid}} &\propto n \\ R_{\text{planar}} &\propto n^2, \end{aligned} \tag{5.2-9}$$

which means that the optimal inductance/resistance ratio is obtained for a 3D solenoidal coil. However, as current silicon technology is optimized for

2D structures, practical realizations may be restricted to 2D or quasi-3D geometries only. Typical values for inductances and resistances are $0.25\,\mu H/7.1\Omega$ for a planar coil, and $0.035\,\mu H/0.03\Omega$ for a solenoidal coil, both about $0.5\,mm^2$, which gives a flavor of what in currently used technologies is possible.

5. ENERGY SCAVENGING

In a sense, energy scavenging is the focus point of the trends identified and discussed in the preceding sections. In a microsystem based on energy scavenging, a seamless integration of energy source, application, electrical energy conversion, and power management are realized, all with the aim of miniaturization.

In Figure 5.2–15, a schematic representation of the energy and information flow in a typical application based on energy scavenging is depicted.

5.1. Overview of Energy Scavengers

Several kinds of devices have been proposed in the literature to convert environmental energy in electric energy. These devices range from the

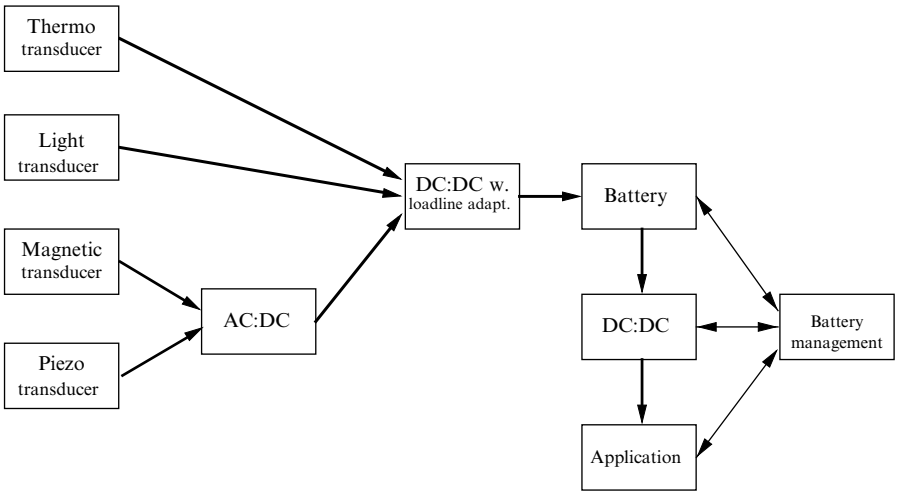


Figure 5.2-15. Schematic diagram of the energy (thick arrows) and information (thin arrows) flow in an application based on energy scavenging. While a multitude of energy scavengers are indicated, in practice, probably only one or two will be employed.

well-known and mature photovoltaic cells to thermoelectric scavengers (based on thermocouples), and devices able to convert mechanical energy present in the environment as vibrations (machines, domestic appliances), large movements (on the human body, for instance), etc . An overview of energy scavengers is presented in [13] and [14]. Some of the most important aspects of these scavengers, like achievable energy density, internal impedance and typical output voltage levels, are summarized in Figure 5.2–16.

5.2. Power Management in a Microsystem Powered by an Energy Scavenger

As shown in Figure 5.2–15, a microsystem that obtains its energy from a scavenger has to comprise at least two DC:DC converters. The first converter (DC:DC load line adaptation in Figure 5.2–15) has two functions:

- It transfers energy from the voltage level available at the scavenger output to the voltage level needed to charge the battery; and
- It offers to the energy scavenger the optimal electric load in order to optimize the transfer of energy to the battery (load line adaptation).

The second converter (Figure 5.2–15) adapts the rather high voltage of the onboard battery to the biasing needs of the other (integrated) electronics in the system (typically < 2 V).

Voltage levels provided by energy scavengers can span a wide range: electromagnetic vibration scavengers typically output 1 to l00 mV [15], while macroscopic generators based on thick piezoelectric layers can produce voltages in the order of tens of kilovolts [16]. Depending on the kind of generator, moreover, the electric power can be available in time as a

Type of scavenger		Max.output power per cm^3 (or cm^2) [μW]	Internal impedance [Ω]	Typical voltage level [V]	Voltage dynamic
Photovoltaic	Indoor (200–1000 Lux)	1–1,000	Nonlinear	0.2 – 0.4	DC
	Sun	15,000	Nonlinear	~0.4	DC
Thermoelectric	ΔT = 2K	40	$4 - 10^6$	4 – 400 m	DC
Vibration	Electromagnetic	400	~100	<100 m	AC(1–10^3 Hz)
	Capacitive	40	$<10^6$ (per cm^2)	>50	Pulse (1–10^3 Hz)
	Piezoelectric	300	$10^3 - 10^5$ (per cm^2)	100 m – 100	Pulse/AC (1–10^3 Hz)

Figure 5.2-16. Characteristics of various energy scavengers.

sinusoidal, pulsed, or DC signal. A summary of the dynamic characteristics of the electric output for different types of energy scavengers is given in Figure 5.2–16.

Batteries used in microsystems powered by energy scavengers must offer a very small form factor together with the possibility to be recharged continuously without suffering damage. Solid-state thin-film Li–ion batteries have both these characteristics [14, 17], thus, they are ideal candidates to provide energy storage for energy scavengers. These batteries are ultrathin, have a typical energy density of $\approx 50\,\mathrm{mWh/cm^3}$, a self-discharge rate of less than 5% per year and a maximum voltage of $\approx 4.2\,\mathrm{V}$. Solid-state thin-film Li–ion batteries can be charged within 4 min, under the condition that the voltage across the battery must be kept above $\approx 3\,\mathrm{V}$ and below $\approx 4.2\,\mathrm{V}$ If these boundaries are exceeded, the battery experiences irreversible degradation mechanisms. This charging behavior simplifies the design of the converter that provides load line adaptation, as the output current waveform and the quality of the DC output voltage are not really a concern. It will be the task of the secondary converter to provide the electronics with bias voltages of the desired quality.

5.3. Load Line Adaptation

An accurate and direct way to achieve the maximum power transfer to the battery is to design the control of the load line adaptation converter so that it maximizes the current flowing into the battery. This control action, however, implies accurate and frequent measurement of the output current and the use of a rather complex control algorithm [18]. This control strategy does not seem to be adapted to an ultralow power implementation of the load line adaptation converter, with sufficient efficiency at power levels in the microwatt range. A simpler and possibly less energy-demanding approach would be to exploit the relationship between the open circuit voltage given by a generator and the voltage it delivers when the power transferred to the load is maximal. If the source impedance can be considered linear (resistive or complex), and the load impedance can be considered resistive, the voltage amplitude at maximum power transfer is just half that of the voltage amplitude provided by the generator without load. This is direct consequence of the maximum power transfer theorem applied to a resistive load [19]. In this case, to perform the load line adaptation, the DC:DC converter has just to control the input voltage V_{in} (Figure 5.2–17) to be half of the open circuit value to achieve maximum power transfer to the load. The open circuit voltage of the generator must be measured periodically, disconnecting the load, in order to adapt to possible variations in the level of available energy in the environment

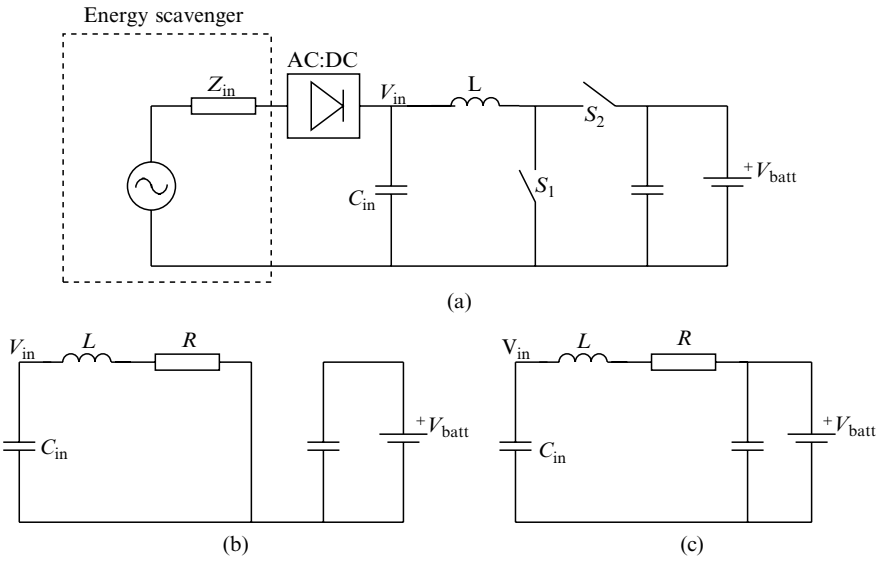


Figure 5.2-17. Schematic representation of an upconverter connecting the scavenger to the battery.

and ensure an optimal energy transfer at any time. This kind of control can be implemented without performing continuous and accurate measurements and is better adapted, in the view of the authors, to microwatt power levels. Such an approach would work, for instance, in the case of the piezoelectric generator presented in [18], where the maximum power transfer is indeed reached at a voltage level approximately equal to half the open circuit voltage. Other energy scavengers, however, do not provide the maximum power at half of the open circuit voltage. Some examples are:

- The energy scavenger has nonlinear output impedance. For instance, photovoltaic cells can be electrically modeled as a current source in parallel to a diode.
- In some vibration scavengers, a variation of the electric load causes an appreciable variation in the mechanic damping, so that the electrical characteristics of the generator depend on the load itself.

In the case of photovoltaic cells it is easy to verify from a set of I-V characteristics that a cell outputs the maximum power at a voltage, which is $\approx$70% its open circuit voltage [20], as shown in Figure 5.2–18. Measurements confirm that this ratio is kept over a wide range of illumination conditions. Many DC:DC converters developed for photovoltaic cells make use of this property and control the output voltage of the cell to a 70% fraction of its open circuit value to achieve maximum power

transfer to the load. A further advantage of the control of the input voltage to achieve maximum power transfer is that the power–voltage characteristic is in general rather flat around the maximum power point (Figure 5.2–18). The accuracy of the control and the size of the input buffer capacitance C_{in} (Figure 5.2–17) can thus be relaxed without losing much in performance.

5.4. Integration of the Load Line Adaptation Converter in a Microsystem Powered by an Energy Scavenger

In this section, issues related to the (physical) integration of a load line adaptation DC:DC converter together with the necessary passives in a $1\,cm^3$ volume at micropower levels are investigated. Calculations will be performed with reference to the following assumptions:

- Scavenger power output: $P_{in} = 10\,\mu W$;
- The input voltage of the load line adaptation circuit (=scavenger output), V_{in} is assumed to be less than the battery voltage V_{batt};
- The inductance of the largest coil that can be integrated on a $1\,cm^2$ area is estimated, on basis of the figures given in Section 4, to $L = 50\,\mu H$. Its resistance is $R_L = 100\,\Omega$;
- The largest capacitance per unit area available is $\approx 20\,nF/mm^2$;

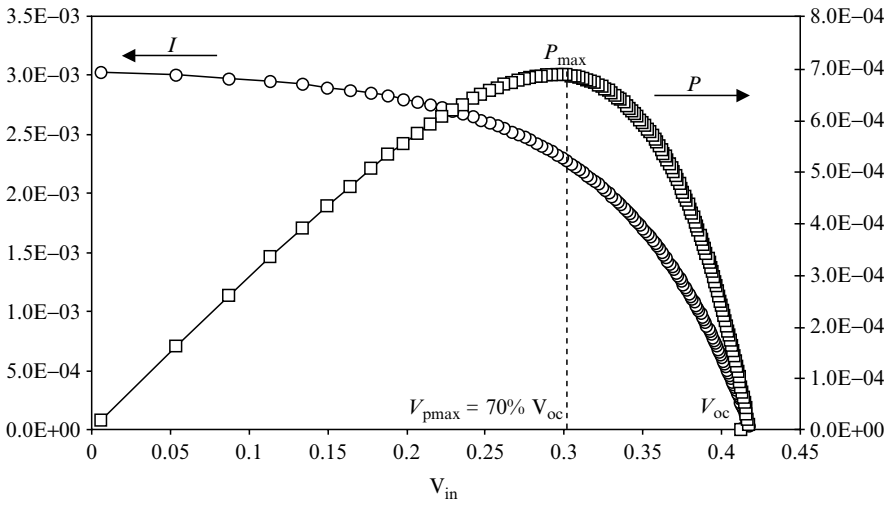


Figure 5.2-18. Measured power output of a photovoltaic cell as a function of its output voltage.

- The energy storage is realized with a solid state thin-film Li–ion battery, $V_{batt} = 4.2\,V$;
- The switches of the converter are small to enable high frequency operation; their resistance is $R_{on} = 100\,\Omega$ and their gate capacitance $C_{gate} = 100\,fF$; and
- The losses in the AC:DC converter will be neglected.

The load line adaptation converter is implemented with a simple boost scheme, being $V_{in} < V_{batt}$ (see Figure 5.2–17). During the first conversion phase, lasting a time t_1 switch S_1 is closed and S_2 is open. The inductor L drains power from the scavenger output (after AC:DC conversion, if this is needed), and charges, increasing its current. A capacitance in parallel with the scavengers output (C_{jn}, Figure 5.2–17(a)) provides the necessary charge, avoiding excessive V_{in} drop. In Figure (5.2–17(b)), an equivalent circuit of the DC:DC converter during this phase is presented. During the second conversion phase, t_2, switch S_1 is open and S_2 is closed. The inductor L can provide its current to the output and charge the battery. In Figure 5.2–17(c), an equivalent circuit of the DC:DC converter during time t_2 is presented. In Figure 5.2–17[(b) and (c)], the inductor and switch resistance have been lumped in a global parasitic resistance $R = R_{on} + R_L$. Assuming that C_{in} is large enough to consider V_{in} approximately constant, the time constant during the phase t_1 (Figure 5.2–17(b)) will be $\tau_{in} = L/R$. If the on-time of switch S_1, t_1 is larger than the time constant τ_{in}, the loop current will approach exponentially the value V_{in}/R with time constant τ_i, and the voltage across the inductor will approach with the same dynamic zero. In this case, most of the current drawn from the scavenger will be dissipated on the resistance R and will not be stored in the inductance. So, the DC:DC converter will be very inefficient. The converter can be efficient only if the on-time of switch S_1, $t_1 \ll \tau_{in}$, so that the current developed in the inductor can be transferred to the output without unnecessary dissipation on R. The time constant $\tau_{in} = L/R$ is in our case 0.25μsec. t_1 can still be made smaller than this value without excessive dynamic dissipation in the control electronics. A smaller value of L would be unpractical, as the control electronics should operate the switches with even shorter phases t_1 and t_2, resulting in too high power consumption. In conclusion, the largest value of L that can be integrated in $1\,cm^2$ is just large enough to allow a realistic speed for the control electronics.

The value of C_{in} required to have, at an average input voltage $V_{in} = 1\,V$, a ripple of l00 mV while the inductor is charging would be in this case 10 nF. This value can be integrated without problems within the given dimensions using high density integrated capacitance techniques (Section 4), or (at the cost of yield, and provided that the input voltage is low enough) with gate capacitance in CMOS chips.

A serious problem arises when the scavenger provides its energy at a voltage level <200 mV. This can happen because of intrinsic small internal impedance of the scavenger coupled to low output power levels (like it is common in electromagnetic vibration scavengers [15]). Even piezoelectric scavengers that normally provide energy at voltage level of volts can have very limited output voltage when the thickness of the piezoelectric layers is aggressively scaled down to the μm size, increasing the capacitance level. If the scavenger energy is available at voltages below 200 mV, its energy can be transferred to a higher voltage level with one of the following techniques:

- If the voltage generated by the scavenger is AC, it can be raised to higher levels with a transformer, to be then rectified (at a voltage level easy to handle), and further processed by the load line adaptation converter. In Figure 5.2–19, it is shown the efficiency of a transformer as a function of the frequency of the AC input signal and for two values of the internal impedance of the source, $Z_{in} = 100\,\Omega$, $10\,k\Omega$. The lines in black are calculated for an hypothetic integrated transformer with a self-inductance of the primary of $50\,\mu H$. The gray curves are calculated for a $4\,cm^3$ iron transformer with 100 turns in the primary and an iron core. Figure 5.2–19 demonstrates that it is impossible to build an efficient transformer in

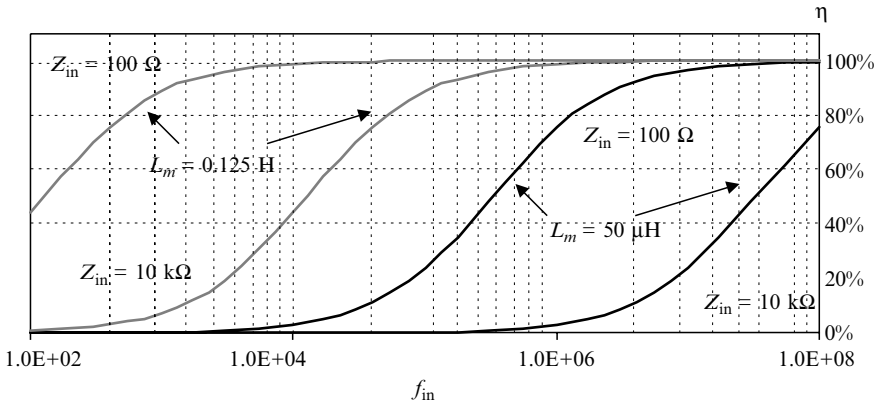


Figure 5.2-19. Efficiency of a transformer as a function of the frequency for different values of the source impedance Z_{in}. Black curves are obtained for a primary self-inductance $L_m = 50\,\mu H$, a value that can be obtained with an integrated implementation. Gray curves are obtained for a primary self-inductance of 0.125 H, which can be obtained with 100 turns on a $4\,cm^3$ iron core. Resistive and magnetic losses are neglected.

the frequency domain of the vibration scavengers (1 Hz–1 kHz) within a very reduced volume.

- If the voltage generated by the scavengers is DC, or it could be efficiently rectified, even at low voltage levels, another possibility would be to use just a DC:DC converter to increase the input voltage to the battery level. Figure 5.2–20 shows the efficiency of an upconverter as a function of the input voltage. The value used for the parasitic resistance of the inductor and the switches are the ones specified in this section. As one can see, the converter efficiency drastically decreases at values lower than 100 mV.

6. CONCLUSIONS

Several issues regarding power management for AmI purposes have been discussed. It has been shown that optimal power management is capable of reducing the power demands of various applications by about a factor of 2–4, depending on the type of application. These improvements shift the trend lines in Figure 5.2–2 to the dashed line position. A further substantial power reduction is physically impossible as the total efficiency of such systems has reached a level close to 100%. Thus, while significant work is in progress to realize this important increase in power

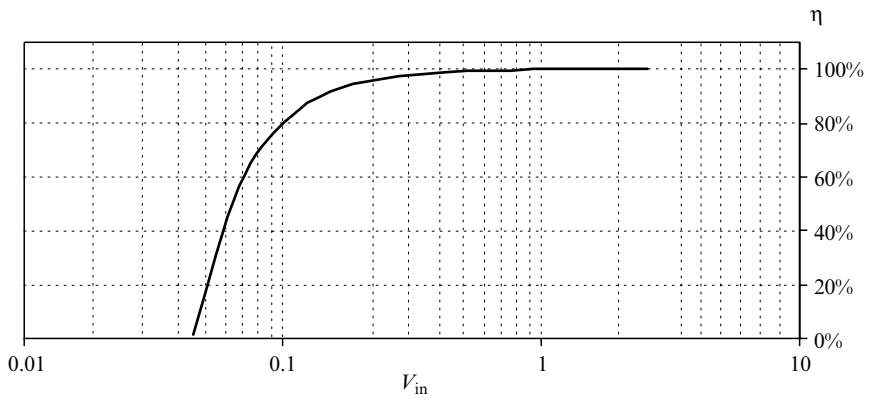


Figure 5.2-20. Efficiency of a DC:DC converter as a function of the input voltage. The parameters used are listed in (Section 5). The efficiency is calculated based on the assumption that the average inductor current is equal to the average output current of the scavenger, and that the inductor current ripple is negligible (continuous conduction mode).

efficiency, *it cannot be expected that applications will penetrate deeply in the μWatt region in Figure 5.2–2 due to further improvements in power management*.

Also, in order to make an important miniaturization step, *integrated coils need to be developed, concurrently with the development of DC:DC converters that switch at high frequency (≫ 10 MHz)*. This development is in contradiction with the requirement of increasing efficiency and calls for substantial work in the area of power management.

Further, it has been shown that the efficiency of energy extraction from energy scavengers is highly dependent on several scavenger characteristics, in particular the level of its output voltage in relation to the frequency of its output voltage. From a practical point of view, it is realistic only to build scavengers with output voltages of several hundreds of mV, or otherwise the total efficiency of the energy scavenger will drop to levels below 10%. Alternatively, the frequency of the scavenger output must be increased to values significantly larger than 10 kHz. Typically, this rules out the use of most miniaturized vibrational energy scavengers, as well as many thermoelectrical energy-based scavengers [21]. *This observation shows that the development of new energy scavenger types must be conducted in close collaboration with the power management electronics*.

Concluding at this point in time, the only feasible power sources for most (miniaturized) systems for autonomous AmI devices are either photovoltaic cells or batteries.

ACKNOWLEDGMENT

We thank the European Commission for funding of this work under contract IST-1–507911 (VIBES)

REFERENCES

[1] Lee, R., 1947, *Electronic Transformers and Circuits,* John Wiley, New York.
[2] Marti, O. K. and Winograd, H., 1930, *Mercury Arc Rectifiers,* McGraw-Hill, Norwell, Massachusetts.
[3] ITT Semiconductors, 1975, *Integrated Circuits TTL 74. series*. ITT Semiconductors.
[4] Bergveld, H. J., Kruijt, W. S. and Notten, P. H. L., 2002, Battery management systems-design by modelling, *Philips Research Book Series*, Volume 1, Kluwer Academic Publishers, Boston.
[5] Schoofs, F. A. M., 2004, Power management electronics, *in* J. H. Huijsing, M. Steyaert and A. Roermund (eds), *Analog Circuit Design,* Kluwer Academic Publishers.

[6] Erickson, R. W. and Maksimovic, D., 2001, *Fundamentals of Power Electronics,* Kluwer Academic Publishers, Norwell, Massachusetts.
[7] Burd, T., and et al., 2000, A dynamic voltage scaled microprocessor system, *IEEE International Solid-State Circuits Conference, Digest of Technical papers,* San Francisco, CA, USA, February 7–9, 2000, pp. 294–295, 466.
[8] Hartman, M., 2002, Powerwise adaptive voltage scaling minimizes energy consumption, *National Semiconductors white paper,* available at *nsc.com*
[9] Roozeboom, F., 2002, Mos decoupling capacitor chips of 100 nf and $20\,\mathrm{nf/mm^2}$ specific capacitance on a printed circuit board, *Proceedings of the 34th Int. Symp. on Microelectronics (IMAPS 2001)*.
[10] Ferroxcube, 2002, *Soft Ferrites and Accessoires–2002 Handbook*. Ferroxcube International Holding.
[11] Waffenschmidt, E., 2003, *Ferrite polymer compounds for printed circuit board integrated inductors,* PFL-Aachen Report 1679/2003, Aachen.
[12] Burkett, F. S., 1971, Improved designs for thin film inductors, *Proceedings of 21st Electronic Components Conference,* May 10–12, 1971,Washington, DC.
[13] Roundy, S., 2004, Energy scavenging in support of ambient intelligence: Techniques, challenges and future directions, *Proceedings of the 37th Int. Symp. on Microelectronics (IMAPS 2004)*.
[14] Ouwerkerk, M., 2004, Ubiquitous power, *IEEE Trans. on Power Electronics,* **19**.
[15] Kulah, H. and Najafi, K., 2004, An electromagnetic micro-power generator for low-frequency environmental vibrations, *Proc. 17th IEEE International Conference on MEMS,* Maastricht, 2004.
[16] Engel, T. G., Keawboonchuay, C. and Nunnally, W. C., 2000, Energy conversion and high power pulse production using miniature piezoelectric compressors, *IEEE Trans. on Plasma Science,* **28**.
[17] Frontedge, *http://www.frontedgetechnology.com*
[18] Ottman, G. K., Hofmann, H. F., Bhatt, A. C. and Lesieutre, G. A., 2002, Adaptive piezoelectric energy harvesting circuit for wireless remote power supply, *IEEE Trans. on Power Electronics,* **17**.
[19] Desoer, C. A., 1983, A maximum power transfer problem, *IEEE Trans. on Circuits and Systems,* CAS-30.
[20] Noh, H. J., Lee, D. Y. and Hyun, D. S., 2002, An improved MPPT converter with current compensation method for small-scaled pv applications, *IEEE IECON Proc., 2002*.
[21] Snyder, G. S., Luu, J. R., Chen-Huo, H. and Fleuriel, J. P., 2003, Thermoelectric micro device fabricated by a mems-like electrochemical process, *Nature Materials,* **2**.

Chapter 5.3

RECHARGEABLE BATTERIES
Efficient Energy Storage Devices for Wireless Electronics

P.H.L. Notten
Philips Research Eindhoven, Eindhoven University of Technology
peter.notten@philips.com

Abstract Batteries are indispensable in our present-day portable society. In order to feet the wide variety of portable and wireless electronic equipment we can nowadays rely on various battery systems, each having its own specific advantages and disadvantages. In this contribution, the basic principles of the most popular battery systems are reviewed, including Nickel–MetalHydride, Nickel–Cadmium, and Lithium–ion.

Keywords rechargeable batteries; Nickel–MetalHydride; Nickel–Cadmium; Li–ion

1. INTRODUCTION

Rechargeable batteries are energy storage devices, which are able to convert chemically stored energy into electrical energy during discharging and vice versa during recharging. The application of batteries to provide portable equipment with electrical energy has been rapidly growing during the last decades. Various types of commercially available batteries are used nowadays, varying from small button-type cells used in small-size electronics to batteries for hybrid cars and other large-scale electrical storage applications. Two classes of batteries can, in principle, be distinguished (see Table 5.3-1).

The first group is formed by the primary cells. These types of batteries can, in general, not be recharged and are therefore considered as non-rechargeable. The most popular member of this class, often applied in

S. Mukherjee et al. (eds.), AmIware, 315–345.

Table 5.3.1. Open-circuit potentials and energy density values for various primary (nonrechargeable) and secondary (rechargeable) batteries.

	System	Open-circuit potential (V)	Energy density*	
			Wh/kg	Wh/l
Primary Batteries	$Zn - MnO_2$	1.6	80–150	300–400
	Zn–air	1.65	300–400	800–1300
Secondary Batteries	SLA	2.1	30–40	70–80
	NiCd	1.3	35–65	100–200
	NiMH	1.3	40–90	160–310
	Li–ion	3.8	100–160	200–400

*Energy densities are strongly dependent on discharge rate.

many portable electronics, is the alkaline Zinc–Manganesedioxide ($ZnMnO_2$) cell. Another example is the Zinc–air (Zn–air) button cell, commonly used in hearing aids, although formally speaking, this system should be positioned in between a battery and a fuel cell system, as the oxygen electrode is based on the fuel cell concept of "external" chemical storage rather than the battery concept, relying on "internal" chemical storage inside the battery electrodes.

The second group is formed by the so-called secondary cells. These can be recharged once they are partly or completely discharged. During recharging, electrical energy is converted again into chemical energy by means of a charging device. It is evident that for applications, which frequently need a lot of "portable energy," such as cellular telephones, laptop computers, PDA's, and electrical shavers, rechargeable batteries are preferred, not only for economical reasons but also for our convenience. This chapter will therefore focus on rechargeable batteries only.

Various types of rechargeable batteries are available and the number is still expanding. The most popular types are, at the moment, the conventional Nickel–Cadmium (NiCd) battery, the high-energy dense Nickel–MetalHydride (NiMH) battery, and the most recently developed Lithium (Li–ion) batteries. There are very large differences in the characteristics and performances between the various systems. This becomes clear when one considers, for example, the battery open-circuit voltage and the energy densities of the considered systems in Table 5.3-1 [1–3]. It is obvious that these parameters may have a different impact on the electronic design of portable equipment. It should, however, be emphasized that in order to make a proper battery choice for a particular application, one has to deal with a wide variety of battery characteristics. Many aspects should already

be considered in the very early stages of the development phase. Figure 5.3-1 gives a glimpse of the various parameters, which may be taken into account.

It is an understatement to say that application of portable energy in cordless versions of consumer electronics will become even more important in the near future than is already the case now. A further expansion into the direction of *Autonomous devices* for *Ambient Intelligence* is also foreseen. Knowledge about the performance of the various battery systems is therefore indispensable. In this contribution, we will focus on the basic electrochemical principles and characteristics of the most important systems, starting with the most popular rechargeable systems. Since Philips Research can be considered as the inventor of the NiMH battery, and since NiMH batteries has taken over the NiCd market during the last decade, we will start explaining the basic electrochemical principles of this battery type. Subsequently, the differences in concepts of NiCd and Li–ion will be described. Apart from the various concepts, some electrochemical characteristics typical for these systems will be addressed as well, including, for example, their charge and discharge performance, self-discharge, and occurring memory effects.

2. NICKEL–METALHYDRIDE BATTERIES

2.1. Basic Reactions

A schematic representation of a NiMH battery containing an AB_5-type hydride-forming electrode is shown in Figure 5.3-2 [2, 3]. The electrodes

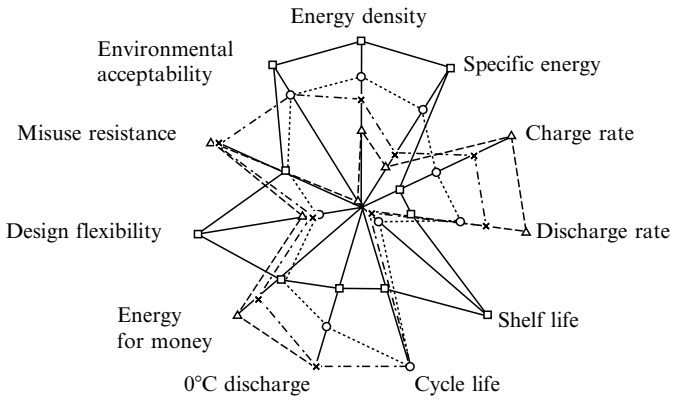


Figure 5.3-1. A wide variety of parameters characteristic for conventional and future type of batteries, which should be taken into account already in the very early development stage of new products.

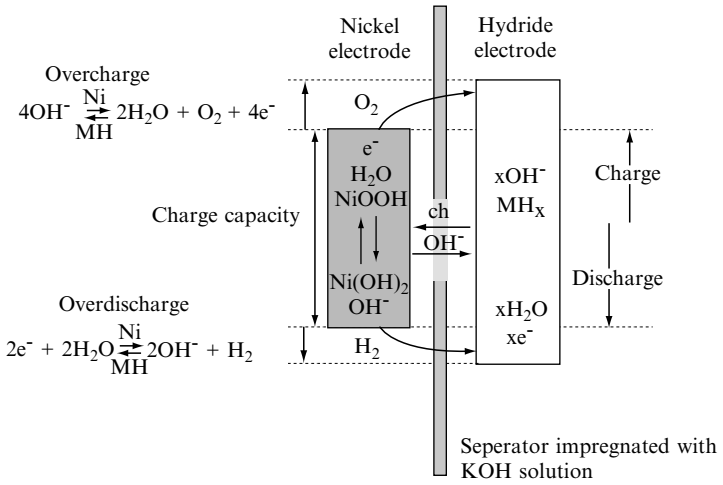


Figure 5.3-2. Schematic representation of the concept of a sealed rechargeable NiMH battery.

are electrically isolated from each other by a separator. Both separator and electrodes are impregnated with an alkaline solution, which provides for the ionic conductivity between the two electrodes. The overall electrochemical reactions, occurring at both electro during charging (ch) and discharging (d) can, in their most simplified form, be represen by:

$$Ni(OH)_2 + OH^- \underset{d}{\overset{ch}{\longleftrightarrow}} NiOOH + H_2O + e^-, \tag{5.3-1}$$

$$AB_5 + xH_2O + xe^- \underset{d}{\overset{ch}{\longleftrightarrow}} AB_5H_x + xOH^-. \tag{5.3-2}$$

During charging, divalent Ni^{II} is oxidized into the trivalent Ni^{III} state and water is reduced to hydrogen atoms at the metalhydride (MH) electrode, which are, subsequently, absorbed by the hydride-forming compound. The reverse reactions take place during dischar. The net effect of this reaction sequence is that hydroxyl ions in the electrolyte are transported from one electrode to the other, and hence that no electrolyte consumption takes place during current flow. For a proper functioning of a battery, it is thus essential that both electrical and ionic conductivity take place. The basic reactions are also indica in Figure 5.3-2.

In general, exponential relationships between the partial ano/ca currents and the applied electrode potential are obser under kinetic-controlled conditions, as is depicted schematically in Figure 5.3-3 (dashed curves)

[4–8]. The potential scale is given with respect to an Hg/HgO (6 N KOH) reference electrode. The equilibrium potential of the Ni-electrode under standard conditions is far more positive ($E^o_{Ni} = +439\,\text{mV}$) than that of the MH-electrode, which is found to be dependent on the plateau pressure of the hydride-forming material used (E^o_{MH} ranges between -930 and $-860\,\text{mV}$). This implies that the theoretical open-circuit potential of an NiMH battery is approximately 1.3 V, very similar to that of an NiCd battery. This makes these two different battery systems indeed very compatible, although it should already be mentioned here that small differences in performance exist between both systems.

During galvanostatic charging of the NiMH battery with a constant current, an overpotential (η) will be established at both electrodes. The magnitude of each overpotential component (η_{Ni} and η_{MH} in Figure 5.3-3) is determined by the kinetics of the charge transfer reactions. An electrochemical measure for the kinetics of a charge transfer reaction is generally considered to be the exchange current I^o, which is defined at the equilibrium

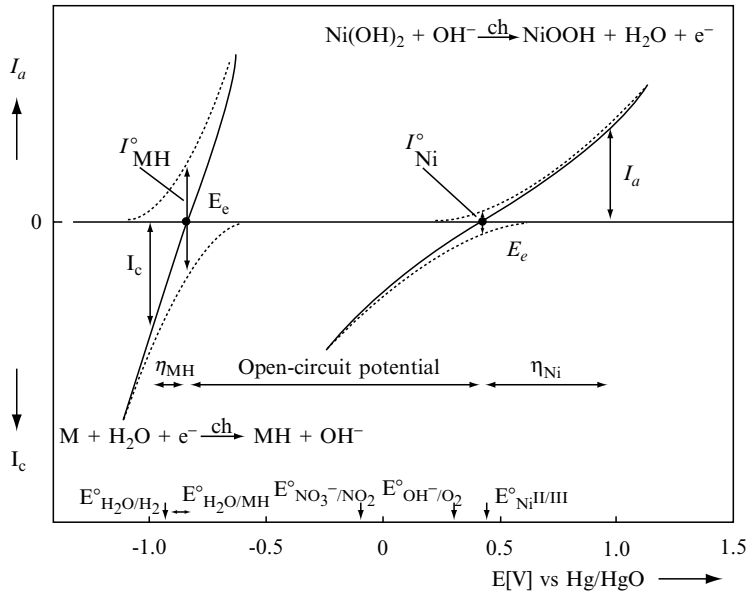


Figure 5.3-3. Schematic representation of the current-potential curves for an Ni and MH electrode (solid lines), assuming kinetically controlled charge transfer reactions. The partial anodic and cathodic reactions are indicated as dashed lines. The exchange currents (I^o) are defined at the equilibrium potentials (E_e). Potentials are given with respect to a Hg/HgO reference electrode. Besides the redox potentials (E^o) of the main electrode reactions, those of some side-reactions are also indicated.

potential, E_e, where the partial anodic current equals the partial cathodic current (see Figure 5.3-3).

In case of the Ni-electrode I^o is reported to be relatively low ($I^o_{Ni} = 10^{-7}\,A/cm^2$), which implies that at a given constant anodic current, I_a, the overpotential at the Ni-electrode is relatively high (see Figure 5.3-3). In contrast, the kinetics of the MH-electrode is known to be strongly dependent on the materials composition. Assuming a highly electrocatalytic hydride-forming compound, this implies that the current-potential curves, characteristic for the MH-electrode are very steep in comparison to those for the Ni-electrode, resulting in a much smaller value for η_{MH} at the same cathodic current I_c, as is schematically shown in Figure 5.3-3. It is evident that the battery voltage under current flow is a summation of the open-circuit potential and the various overpotential components. This includes the ohmic potential drop (η_{IR}) caused by the electrical resistance of the electrolyte (R_e). The reverse processes occur during discharging, resulting in a cell voltage lower than 1.3 V. Clearly, since the potential of both electrodes may change considerably, the absolute values of these potentials cannot be directly deduced from the cell voltage. The use of a reference electrode is therefore inevitable in order to interpret the current-potential dependencies in an appropriate way.

It can be concluded that the kinetics of the charge transfer reactions can generally be described by exponential relationships, denoted as the Butler–Volmer equations. These nonlinear relationships indicate that the so-called charge transfer resistances do not have constant values but are dependent on the applied current. The Butler–Volmer relationships can simply be characterized by two parameters (i.e., the equilibrium potential E_e and the exchange current I^o). It should, however, be noted that these parameters do not have fixed values but are dependent on the concentration of the electroactive species involved in the charge transfer reactions, and thus change as a function of state-of-charge [2, 8]. This is schematically illustrated for the current-potential characteristics for the MH electrode in Figure 5.3-4.

The anodic oxidation of stored hydrogen atoms is here shown to be dependent on the hydrogen concentration at the electrode surface, whereas the reduction rate of water is independent (no water is consumed) on the state-of-charge. As a result, the complete current-potential curves (bold lines), and hence the indicated values for E_e and I^o, change significantly as a function of state-of-charge (SoC). This holds, of course, not only for the metalhydride electrode, but also for other electrodes. This makes an appropriate mathematical description of an entire battery much more complex [4–8].

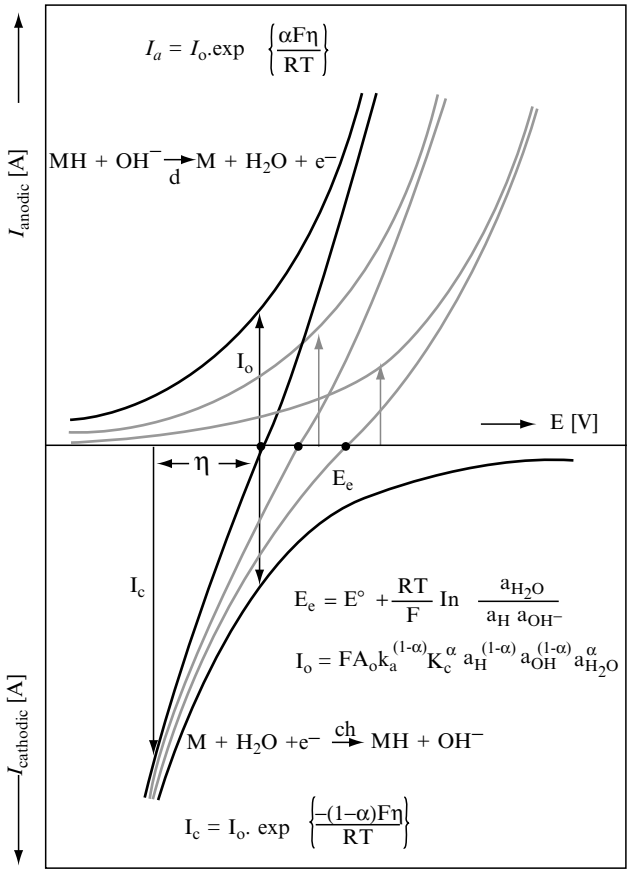


Figure 5.3-4. Schematic representation of the dependence of the partial anodic current-potential curves on the hydrogen concentration within the solid of a metalhydride electrode (i.e., at different states-of-charge). The hydride formation is state-of-charge independent. As a consequence, the overall Butler–Volmer relationships (solid lines) reveal that both the equilibrium potential (E_e) and the exchange current (I^o) change as a function of state-of-charge.

2.2. Side Reactions

2.2.1. Overcharging

To ensure the well-functioning of sealed rechargeable NiMH batteries under a wide variety of conditions, the battery is designed in such a way that the Ni electrode is the capacity-determining electrode, as is schematically indicated in Figure 5.3-2. Such a configuration forces side-reactions

to occur at the Ni electrode, both during overcharging and overdischarging, as is shown later.

During overcharging OH^- ions are oxidized at potentials more positive with respect to the standard redox potential of the OH^-/O_2 redox couple (about 0.3 V with respect to the Hg/HgO reference potential), and oxygen evolution starts at the Ni electrode, according to:

$$4OH^- \overset{\text{Ni}}{\longleftrightarrow} O_2 + 2H_2O + 4e^-. \tag{5.3-3}$$

Again, an exponential relationship between the current and potential is to be expected, as is shown in curve (a) of Figure 5.3-5 and, a more or less constant oxygen overpotential (η_{o_2}), will be established at the Ni electrode.

As a result, the partial oxygen pressure inside the sealed cell starts to rise. Advantageously, oxygen can be transported to the MH electrode, where it can be reduced at the MH/electrolyte interface in hydroxyl ions at the expense of the hydride-formation reaction (Equation 5.3-2),

$$O_2 + 2H_2O + 4e^- \overset{\text{MH}}{\longleftrightarrow} 4OH^-. \tag{5.3-4}$$

When this reduction reaction is kinetically controlled, an exponential dependence is also to be expected for this reduction reaction, as is shown in curve (b) of Figure 5.3-5.

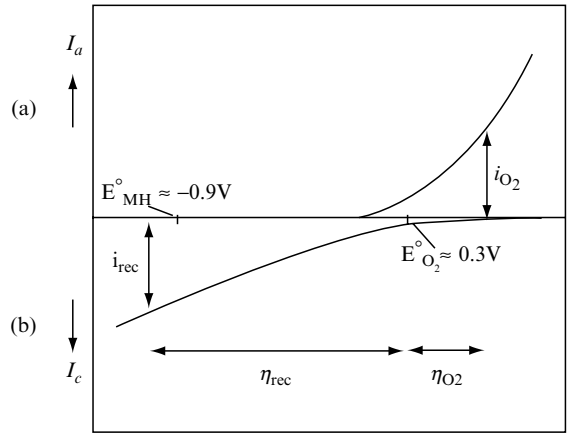


Figure 5.3-5. Schematic representation of the overcharging process inside an NiMH battery. The anodic oxygen evolution reaction (curve (a)) takes place at the Ni electrode, whereas the cathodic oxygen recombination reaction (curve (b)) occurs at the MH electrode.

It should, however, be noted that the steepness of this curve not necessarily need to be the same for both the oxygen evolution (curve (a)) and recombination reaction (curve (b)). It is even very unlikely that the kinetics of both reactions are similar as the oxidation and the reduction reactions take place at chemically different electrode surfaces, resulting in different values for the exchange current densities. All together, the above mechanism ensures that the partial oxygen pressure inside the battery can be kept very low, assuming that the recombination mechanism is functioning properly.

The parasitic oxygen evolution reaction takes place at more positive potentials than the basic Ni reaction (Eq. 5.3-1). This generally results in a rather sharp increase in the battery voltage at the end of the charging process at the point, where the overcharging process takes over. This is indeed confirmed experimentally as Figure 5.3-6 reveals.

This voltage rise is often exploited to detect the end of the charging process, although it should be noted that in general, the battery is not fully charged using this end-of-charge detection point. It is also clear from Figure 5.3-6 that the pressure inside the battery is sharply raising at the end of the charging process, around a 100% state-of-charge level, and tends to level off at higher states-of-charge. In the steady state during overcharging, the amount of oxygen evolved at the Ni electrode (represented by I_{O_2} in curve (a) of Figure 5.3-5) is equal to the amount of oxygen recombining at the MH electrode (I_{rec} in Figure 5.3-5), resulting in a

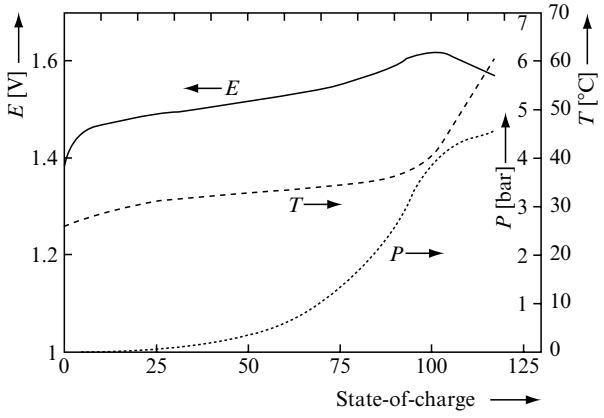


Figure 5.3-6. Development of the cell voltage (E), the internal gas pressure (P), and the cell temperature (T) as a function of state-of-charge for an NiMH battery during charging and overcharging with a high (3 A) current. The battery is then fully charged within 20 min.

constant gas pressure. Furthermore, this implies that all electrical energy supplied to the battery during overcharging is completely converted into heat. From Figure 5.3-5, it is clear that the complete battery voltage is used under these circumstances to build up the two overpotential contributions. The kinetics of this so-called recombination mechanism has been extensively studied [9].

Besides the gas pressure build-up inside a battery, the development of the battery temperature is also of considerable importance and may influence various factors in a negative sense (e.g., cycle-life). The formation of heat (W) inside a battery has been represented by

$$W = I_i \left\{ \sum_i \frac{-T\Delta S_i}{nF} + \sum_i \eta_i + I_i R_e \right\}, \tag{5}$$

where I_i is the local current flowing through the various reaction paths (i) through the battery, T the temperature, n the number of electrons involved in the overall charge transfer reaction (summation of Equations (5.3-1) and (5.3-2)), and F is the Faraday constant [5, 8]. The factors, which contribute to the evolved heat during current flow, can be easily recognized in Equation (5.3-5):

(1) The entropy change (ΔS_i) brought about by the electrochemical reactions;
(2) *The summation term is composed of the various overpotential components (η_i) and has to include the various electrochemical reactions; and*
(3) The internal battery resistance, whose contribution may be significant, especially when high currents are applied, as the heat evolution due to this effect is proportional to the square of the current.

As long as the basic electrochemical reactions (Equations (5.3-1) and (5.3-2)) proceed inside the battery, both overpotential components are relatively small. This implies that the heat contribution, resulting from the electrode reactions is limited. The temperature rise during the normal charging procedure is therefore limited, as Figure 5.3-6 reveals. However, this situation changes drastically as soon as the oxygen recombination cycle at the MH electrode starts. Since the MH electrode potential is at least 1 V more negative with respect to the standard redox potential of the OH^-/O_2 couple (see Figure 5.3-5), this implies that the established overpotential for the oxygen recombination reaction is extremely high ($>$1.2 V). Considering Equation (5.3-5), it is therefore to be expected that the heat evolved inside a battery will sharply increase as soon as the

oxygen recombination cycle starts. This is indeed in agreement with the pronounced temperature increase found during overcharging in the experiments (see Figure 5.3-6). Although the recombinati cycle moderates a considerable pressure rise inside the NiMH battery, it is essential to avoid prolonged overcharging in order to prevent a considerable temperature rise, which may negatively affect other electrode properties. In conclusion it can be said that, dependent on the kinetics of the oxygen recombination reaction (i.e., dependent on the competition between reaction in Equations (5.3-2) and (5.3-4)), the gas pressure and/or temperature of the battery will rise during overcharging. In fact the gas pressure and the temperature rise counterbalance one another when the recombination rate is very poor; the large pressure rise will be combined with a small temperature increase. On the other hand when the recombination rate is excellent, the internal pressure will be limited, while the temperature increase will be rather pronounced. Under extreme charging conditions, both effects do have a negative influence on the battery performance. It should, however, be emphasized that an exorbitant pressure rise may be fatal for the battery performance once the safety vent, with which rechargeable batteries are always equipped, has been opened. The reason for this is that during venting not only the surplus of gases, but simultaneously a significant amount of electrolyte is also released from the battery. This has a negative influence on, for example, the battery cycle-life and recombination kinetics.

2.2.2. Overdischarging

Protection against overdischarging is another factor of importance, especially when NiMH batteries, which inevitably reveal small differences in storage capacities, are used in series. This implies that some batteries are already completely discharged, while others still contain small amounts of electrical energy. Continuation of the discharge process induces over discharging to occur with the already fully discharged batteries. Under these circumstances, water is forced to reduce at the Ni electrode (see Figure 5.3-2), according to:

$$2\ H_2O + 2e^- \overset{\text{Ni}}{\longleftrightarrow} 2OH^- + H_2, \tag{5.3-6}$$

which also results in a pressure build-up inside the battery when no precautions are taken. This decomposition reaction takes place at rather negative potentials at the Ni electrode (i.e., more than 1.3 V more negative with respect to the Ni^{II}/Ni^{III} redox potential), as is indicated in curve (a) of Figure 5.3-7.

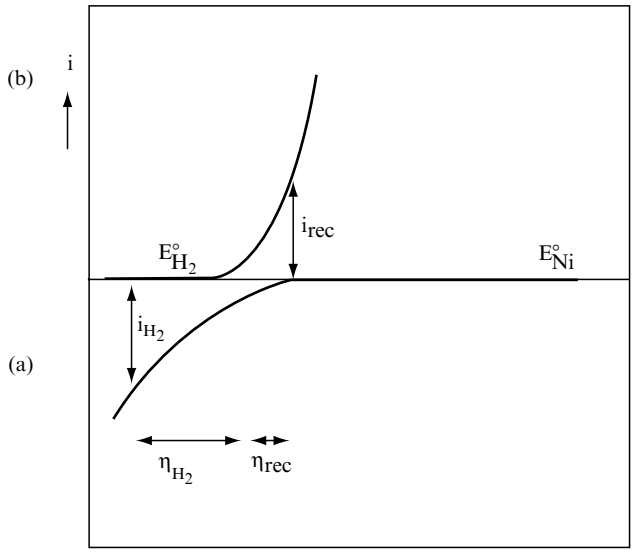


Figure 5.3-7. Schematic representation of the overdischarging process inside an NiMH battery. The hydrogen evolution reaction occurs at rather negative potentials at the Ni electrode (curve (a)), and oxidation of hydrogen occurs at the MH electrode (curve (b)) in the same potential region, resulting in a battery voltage close to 0 V.

As the (electro)chemical affinity of the MH electrode towards hydrogen gas is excellent, it is evident that this gas can be again converted into water at the MH electrode during overdischarging, according to:

$$H_2 + 2OH^- \leftrightarrow 2H_2O + 2e^-. \tag{5.3-7}$$

Whether conversion of molecular hydrogen occurs directly at the MH electrode or atomic hydrogen is oxidized indirectly after chemical adsorption and/or absorption has taken place, is not clear. It is, however, obvious that in both cases, high demands are put on the physical properties of the electrode/electrolyte interface [10]. The electrochemical oxidation also occurs close to the OH^-/H_2O redox potential (curve (b) of Figure 5.3-7). This means that the cell voltage of NiMH batteries is expected to be close to 0 V under these overdischarging conditions, or even invert to some minor extent when the overpotential contributions of both reactions are taken into account. The experimental result of such a process is shown in Figure 5.3-8, and is in agreement with these expectations. During normal discharging the battery voltage is located around 1.2 V and drops indeed towards an inverted voltage of -0.2 V when the overdischarge reactions

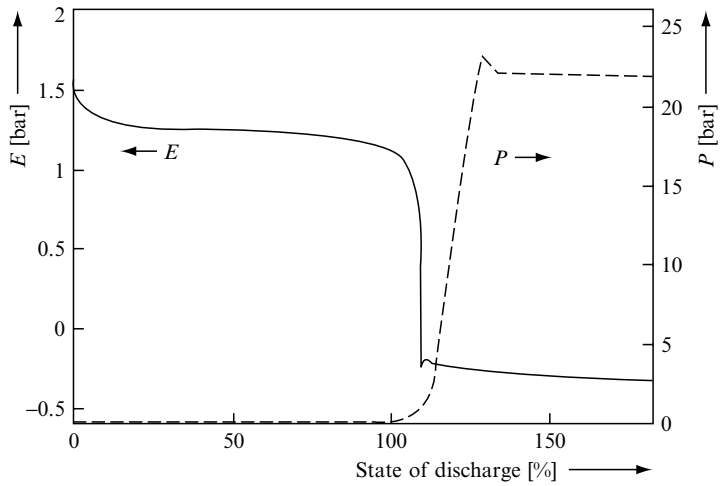


Figure 5.3-8. Experimental result of the development of the cell voltage (E) and the gas pressure (P) as a function of state-of-discharge for an NiMH battery during discharging and overdischarging.

take over. Figure 5.3-8 also reveals that the pressure rises due to hydrogen evolution may be considerable. In the example given, the pressure was quickly built up to the critical level of approximately 20 bars.

At that level the safety vent was forced to open, which can be recognized on the small pressure decrease. After a short period of time the vent closes again. Since the overpotentials of both the hydrogen evolution and hydrogen recombination reaction are relatively low during overdischarging (see Figure 5.3-7), their contribution to the heat evolution will be rather limited. This strongly contrasts to the overcharging situation described above.

In conclusion we can say that a hydrogen recombination cycle controls the pressure rise inside NiMH batteries under overdischarging conditions. As hydrogen evolution and hydrogen oxidation at the separate electrodes take place in the same potential region (see Figure 5.3-7), it is obvious that the battery voltage is very close to 0 V under these conditions. This strongly contrasts to NiCd batteries, for which a large potential reversal is generally observed during prolonged overdischarging as is shown in the subsequent section.

2.2.3. Self-discharge

It is well-known that charged NiMH batteries, similar to all rechargeable batteries, lose their stored charge under open-circuit conditions to a certain extent. The self-discharge rates are strongly dependent on external

conditions, such as, for example, the temperature of the batteries. Typical self-discharge rates at room temperature are of the order of 1% of the nominal storage capacity per day. An example of the variation in self-discharge rate for an NiMH battery as a function of temperature is shown in Figure 5.3-9.

Various mechanisms contribute to the overall self-discharge rate. These mechanisms are all electrochemical in nature. The mechanism operative in NiMH batteries occurs mainly via the gas phase, and can be divided into processes initiated by the Ni electrode or by the MH electrode. The most important mechanism contributing to the overall self-discharge rate will be treated below:

(i) Considering the redox potentials of the Ni electrode (+439 mV) and that of the competing oxygen evolution reaction (+300 mV, see Figure 5.3-3), it is obvious that trivalent Ni^{III} is thermodynamically unstable in an aqueous environment. As a consequence, NiOOH will be reduced by hydroxyl ions at the open-circuit potential, according to:

$$NiOOH + H_2O + e^- \rightarrow Ni(OH)_2 + OH^-, \tag{5.3-8}$$

$$4\ OH^- \xrightarrow{Ni} O_2 + 2H_2O + 4e^-. \tag{5.3-9}$$

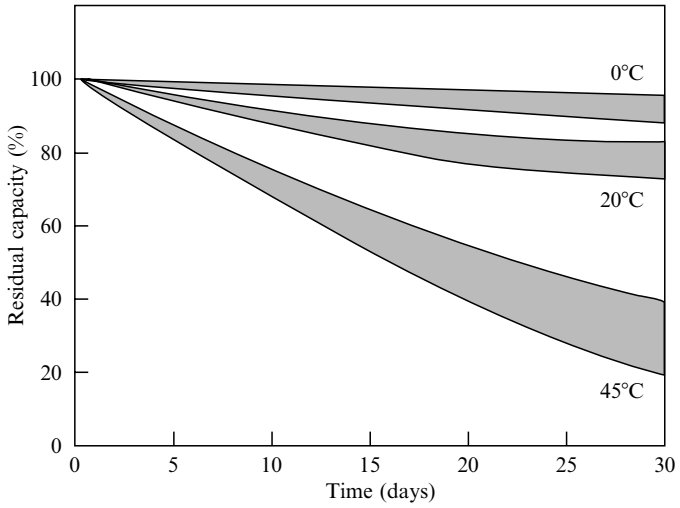


Figure 5.3-9. Dependence of the self-discharge rate on the temperature for an NiMH battery.

These reactions, occurring at the Ni electrode at a rate given by the exchange current, are represented by curves (a) and (b) in Figure 5.3-10, respectively.

The electrons released by the OH^- ions are transferred to the Ni electrode at the electrode/electrolyte interface. Although the Ni^{III} species are thus in principle unstable, electrical charge can, however, be stored in the form of chemical energy in the Ni electrode. This is due to the fact that the kinetics of the oxygen evolution reaction are, fortunately, relatively poor, so that it takes quite a while before capacity loss due to battery self-discharge becomes appreciable. Subsequently, the produced oxygen gas can be transported to the MH electrode, where it can be converted again into OH^- ions at the expense of charge stored in the MH electrode, that is,

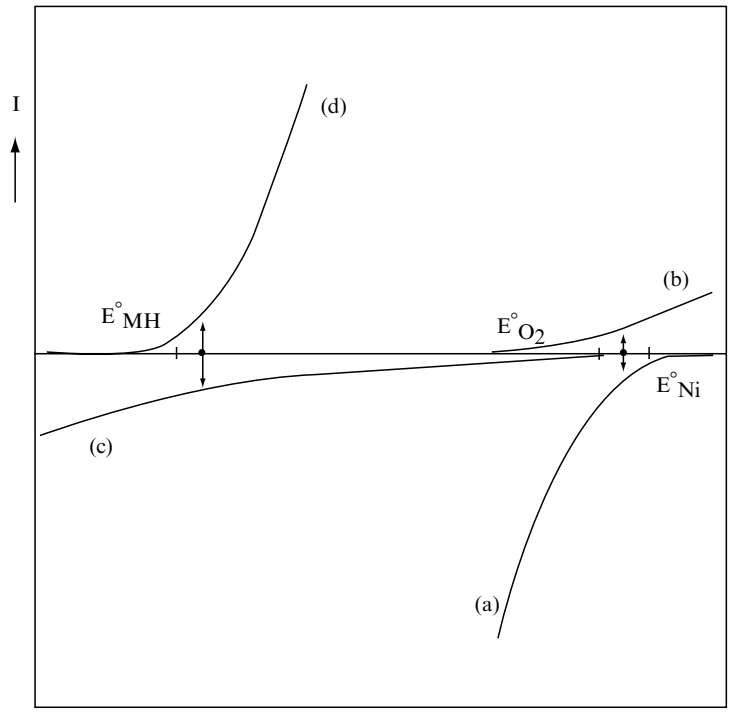


Figure 5.3-10. Schematic representation of the "oxygen gas phase shunt," partly responsible for the self-discharge behavior under open-circuit conditions of aqueous rechargeable batteries, like NiMH. Oxygen evolution (curve (a)) is initiated at the Ni electrode, which is simultaneously discharged (curve (b)) at the open-circuit potential. Consequently, O_2 can be reduced at the MH electrode (curve (c)) at the expense of electrochemically stored hydrogen (curve (d)).

$$O_2 + 2H_2O + 4e^- \xrightarrow{MH} 4OH^-, \tag{5.3-10}$$

$$MH + OH^- \rightarrow M + H_2O + e^-. \tag{5.3-11}$$

These reactions, also occurring at the open-circuit potential at the MH electrode, are represented by curves (c) and (d) in Figure 5.3-10. In the steady-state, all reaction rates are equal. Furthermore, Figure 5.3-10 indicates that the battery open-circuit voltage is around 1.2 V. The ultimate result is that charge stored in both the Ni and MH electrode is slowly released through a gas-phase shunt, in this case oxygen gas.

(ii) A different type of gas-phase shunt is initiated by the MH electrode and is caused by the presence of hydrogen gas inside the battery. As the storage capacity of the MH electrode is considerably larger than that of the Ni electrode (see battery concept in Figure 5.3-2), and the MH electrode contains a certain amount of precharge in the form of hydride, a minimum partial hydrogen pressure is inevitably established inside the NiMH battery, according to the chemical equilibrium

$$MH \leftrightarrow M + 1/2H_2 \uparrow. \tag{5.3-12}$$

The minimum H_2 pressure is dependent on the condition of the battery, but will often be determined by the hydrogen plateau pressure, which is characteristic for many hydride-forming compounds. As a result, H_2 is in contact with the Ni electrode. Since the standard redox potential of the OH^-/H_2 redox couple is much more negative than that of the Ni^{II}/Ni^{III} couple, hydrogen can be oxidized at the Ni electrode, whereas the Ni-electrode is simultaneously reduced, according to

$$H_2 + 2OH^- \xrightarrow{Ni} 2H_2O + 2e^-, \tag{5.3-13}$$

$$NiOOH + H_2O + e^- \rightarrow Ni(OH)_2 + OH^-. \tag{5.3-14}$$

Since the oxidation of hydrogen gas takes place more than 1.2 V more positive with respect to its standard redox potential and the kinetics of this reaction is very favorable, it is likely that the oxidation reaction at the Ni electrode becomes diffusion-controlled. This implies that the oxidation current, which generally reveals an exponential dependence on the voltage,

will level off to become constant at higher voltage levels. Such diffusion-controlled oxidation process, as represented by Equation (5.3-13) is schematically indicated in curve (a) of Figure 5.3-11.

The Ni reduction reaction (Equation 5.3-14) is represented by curve (b). The overall electrochemical process occurs under open-circuit conditions at the Ni electrode and will be strongly influenced by the partial hydrogen pressure inside the battery. It has indeed been reported that the self-discharge rate at the Ni electrode is proportional to the partial hydrogen pressure [10]. For this reason it is important that the hydrogen pressure inside the battery is kept as low as possible (i.e., to employ hydride-forming compounds), which are characterized by a relatively low hydrogen plateau pressure. Again, according to Equations (5.3-12) and (5.3-14), the chemical energy stored in both the MH and Ni electrode is wasted by a gas-phase shunt and can no longer be employed for useful energy supply.

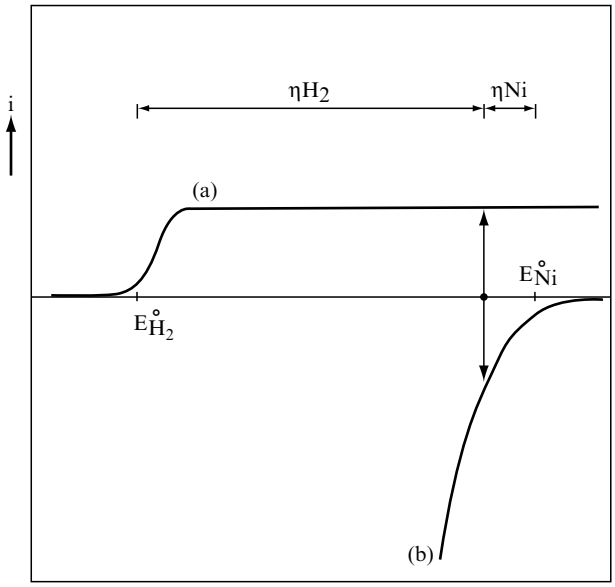


Figure 5.3-11. Schematic representation of the "hydrogen gas phase shunt," occurring during self-discharge inside an NiMH battery under open-circuit conditions. The hydride stored in the MH electrode is inevitably in equilibrium with hydrogen in the gas phase. H_2 gas can be oxidized at the Ni electrode (curve (a)) at positive potentials, where NiO_2H simultaneously is reduced (curve (c)). The oxidation current is shown to be diffusion-controlled, resulting in an anodic current plateau. It has indeed been shown that the self-discharge rate is proportional to the partial hydrogen pressure inside NiMH batteries.

(iii) The third self-discharge mechanism is related to the fabrication process of the Nickel-oxide electrode. These solid-state electrodes are generally prepared by electrolytic reduction of an acidic salt electrolyte, often $Ni(NO_3)_2$ [2]. During this process, NO_3^- ions are reduced to NH_4^+ ions. This results in a significant increase in pH near the electrode/electrolyte interface. The solubility product of $Ni(OH)_2$ will be exceeded and, as a result, $Ni(OH)_2$ will subsequently precipitate on the substrate. A consequence of this process is that, despite the fact that the as-prepared electrodes are thoroughly washed, a certain amount of nitrate ions are inevitably incorporated into the Ni electrodes, which can be leached out during the battery cycle-life. These NO_3^- ions, dissolved in the liquid phase, form the basis of this third self-discharge mechanism. These ionic species can be reduced to lower oxidation states [11]. It is generally assumed that a so-called nitrate/nitrite shuttle is operative in alkaline rechargeable batteries [12]. The standard redox potential of the nitrate/nitrite redox couple [11] is much more positive than that of the MH electrode (−91 mV versus Hg/HgO, see also Figure 5.3-3). This implies that ions delivered by the Ni electrode can be reduced at the MH electrode under open-circuit conditions, according to

$$NO_3^- + H_2O + 2e^- \xrightarrow{\text{MH}} NO_2^- + 2OH^-, \tag{5.3-15}$$

$$\mathrm{MH + OH^- \rightarrow M + H_2O + e^-}. \tag{5.3-16}$$

These reactions are schematically indicated by curves (a) and (b), respectively, in Figure 5.3-12.

The produced nitrite ions can diffuse to the Ni electrode. As the electrode potential of the Ni electrode is more positive than the redox couple of the nitrate/nitrite couple can be converted to nitrate again (curve (c) of Figure 5.3-12), while NiOOH is simultaneously reduced (curve (d)), according to

$$NO_2^- + 2OH^- \xrightarrow{\text{Ni}} NO_3^- + H_2O + 2e^-, \tag{5.3-17}$$

$$\mathrm{NiOOH + H_2O + e^- \rightarrow Ni(OH)_2 + OH^-}. \tag{5.3-18}$$

This reaction sequence can proceed continuously, as the electroactive nitrate and nitrite species are continuously produced at both electrodes. The

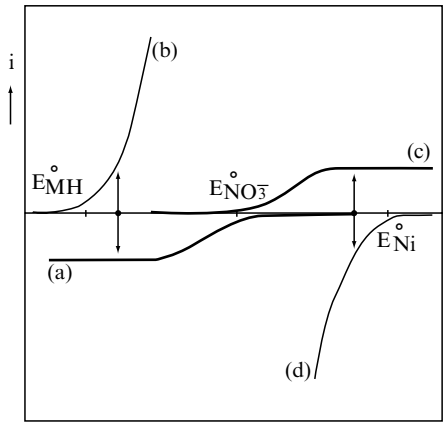


Figure 5.3-12. Schematic representation of the so-called nitrate/nitrite (NO_3^-/NO_2^-) shuttle, which takes place in the electrolyte phase. This "electrolyte shunt," induced by leaching out the Ni electrode starts with the NO_3^- reduction at the open-circuit potential of the MH electrode (curve (a)), at which the stored hydrogen simultaneously is oxidized (curve (b)). The produced NO_2^- ions are transported and converted again to NO_3^- (curve (c)) at the Ni electrode, which itself is reduced (curve (d)).

final result is again that, charge stored in both the MH and Ni electrode is consumed and no longer is available for useful energy supply.

2.2.4. NiMH charging and discharging characteristics

The most common charging method for NiMH batteries is constant current charging. The current has, however, to be limited in order to avoid an excessive rise of temperature and/or internal gas pressure (see Figure 5.3-6). This means that severe overcharging has to be terminated by a reliable end-of-charge detection. The most commonly applied method is based on a voltage drop ($-dV/dt$) induced by the temperature increase during overcharging. Figure 5.3-13 shows the impact of the charging current on the overall battery voltage at ambient temperature. According to kinetic considerations, the battery voltage is indeed expected to increase with increasing currents not only in the "Ni region" at low depth-of-charge (DoC), but also in the "oxygen region" at high DoC.

The transition between these two regions is somewhat dependent on the charging current in that it is shifting towards lower DoC when the current is increased, making the competition between the Ni and O_2 reaction more severe. Furthermore, it should be noted that the $-dV/dt$ effect is much more pronounced at higher currents, resulting from a higher heat production. This is in agreement with the relationship given in Equation (5.3-5).

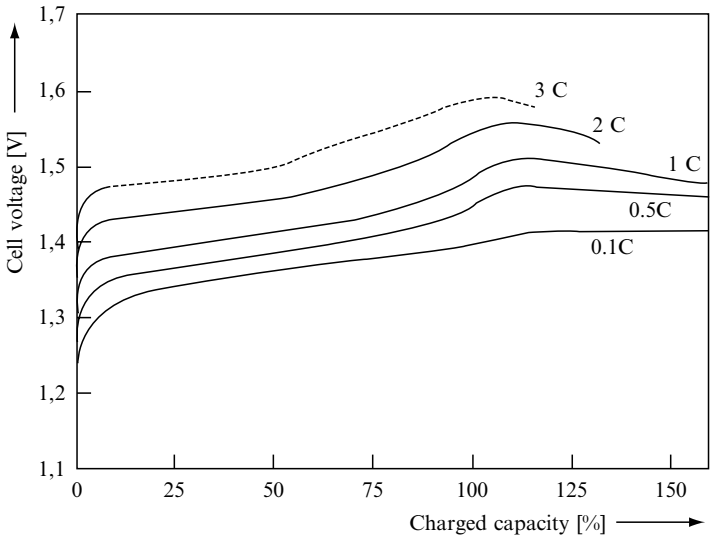


Figure 5.3-13. Cell voltage versus time during charging of NiMH batteries at 25°C at various charging rates.

The capacity of an NiMH battery depends strongly on the conditions, such as the rate of discharge and the ambient temperature. Parameters like cut-off voltage, cycle-life, and general cell condition have a minor effect. The influence of the discharge rate is shown in Figure 5.3-14.

For every battery system, energy is "lost" in two different ways during battery use. In the first place, there is a "virtual los" in discharge capacity, which is fairly limited up to currents of 1 A. The term virtual means that the capacity is not really lost, but is inaccessible at high currents due to the established concentration gradients. Secondly, further energy is lost by the drop of the discharge voltage with increasing currents, which may become very pronounced at high currents (Figure 5.3-14). This energy is really lost and is dissipated as heat (see Equation 5.3-5).

These effects are further accentuated at lower temperatures, as is depicted in Figure 5.3-15. The combination of high discharge rate and low temperature boost the virtual capacity loss. Note the increasing effect of the cut-off voltage level at higher discharge rates, particularly at low temperatures. An arbitrary cut-off voltage level of 1 V is adopted in Figure 5.3-15, and it shows that it makes quite some difference for the discharged capacity whether one discharges at 25 or 0°C, especially at a high drain current of 4 A required for (e.g., power tools).

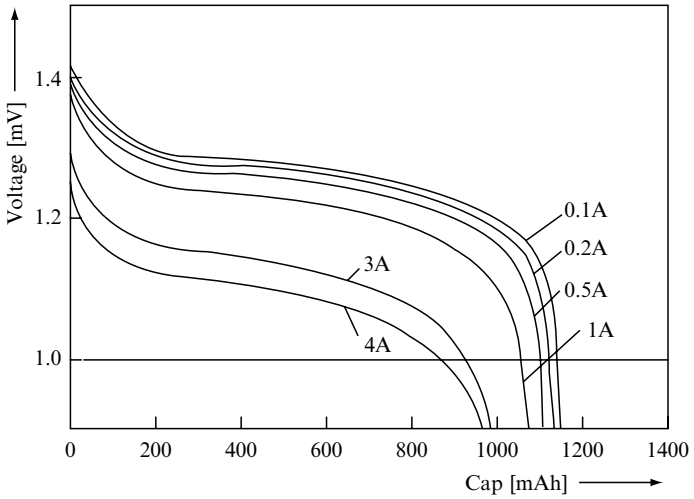


Figure 5.3-14. Influence of the discharge current on the cell voltage and discharge capacity for an AA-size NiMH at 25°C.

2.3. Nickel–Cadmium Batteries

The principles outlined in the previous section for NiMH batteries are, in several aspects, very much alike for NiCd batteries. The concept of an NiCd battery and the basic electrochemical reactions are represented in Figure 5.3-16.

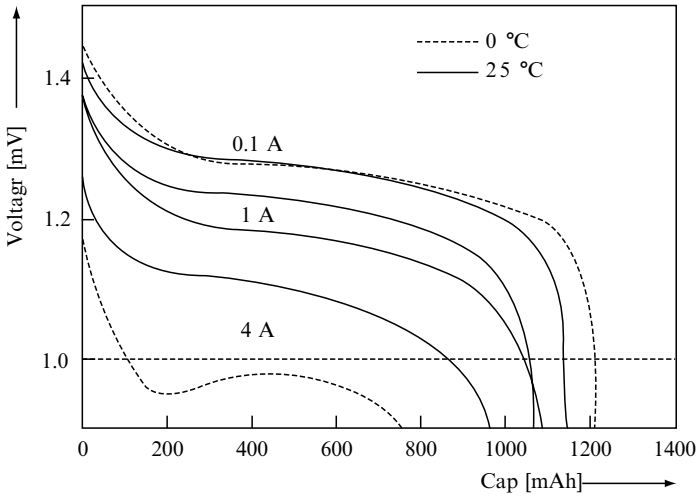


Figure 5.3-15. Influence of the discharge current on the development of the cell voltage for an AA-size NiMH battery at 0 (dashed lines) and 25°C (solid lines).

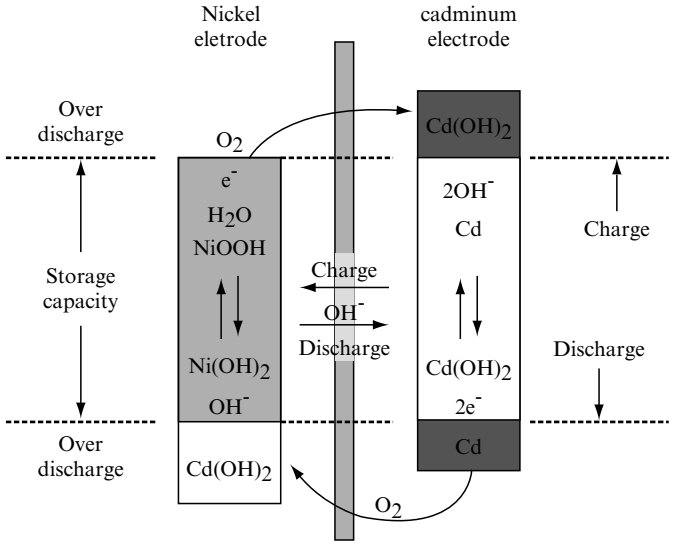


Figure 5.3-16. The concept of a sealed rechargeable NiCd battery.

Evidently, the electrochemical behavior of the nickel electrode is similar to that employed in NiMH batteries, although it should be noted that the performance may differ considerably due to relatively small variations in chemical composition. The electroactive cadmiumhydroxide ($Cd(OH)_2$) species are converted into metallic cadmium (Cd) via a complex series of intermediate chemical dissolution/precipitation reactions during charging. Since the charge transfer reaction is reversible, the opposite reaction occurs during discharging. The white areas of both electrodes represent the nominal storage capacity of the battery. The basic electrochemical charge transfer reactions in its most simplified form are as follows:

$$Ni(OH)_2 + OH^- \underset{d}{\overset{ch}{\longleftrightarrow}} NiOOH + H_2O + e^-, \qquad (5.3-19)$$

$$Cd(OH)_2 + 2e^- \underset{d}{\overset{ch}{\longleftrightarrow}} Cd + 2OH^-. \qquad (5.3-20)$$

In order to ensure the overcharge and overdischarge ability of sealed rechargeable NiCd batteries, these batteries are designed in a very specific way, which is partly based on the same concepts as explained for NiMH batteries. Especially, the gas recombination cycle initiated during over-

charging is exactly the same for NiCd and NiMH batteries. Since both systems are based on aqueous electrolytes, and since the Ni electrode is the capacity-determining electrode in these systems, it is obvious that oxygen evolution is forced to take place at the Ni electrode during overcharging. Simultaneously, reduction of $Cd(OH)_2$ in excess present in the Cd-electrode still continues according to reaction (Equation 5.3-20). As a result, the partial oxygen pressure within the cell starts to rise and induces electrochemical conversion of O_2 at the Cd-electrode. The same reaction sequence takes place, as represented by Equations (5.3-3) and (5.3-4), and was illustrated in Figure 5.3-5. As both reactions also occur at very similar electrode potentials during overcharging (i.e., at relatively high overpotentials), the temperature development is expected to be very similar to that of NiMH batteries. Dependent on the competition between reactions in Equations (5.3-20) and (5.3-4), the gas pressure and/or the temperature of the battery will rise during overcharging. Figure 5.3-17 indeed shows that both the development of the internal gas pressure and temperature increases at the end of the charging process where the overcharging reactions start to proceed.

Differences between these characteristics may be due to differences in reaction kinetics and also to different charging conditions, such as overcharging current. The potential decrease $(-dV/dt)$, for example, is generally more pronounced for NiCd than for NiMH, which makes an end-of-charge detection based on this phenomenon easier for NiCd batteries than for NiMH batteries.

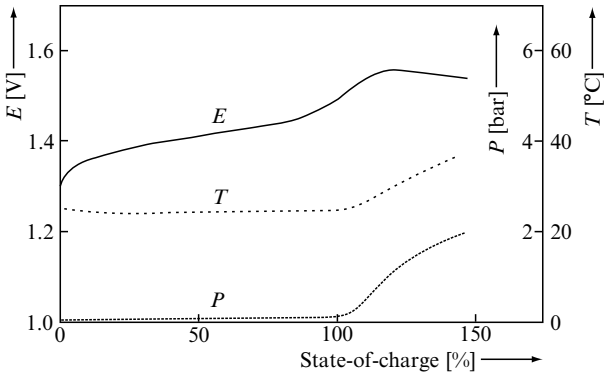


Figure 5.3-17. Development of the cell voltage (E), the internal gas pressure (P), and the battery temperature (T) of an NiCd battery as a function of state-of-charge during charging with a current of 0.6 A (1 C-rate). Clearly, overcharging has quite a different impact on these parameters than the normal charging process.

Protection against overdischarging NiCd batteries is somewhat more complex for NiCd than for NiMH batteries. Since a hydrogen recombination cycle has to be avoided due to the poor electrocatalytic activity of the Cd-electrode with respect to the H_2 oxidation reaction, battery manufacturers adopted another elegant approach. A significant amount of $Cd(OH)_2$, generally denoted as depolarizer, is added to the Ni-electrode and, to a lesser extent, some metallic Cd is added to the Cd-electrode as discharge reserve (see shaded overdischarge areas in Figure 5.3-16). During overdischarge, the $Cd(OH)_2$ present in the Ni-electrode is reduced to metallic Cd (Equation 5.3-20), while the excess of metallic Cd is still being oxidized at the Cd-electrode. As the amount of extra Cd with respect to $Cd(OH)_2$ is limited, oxygen gas will be evolved at the Cd-electrode during continuation of the discharge process. Again an oxygen recombination cycle is established, now starting at the cadmium electrode, as is indicated in Figure 5.3-16.

The theoretically expected development of the electrode potential of the Ni (curve (a)) and the Cd-electrode (curve (b)) during both the discharge and overdischarge process are shown in Figure 5.3-18.

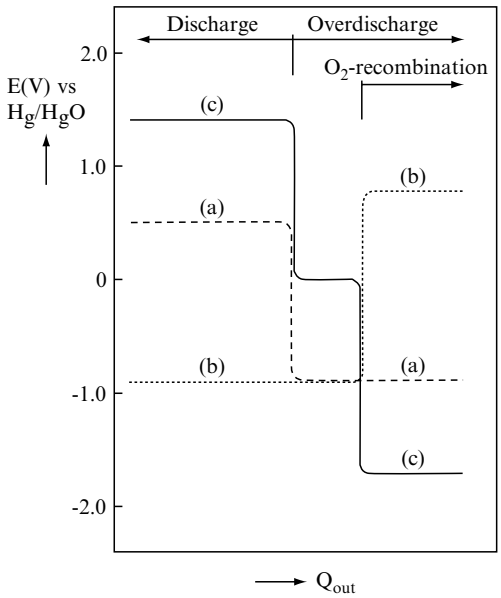


Figure 5.3-18. Schematic representation of the potential dependence of an NiCd battery during discharging and overdischarging. The potentials of the separate Ni and Cd electrodes are represented by the dashed curves (a) and (b), respectively, and are given with respect to an Hg/HgO reference electrode. The development of the total cell voltage is shown by the solid line (curve (c)). The two stages of the overdischarge process can clearly be recognized.

The two different stages, characteristic for the overdischarge process, can clearly be recognized in this figure and have a pronounced influence on the total cell voltage ($E_{Ni} - E_{Cd}$), as is schematically shown in curve (c) of Figure 5.3-18. During the first stage of overdischarging, the same redox reaction takes place at both electrodes and a cell voltage close to 0 V is therefore to be expected. When the oxygen recombination cycle starts (i.e., when O_2 is evolved at the Cd-electrode at positive potentials (curve (b) of Figure 5.3-18)), and converted again into OH^- at the Ni-electrode at very negative potentials (curve (a)), a potential reversal of the battery is indeed to be expected (curve (c)).

An experimental result of such overdischarge process is shown in Figure 5.3-19. The potential dependence is very much the same as for the predicted potential. Once the NiCd battery is fully discharged, electrochemical transition of Cd-species occurs at both electrodes and, consequently, the cell voltage is close to 0 V at the first overdischarge plateau. Evidently, as the overpotential contributions of both electrochemical reactions are very limited, the heat production is relatively small. This results in a very limited temperature increase in this first plateau region (see T-curve in Figure 5.3-19). Continuation of overdischarging leads to cell potential reversal; a second potential plateau, in this case around -1.8 V, is established, at which the oxygen recombination takes place. In agreement with the theoretical considerations given above, the temperature rise is here much more pronounced and is very similar to that under

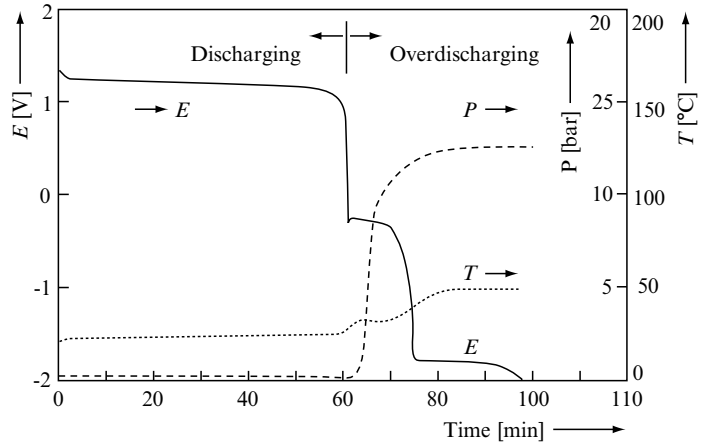


Figure 5.3-19. The characteristic response of the cell voltage (E), internal gas pressure (P), and temperature (T) of an NiCd battery during (over)discharging with a current of 1.2 A (2 C-rate).

overcharge conditions. Although the processes occurring during overcharging and overdischarging are essentially the same, the kinetics of these processes need not necessarily be the same. This becomes clear when one considers the pressure build-up during both processes. The experimental results generally reveal that the pressure during overdischarging rises more rapidly than during overcharging, indicating that the oxygen recombination kinetics at the Cd electrode is more favorable than at the Ni electrode. This sometimes results in a very steep pressure increase, as is shown by the P-curve in Figure 5.3-19.

The well-known memory effect is related to the complex dissolution/precipitation mechanism of the Cd electrode. As has been pointed out, the Cd electrode is composed of a two-phase morphology and the crystallite sizes determine the kinetics of these electrodes. Under normal operation conditions, these crystallite sizes will be small and, consequently, the surface area reasonably high. When the Cd electrode is operated under these normal conditions, the current density will be relatively low for the entire electrode. On the other hand, when only a small part of the storage capacity is frequently used, that part, which is not under continuous current flowing conditions, are allowed to recrystallize. As a result, those effectively ''unused'' electrode parts recrystallize into much larger morphological structures, which results in a much lower surface area. When at a certain moment, this part of the electrode is charged and/or discharged, the current density and consequently the overpotentials will be much larger than for those parts, which were frequently under (de)loading conditions. This results in a virtual capacity loss. This virtual capacity loss can be diminished by charging and discharging the battery over the entire capacity for a few cycles.

Some discharging characteristics for NiCd are shown in Figure 5.3-20. Similar trends as for NiMH (Figure 5.3-14) can be seen: (1) the voltage drops when the current is increased due to the inevitable overpotential losses and (2) a virtual capacity loss is observed as transport limitations are playing a more dominant role at higher currents. As discussed before, this is not a real capacity loss, as all capacity can be discharged when the system is allowed to equilibrate so that the built-up concentration gradients can be minimized again.

2.4. Li–Ion Batteries

So far, this contribution has been restricted to aqueous rechargeable battery systems. The battery voltage of these systems is dominated by the decomposition potentials of water. In this section, we concentrate on a non-aqueous and relatively new system, which has reached its mature

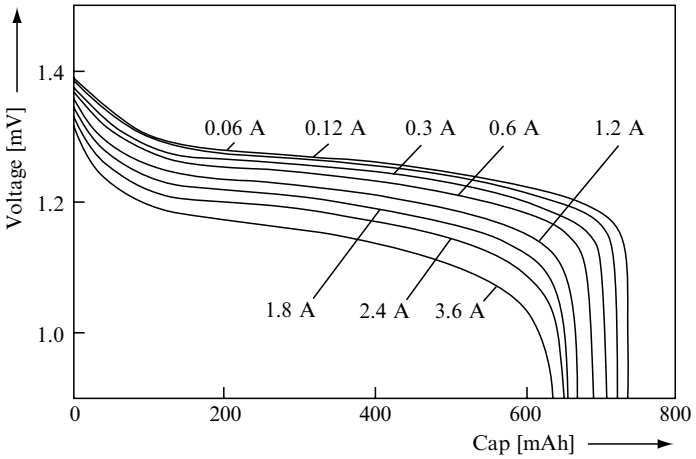


Figure 5.3-20. Development of the cell voltage of an AA-size NiCd battery at 25°C and various discharge currents.

commercialization stage for one decade now: the lithium–ion battery. This battery type differs from aqueous systems in several aspects. In the first place, lithium is a base metal having a very negative value for the standard redox potential (less than −2.5 V versus Hg/HgO; compare with values indicated in Figure 5.3-3). Combining an electrode based on this redox system with a second electrode having a positive redox potential leads to a battery concept with an extremely high cell voltage, of the order of 3.5 up to 4.0 V. In the second place, the molecular weight of some lithium and some lithium-host materials are relatively low, which may result in battery systems with potentially high energy densities, as was already mentioned in relation to Table 5.3-1. In order to make use of lithium as negative electrode material, water and air have, however, to be excluded. Aqueous electrolyte solutions can therefore not be employed, as Li is unstable in this environment. Consequently, all lithium systems are based on organic electrolytes, either in the liquid or the solid form. Conventional Li–ion batteries are nowadays exclusively based on liquid organic electrolytes. All-solid-state Li–ion batteries will become more important in the near future for smaller applications in the field of *Autonomous devices* in the context of *AmI*. But the basic principles apply for both systems.

The concept of a rechargeable Li–ion battery is relatively simple and is shown in Figure 5.3-21.

The positive electrode generally consists of trivalent cobaltoxide species, in which lithium ions are intercalated ($LiCo^{III}O_2$). During charging, trivalent $LiCoO_2$ is oxidized into four-valent $Li_{1-x}Co^{IV}O_2$ and the excess

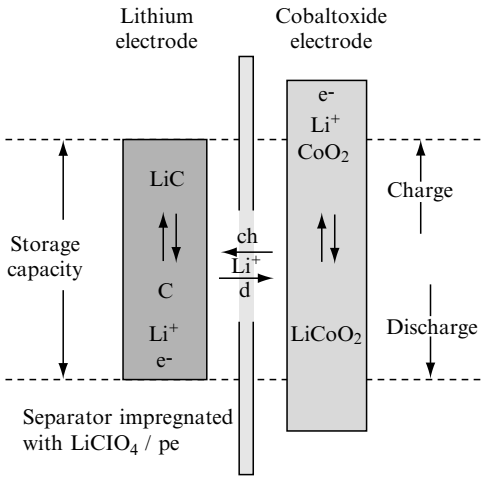


Figure 5.3-21. Concept of a sealed rechargeable Li–ion battery.

of positive charge is liberated from the electrode in the form of Li ions, according to

$$LiCo^{III}O_2 \underset{d}{\overset{ch}{\longleftrightarrow}} Li_{1-x}Co^{IV}O_2 + xLi^+ + xe^-. \qquad (5.3-21)$$

The Li^+ ions dissolve into the electrolyte. For a proper reversible functioning of a Li–ion cell, not all Li^+ ions can be removed from the solid. This implies that x can, in practice, not become lower than 0.45. The electrolyte generally consists of an organic solution, like propylenecarbonate (PC), containing a high concentration of an Li salt (e.g., $LiClO_4$, $LiAsF_6$, or $LiPF_6$) to ensure the electrolytic conductivity between the two electrodes. Arriving at the negative electrode, Li^+ ions can be reduced. This would result in metallic lithium. It was, however, found that the formation of metallic Li unfortunately results in a poor cycle-life. Furthermore, it was recognized that the risk of dendrite formation at the Li electrode surface and, consequently, the risk of short-circuiting resulted in an unsafe design. In order to circumvent these problems, the following electrode construction has been proposed: the Li ions are transported inside a carbon (C_6) electrode and are subsequently reduced according to

$$Li_yC_{6d} + xLi^+ + xe^- \underset{d}{\overset{ch}{\longleftrightarrow}} Li_{y+x}C_6, \qquad (5.3-22)$$

where the value for y, which may range from 0 to 0.7 is dependent on the nature of the used graphite. From the overall reaction;

$$LiCoO_2 + Li_yC_6 \underset{d}{\overset{ch}{\longleftrightarrow}} Li_{1-x}CoO_2 + Li_{y+x}C_6, \qquad (5.3-23)$$

it is clear that the essence is that lithium ions are transported from one electrode through the organic electrolyte to the other electrode. This basic principle is obviously very similar to that of the NiMH battery except that Lithium is involved instead of hydrogen. This transportation concept has often been denoted in speaking terms, like the "rocking chair" model or the "swing" concept to illustrate the swinging movements of the Li-ions. The advantage of this concept is that lithium is safely stored within both electrodes. A disadvantage is, of course, that the energy density has been reduced significantly with respect to potentially possible values. Inspection of Table 5.3-1 reveals that the energy density per volume is hardly higher with respect to that of, for example, the NiMH battery. The advantage can be found in the energy density per weight (Table 5.3-1). It should, however, be noted that this characteristic is often not of major importance in many electronic applications.

One of the remaining problems of Li–ion is that a thorough (electro)-chemical protection against both overcharging and overdischarging is not (yet) available. Recombination cycles, as employed in NiMH and NiCd batteries, do not exist at the moment for Li–ion batteries. However, it should be emphasized that the organic electrolyte can also be decomposed, whereas recombination of the decomposition products into the original organic solvent has not been realized yet. This implies that overcharging and overdischarging has to be avoided under all circumstances, and care should be taken that the cell potential is never outside the potential range in which the basic electrochemical reactions occur. An excellent method to accomplish this is, charging and discharging at a constant potential until the voltage limits are reached. An example of a typical constant-current-constant-voltage (CCCV) charging profile is shown in Figure 5.3-22 reveals the development of the cell voltage (E), the applied current (I), and the stored capacity. In the initial stage, the battery is charged rapidly with a moderate constant current.

The maximum height of the current is, for safety reasons, generally prescribed by the battery manufacturer. Figure 5.3-22 shows that the cell voltage is gradually increased up to a value of 4.2 V. This value is considered to be the upper allowable limit and ensures a proper functioning of the battery. As soon as this limit is reached, the charging regime changes from amperostatic (CC) to potentiostatic (CV-mode). As a result the current is adapted in this region and decreases rapidly to lower values.

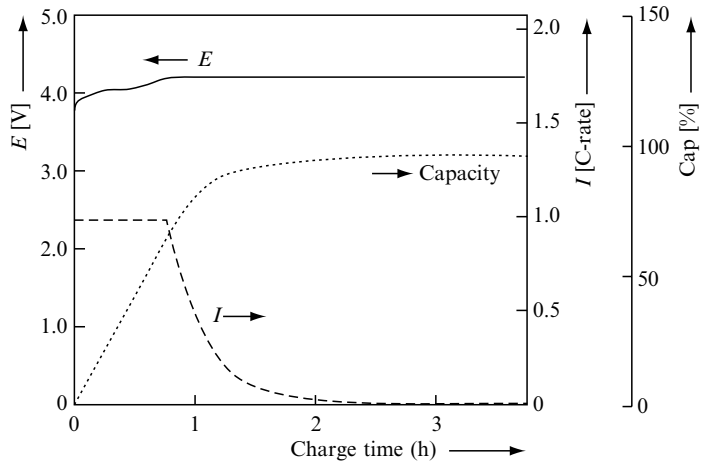


Figure 5.3-22. Constant -current-constant-voltage (CCCV) charging regime typical for Li–ion batteries.

Evidently, charging proceeds more slowly which can be recognized by the inhibited increase of the storage capacity. At the end of the charging process, the current diminishes to very low values when the battery is fully charged. In total charging takes more than 1 hour. Recently, we have proposed the concept of boostcharging in order to cope with the customer request to more quick charge Li–ion without introducing any negative cycle-life effects [13]. In this way, for example, one third of its rated capacity can be charged within 5 min [13].

Some discharge characteristics as a function of the current level are shown in Figure 5.3-23. Again, the same two trends are observed. Firstly, due to larger overpotentials the voltage drops at higher currents, and secondly, the virtual capacity goes down rapidly due to transport limitations. In general, the kinetics of Li–ion is somewhat poorer compared to those of NiCd and NiMH.

The self-discharge rates of commercially available Li–ion batteries are, on the other hand, generally somewhat lower than that of the aqueous systems. Although the self-discharge mechanism is not fully understood, it is likely that the origin must also be sought in the electrochemical instability of the organic electrolyte. The specified self-discharge rates are nowadays of the order of 0.1% per day at room temperature. Following the reasoning of the previous two sections, it is clear that the large temperature increase found for the aqueous systems during overcharging is not observed for Li–ion batteries, because a corresponding recombination cycle does not exist.

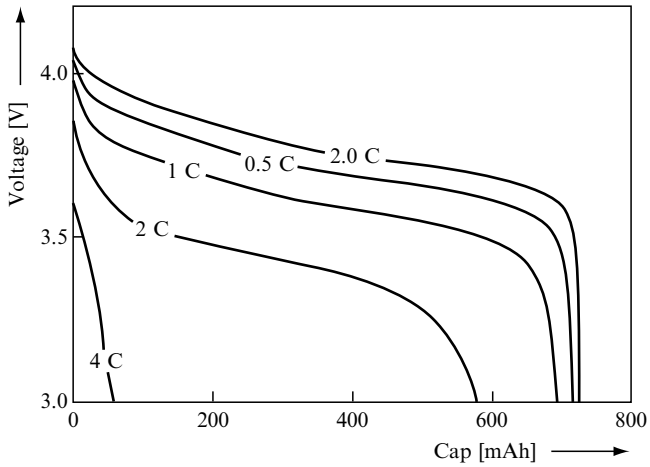


Figure 5.3-23. Development of the discharge voltage of a cylindrical Li–ion battery at 25°C at various discharge rates.

REFERENCES

[1] Linden, D. (ed), 1995, *Handbook of Batteries*, 2nd edition, McGraw-Hill, New York.
[2] Notten, P. H. L., 1994, Rechargeable Nickel-Metalhydride Batteries: A Successful New Concept, Chapter 7, *in* Grandjean, et al. (eds), *NATO ASI Series E*, Volume 281, London.
[3] Notten, P. H. L. and van Beek, J. R. G., 2000, *Chem. Ind.*, **54,** 102–115.
[4] Kruijt, W. S., Notten, P. H. L. and Bergveld, H. J., 1998, *J. Electrochem. Soc.,* **145,** 3764.
[5] Notten, P. H. L., Kruijt, W. S. and Bergveld, H. J., 1998, *J. Electrochem. Soc*., **145,** 3774.
[6] Bergveld, H. J., Notten, P. H. L. and Kruijt, W. S., 1999, *J. Power Sources*, **77,** 143.
[7] Ledovskikh, A., Verbitskiy, E., Ayeb, A. and Notten, P. H. L., 2003, *J. Alloys Comp*., **742,** 356–357.
[8] Bergveld, H. J., Kruijt, W. S. and Notten, P. H. L., 2002, *Battery Management Systems: Design by Modelling*, Kluwer Academic Publishers.
[9] Notten, P. H. L., Verbitskiy, E., Kruijt, W. S. and Bergveld, H. J., *J. Electrochem. Soc*., **152** (2005) A 1423.
[10] Kim, Y. J., Vinsintin, A., Srinivasan, S. and Appleby, A. J., 1992, *J. Electrochem. Soc*., **139,** 351.
[11] Bard, A. J., Parsons, R. and Jordan, J. (eds), 1985, *Standard Redox Potentials in Aqueous Solutions*, Marcel Dekker, New York.
[12] Kohler, U. and Dekker, Ch., 1993, *Dechema Monochraphien,* **128,** 213.
[13] Notten, P. H. L., van Beek, J. R. and Op het Veld, J. H. G., *J. Power Sources*, **145,** 89, 2005.

Section 6

Enabling Technologies and Devices

Chapter 6.1

PERSONAL HEALTHCARE DEVICES

Steffen Leonhardt
Helmholtz Institute of Biomedical Engineering, RWTH Aachen University
medit@hia.rwth-aachen.de

Abstract This chapter gives an overview on state-of-the-art devices for personal healthcare. Due to the expected large impact on health economy and on the biomedical engineering device industry, a discussion of the ongoing demographic changes that will presumably push the future development of personal healthcare devices is included. Finally, some trends in ongoing research and suggestions for open problems are presented.

Keywords definition of personal healthcare; demographic changes; geriatric patients; cardiovascular diseases; motion sensors; ECG monitoring; oxygen saturation

1. INTRODUCTION

What first comes to mind when talking about "biomedical engineering" these days are large technical devices, like magnetic resonance imaging (MRI) tomographs, computer tomographs (CT), anaesthesia workstations or dialysis machines. Such devices were and still are produced mainly for a clinical setting, and only a few of today's biomedical engineering devices aim at being used directly by the patient at home. This scenario will soon change. The following chapter presents an overview and some background information on factors that are expected to influence the future demand for personal healthcare devices and may, therefore, fuel further technological developments.

In Section 2, an attempt to define personal healthcare devices is made. Due to its large impact, the subsequent section exclusively deals with the expected changes in the world population age distribution with a special focus on the situation in Germany. Section 4 presents some state-of-the-art

S. Mukherjee et al. (eds.), AmIware, 349–370.

devices, while Section 5 tries to find areas of future research and identify major trends for personal healthcare. Finally, a summary is given in Section 6.

2. DEFINITIONS

Attempting to define the term "personal healthcare devices," the following statement seems well-suited: personal healthcare devices are "*smart, wearable medical devices supporting diseased people in their domestic environment*." Some of these attributes need further discussions.

2.1. Personal Healthcare Devices Are Smart

Most personal healthcare devices are able to measure vital signals or special physiological parameters of the patient. The sensors employed are especially suitable for this rather rough environment and can cope with humidity, mechanical stress, etc. The crucial question of coupling between the device and the body (e.g. mechanically, optically, resistively, magnetically, etc.) has been solved. Sensor fusion (i.e. extraction of information from several sources) may generate added value. Motion artefacts represent a typical and major challenge for measurement tasks in personal healthcare situations.

Other personal healthcare devices may provide therapeutic measures like electric stimulation (e.g. for induced heart contraction, acoustic sensation, or heart defibrillation) or infuse drugs. One prominent example for such a therapeutic device is the heart pacemaker (Figure 6.1-1).

In addition, the term "smart" implies some local computing power enabling the device to extract information from measured signals, detect artefacts, store valuable fractions of information in memory, diagnose the patient status, and possibly generate alarms automatically. Also, smartness may include the option to individually calibrate the device when necessary and store this information.

2.2. Personal Healthcare Devices Are Wearable

This property has several consequences. First of all, many personal healthcare devices are worn on or in the body. In fact, such devices may be either *implants* (like e.g. heart pacemakers or cochlear implants) or *external devices* integrated into clothing or items of daily use (like watches, glasses, etc.). Secondly, personal healthcare devices tend to be small and light to avoid discomfort while wearing them. Thus, miniaturization and

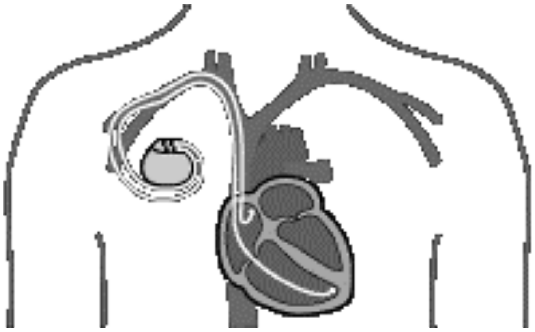


Figure 6.1-1. An implanted heart pacemaker with two electrodes. (Copyright Biotronik GmbH. Modified from [1], reprinted by permission.)

integration are important issues. Thirdly, personal healthcare devices usually have some onboard energy supply system. In some case, a battery may be sufficient while other devices rely either on a smart recharging strategy (e.g. inductive power transfer) or employ an autonomous body power generation and distribution technology.

In a network environment, all body-worn devices and all implants have to be connected. Either a point-to-point or a bus connection may be established by means of electrical wires or optical fibers woven into the garment (*textile integration*, see [2]), or by means of a *wireless body area network* (BAN, see [3, 4]).

However, in a broader and less integrated sense, the definition of personal healthcare devices may be extended to include today's desktop devices that can already be used in one's own domestic environment. Figure 6.1-2 gives an example of a commercial telemonitoring service already in operation allowing domestic measurement of weight, electrocardiogram (ECG), pulse rate, blood pressure and blood sugar concentration, and transmission of that information to a medical consultant team via the Internet.

2.3. Personal Healthcare Devices Support Diseased People

This property addresses the fact that the target group of personal healthcare devices may be handicapped, have developed deficiencies, have minor but chronic diseases, or is expected to get them. One major fraction of that target group is the "elderly patients," a group that will significantly increase in numbers during the next few decades. As these demographic changes will occur throughout the whole world, the next chapter will be completely devoted to a discussion of these changes.

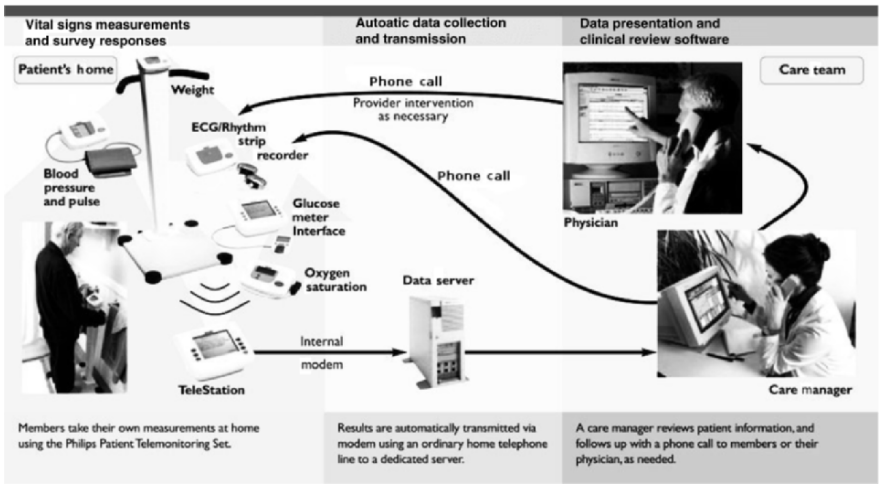


Figure 6.1-2. Commercial Telemonitoring System measuring several physiological parameters and obtaining feedback from a medical consultant team. (Copyright Philips Telemonitoring Solutions (PTS), modified and reprinted by permission.)

Furthermore, it is certainly worthwhile to study the typical handicaps and chronic diseases of elderly persons to identify adequate areas of future research.

However, other patient populations, like middle-aged subjects at a high cardiac risk, people with sleep disorders, diabetics, or other users of personal healthcare devices (like users in sportive scenarios or specific working conditions), should not be excluded.

2.4. Personal Healthcare Devices Are Used Especially in People's Domestic Environment

This property relates to the question how well personal healthcare devices fit into people's domestic environment. In some cases, it may be possible to change the homes and install special equipment for 24/7 monitoring (like motion sensors in the room corners, infrared sensors at the ceiling, microphones in the bedroom to detect breathing sounds, etc.).

Note that the domestic environment is an important area of application, but that this focus should not exclude other mobile applications like sportive ones. It must also be pointed out that although some patients may accept ''total monitoring,'' others may be reminded of Orwell's famous novel ''1984,'' and refuse to accept such a violation of privacy at home. Thus, there certainly are several psychological questions related to this issue, which need to be further investigated.

3. DEMOGRAPHIC CHANGES IN THE WORLD

A seemingly general law of population dynamics is that population growth is often inversely proportional to average income and wealth of a nation. From a prediction of the World Health Organization (WHO) for 2025, we can directly conclude that many of the so-called first and second world countries are coming to age, see Figure 6.1-3.

When focussing on Germany, it turns out that the population pyramid has dramatically changed its shape over the last century (see Figure 6.1-4). In 1901, the population age–group distribution clearly was a ''population pyramid'' (see Figure 6.1-4, left). Such a distribution has many young and only a few old people being supported by a large group of middle-aged people. Keep in mind that the average life expectancy in those days was below 50 years, indicating that many people died before reaching the typical retirement age.

By 1950, the shape of the age–group distribution had dramatically changed (see Figure 6.1-4, 2nd left), mainly due to the natality and losses in specific age-groups (especially the 20–35-year-old world war veterans; men more than women), during the *World War I and World War II* years.

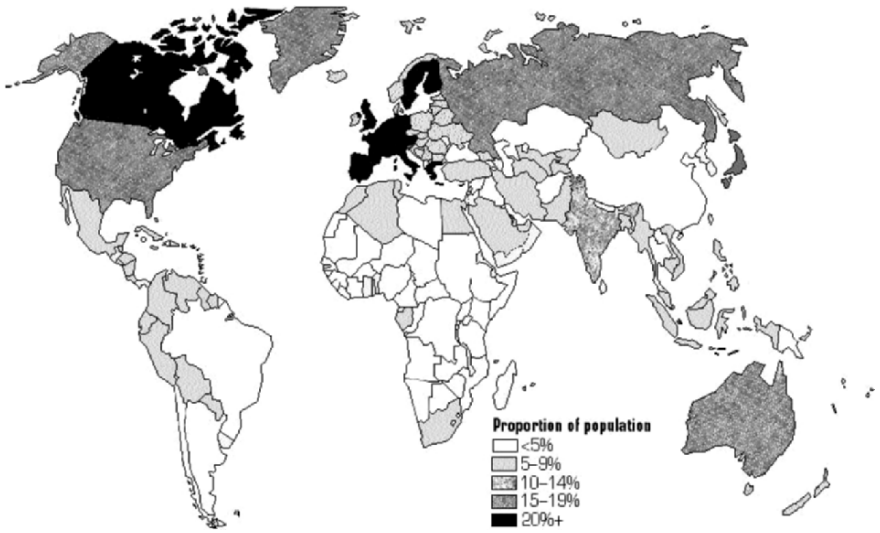


Figure 6.1-3. Prediction of world population aged 65 years and above for the year 2025. (Copyright WHO. Modified from [35], reprinted by permission.)

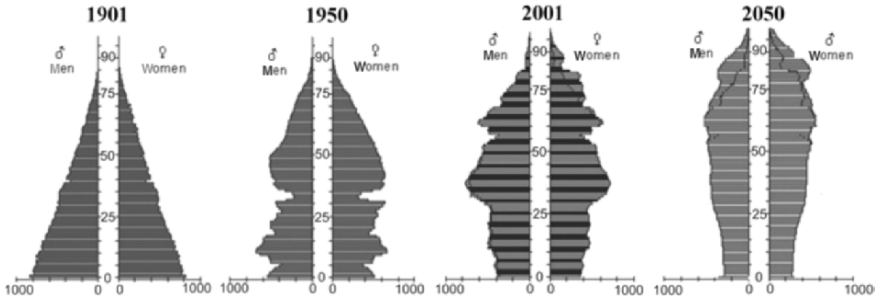


Figure 6.1-4. Population pyramid in Germany over the last century and predicted for 2050. Numbers on x-axis in 1000 persons; age on y-axis in years. (Copyright Statistisches Bundesamt. Modified from [5], reprinted by permission.)

In the 2001 age distribution, two other major phenomena can be identified: (1) the enormous number of the *baby boomer generation* now aged 35–45 and (2) the effect of *contraceptives,* which became available in the early 1960s and allowed effective birth control for an increasing number of women. As a result, the number of people in the age group below 30 significantly dropped (see Figure 6.1-4, 2nd right).

In 2050, the shape of the predicted German age–group sizes shows an almost uniform distribution. Possibly, in the future we may have to name such a distribution a ''population column'' rather than a ''population pyramid.'' In Germany and in many other countries, there seem to be three major factors influencing this demographic change.

3.1. Birth-Rate

(syn. *natality* or *total fertility rate*, TFR) To keep the German population at its current level, a total fertility rate of 2.1 children per woman would be required. However, the current natality in Germany is only 1.38 babies per woman [5].

It is interesting to note that the total fertility rate between East and West Germany has almost equalized by now, but the history of East German and West German reproductive behavior shows some differences over the last few decades reflecting the different political systems and economic incentives (see Figure 6.1-5).

By inspection of Figure 6.1-5, major differences of birthrates are noticeable in the 1980s (when the East German government supported the decision for children with financial incentives), and in the 1990s (when the social changes due to the German unification seem to have frightened potential East German parents).

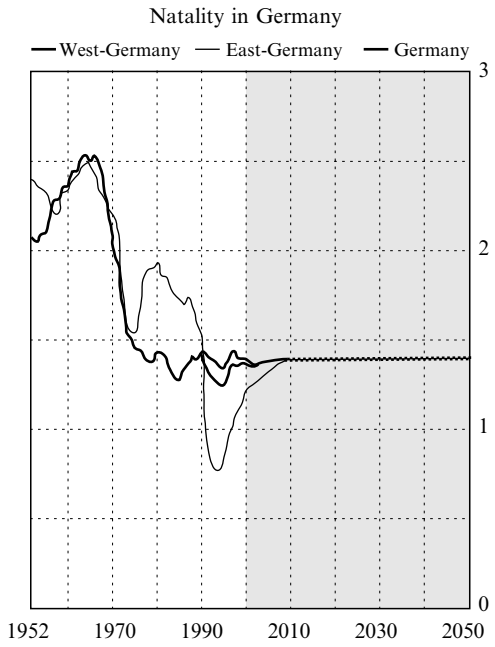


Figure 6.1-5. Birth-rates for West-, East-, and total Germany from 1952 to 2050. Predictions for the years >2002 are based on the 10th coordinated Federal German population forecast. (Copyright Statistisches Bundesamt. Modified from [5], reprinted by permission.)

While the TFR average value of the European Union is only slightly larger (1.48 in 2000) than that in Germany, due to several other low birthrate EU member states (e.g. Greece: 1.29; Italy: 1.24; and Spain: 1.23), some other first world countries deviate significantly from this tendency. For example, Island has a TFR of 2.10, USA 2.06, France 1.88, and Denmark 1.77. But Figure 6.1-3 reveals that even some of the second world countries (e.g. China, Brazil, and Russia, but not India) will grow older. By contrast, in some sub-Saharan African countries, today's women bear an average of 6 children each.

3.2. Life Expectancy

The second factor influencing the demographic situation is life expectancy, which, in Germany, has dramatically increased over the last century, especially in the first half (see Figure 6.1-6).

One major reason that has already been identified was the reduced infant mortality: while in 1901, 200 out of 1000 infants born alive died

Figure 6.1-6. Life expectancy of German infants at birth over the last century. Scenarios 1–3 consider different assumptions regarding future reduction in mortality. (Copyright Statistisches Bundesamt. Modified from [5], reprinted by permission.)

within the first year (20%!), only 4 out of 1000 infants did not survive in 2001. To a large extent, this decline in infant mortality was due to improved health care, a continuous enhancement in hygiene, and access to a better diet. But, of course, the improved living standards (easier working conditions, better housing conditions) throughout life have contributed to life expectancy as well.

3.3. Migration

The third factor influencing demographic changes in Germany is migration. In the 1950s and 1960s, Germany invited foreign workers to move to Germany, which induced a notable population rise. Another notable migration occurred during the early 1990s when many people of German origin from GUS states decided to move to Germany.

Currently, three scenarios on the impact of migration upon the total German population are discussed. These estimates range from a growth of

100.000 to 300.000 (see [5]), which certainly will influence the overall population dynamics and may shift the aging dynamics somewhat, but is not sufficient to keep the population in Germany constant.

4. STATE-OF-THE-ART PERSONAL HEALTHCARE DEVICES

In the later sections, we present a selection of rather typical devices, which are already in the market, or are in the process of becoming commercially available. This collection may give an impression of possible applications and of the variety of designs. Keep in mind, though, that such a selection may be personally biased and is by no means comprehensive. For example, hearing aids or blood glucose monitors certainly are personal healthcare devices, but are omitted due to space restrictions.

4.1. MOTION SENSORS

One example for this category is the Actiwatch device manufactured by Cambridge Neurotechnology Ltd. ([6], see Figure 6.1-7). The basic design features a battery-powered, wrist-worn device weighing about 10 g and containing a miniature uniaxial accelerometer that measures a signal as the wearer produces accelerations above 0.05 g. Applications include general monitoring of activity, insomnia (with additional pressure sensor to be actively pressed between thumb and fingers), mood (with additional light sensor), energy expenditure (with additional temperature sensor), fatigue/alertness and sleep analysis like snoring levels (with additional microphone), or detection of periodic limb movement during sleep (PLMS).

For detection of tremor associated with Parkinson's disease and for monitoring and assessing the effect of drugs on the CNS, a special design named Actiwatch Neurologica is available (focusing on readings in the 3–11 Hz range).

Another example of monitoring energy expenditure with motion sensors is the ''HealthWear Armband'' by BodyMedia, Inc. [8], which senses temperature gradients from skin to environment, mechanical activity and bioimpedance of the skin (see Figure 6.1-8). The targeted application is diet control, where the device and a connected PC act as personal assistants.

Figure 6.1-7. Actiwatch system as a wrist-worn device (left) and applied to the feet for sleep analysis (right). (Modified from [7], reprinted by permission.)

Note that in heart pacemakers, activity sensors based on accelerometers or on piezoelectric elements were already introduced more than a decade ago [10–12], and are commercially employed to control the pacing frequency based on the individual's activity level.

4.2. Heart Rate and ECG Monitoring

The ECG monitoring is a task which is very well-suited for personal healthcare devices. There are many commercial devices already available on the market, and only a few selected examples will be presented here. Such a selection certainly is very subjective and by no means comprehensive.

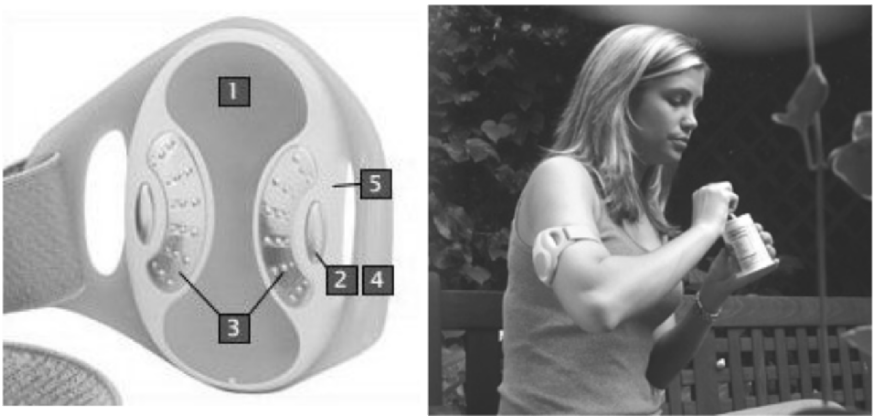


Figure 6.1-8. HealthWear Armband (left) and application to diet control (right). Numbers indicate accelerometer (1), heat flux (2), galvanic skin response (3), skin temperature (4), and near-body temperature (5). (Modified from [9], reprinted by permission.)

One prominent representative of heart rate monitoring devices is the Polar monitoring system [13], originally aiming at sportive applications (see Figure 6.1-9).

The Polar system consists of a thoracic belt featuring dry electrodes to electrically measure heart rate and a wrist-worn watch-like monitor for display of the heart rate. The heart rate is transmitted telemetrically from the belt to the watch.

Note that devices with dry electrodes benefit from sweat, making them especially suitable for sportive applications. On dry skin, however, skin resistance may become a problem.

Somewhat different from the Polar system are ECG multilead measurement systems which include aids to position the electrodes, a big advantage for end users using such a device on their own. One representative of this class of devices was the classical CardioScout system (see Figure 6.1-10), a measurement device consisting of a preshaped multielectrode disposable and a reusable electronic board (transmitter) connected to the electrode disposable.

Another commercial example of a positioning aid for ECG monitoring comes from Tapuz, Inc. [16]. This device has rather large dry electrodes and relies on the elasticity of the silicone apron, which can be pressed against the chest by pulling back both arms (see Figure 6.1-11, right). Through a simple connection to a standard cell phone (see Figure 6.1-11, left), this device can be easily used to transmit a multilead ECG around the world.

As another design example, the Actiheart device should be mentioned briefly. Like the Actiwatch, this use-and-forget device is very small and weighs only 10 g. In order to use it, standard wet ECG electrodes are used for the measurement (see Figure 6.1-12, left).

Figure 6.1-9. POLAR thoracic belt (left) and wrist-worn display (right). (Modified from [14], reprinted by permission.)

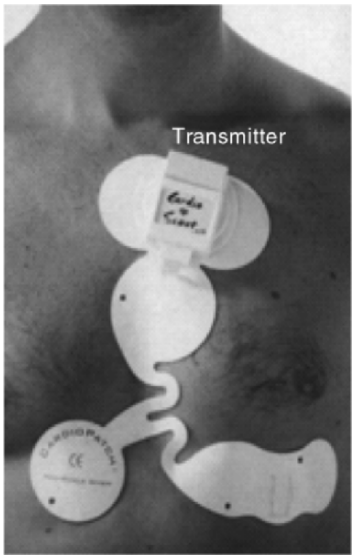


Figure 6.1-10. Old CardioScout system with disposable electrode pad and reusable electronic board including a telemetric system. (Copyright Picomed GmbH from [15], modified and reprinted by permission.)

The primary function of this device is to collect ECG data and store it in memory (data logging, up to 11 days). According to the manufacturer, lifetime of the rechargeable Lithium batteries is claimed to last 14 days. The measured data can be read with an Actiheart reader (Figure 6.1-12, right) and transmitted to a PC for further analysis.

Figure 6.1-11. Multilead electrode apron and measurement box connected to a standard mobile phone (left) and application of apron to chest (right). (Modified from [16], reprinted by permission[1].)

[1] Note: *the device(s) portrayed in this figure are proprietary products of Tapuz Medical Technology (T.M.T 2004) Ltd. All rights of the owner are reserved.*

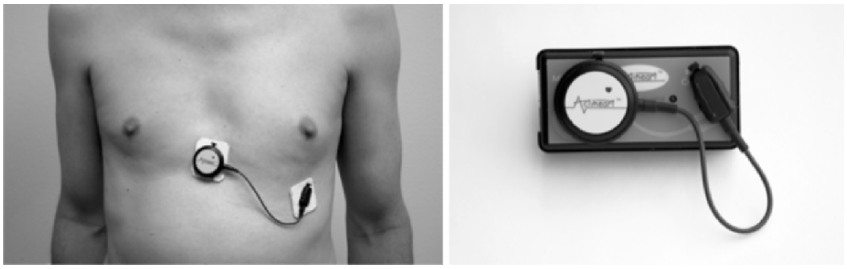


Figure 6.1-12. Actiheart ECG data logger applied to the body (left) and Actiheart Reader (right). (Provided by [7], reprinted by permission.)

Another example for ECG monitoring devices, the CardioOnline belt from Philips Research Aachen, has recently been announced. It features dry electrodes, an integrated accelerometer, and is equipped with a rechargeable Li-Polymer battery. Wireless communication to external monitors is possible via a 5 mW RF front-end link allowing communication at 433 MHz, 868 MHz, 1.88 GHz (DECT), and 2.4 GHz (Bluetooth). Figure 6.1-13 give some impressions of possible textile integration.

4.3. Blood Pressure Monitoring

Blood pressure is a very important physiological parameter and frequently measured in clinical settings, often invasively with direct access to arteries. For private use, noninvasive arm and wrist monitoring devices are available, but, especially wrist devices, are often subject to intrinsic meas-

Figure 6.1-13. Flexible ECG monitoring device appropriate for integration into underwear (left), textile belts or bras (right). (Copyright Philips Research Aachen, modified and reprinted by permission.)

urement errors due to a hydrostatic pressure difference between the heart and location of the monitor. To overcome this foreseeable and frequently occurring measurement mistake, Braun has recently launched a price-winning blood pressure monitor equipped with an active positioning sensor, which helps to find the right arm position during measurements (see Figure 6.1-14).

4.4. Monitoring of Oxygen Saturation

Blood oxygen saturation is a very important vital parameter and holds information on the quality of ventilation, as well as on perfusion and status of the cardiovascular system (e.g. heart rate). In addition, derived quantities, like heart rate variability, give valuable information. In this area, integration and miniaturization are important driving forces. For example, Asada and coworkers [18] have designed several ring devices wearable on a finger, where the device carries not just the saturation sensors, but also energy supply and a RF link at 915 MHz. Figure 6.1-15, right, illustrates this trend.

4.5. Textile Integration

Textile integration is a promising approach to make personal health-care devices wearable. Besides putting flexible electronic boards into shirt pockets (which should be considered a rather low level of integration), a

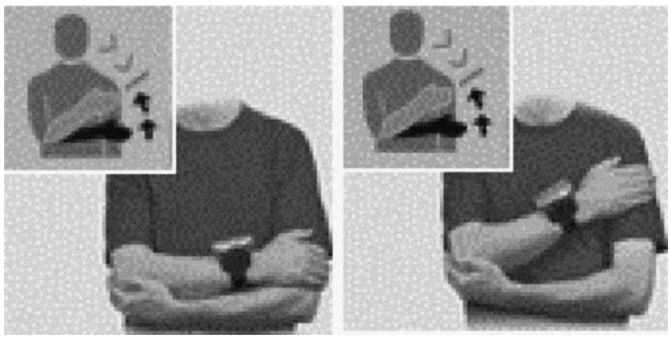


Figure 6.1-14. Proper positioning with the wrist-worn blood pressure monitor *Sensor Control*. (Copyright Braun GmbH from [17], modified and reprinted by permission.)

Figure 6.1-15. Classical handheld pulse oximeter (left, [19]) (Copyright Nellcor Puritan Bennet Inc. USA and the miniaturized Harvard saturation measurement ring right, from [18]. Modified and reprinted by permission.)

few truly integrated designs are becoming available, like the LifeShirt manufactured by Vivometrics Inc. ([20], see Figure 6.1-16).

The sensory principle of the LifeShirt is the so-called *respiratory inductive plethysmography* (RIP), known from the classical Respitrace device and widely accepted as a gold standard for respiratory monitoring. In fact, the electric coils used as sensing elements for respiration are integrated

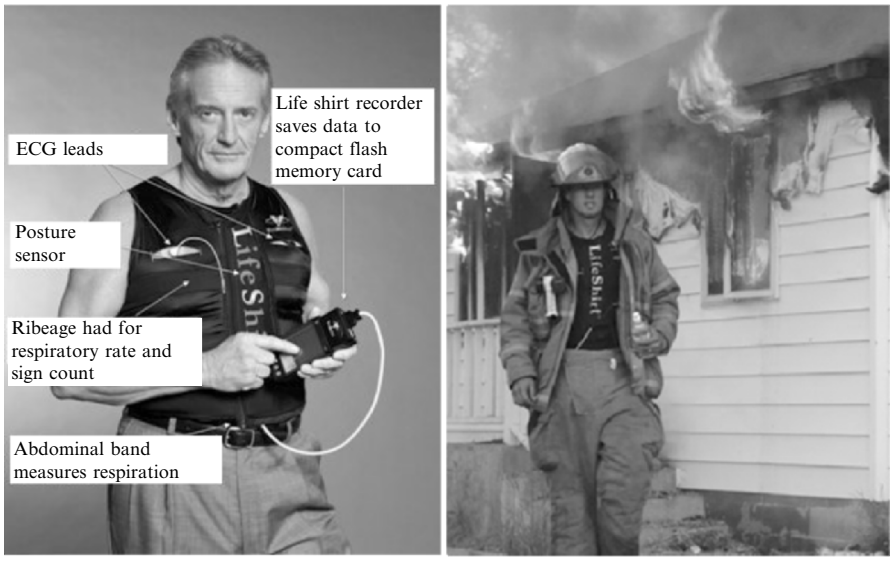


Figure 6.1-16. The concept of the LifeShirt (left) and its application to monitoring of first responders. (Modified from [20], reprinted by permission.)

into the woven fabric. Several peripheral devices (like a wearable ECG monitoring) can be connected to the system.

A competing design, the so-called SmartShirt, has been developed at the Georgia Institute of Technology in the late 1990s [2], and is now manufactured by Sensatex, Inc. [21]. This sometimes called "wearable motherboard" technology features an integrated microcontroller and an optical fiber bus woven into the fabric. Several sensors can be connected via a specifically designed interconnection technology. The targeted areas of applications include first responders, sportive and military use, but there are also first reports on baby clothing using this technology to prevent sudden infant death syndrome (SIDS). Figure 6.1-17 shows some details of the concept.

5. OPEN PROBLEMS AND ONGOING RESEARCH

In this chapter, some open research problems shall be addressed. Afterwards, some of the ongoing research programs are shortly presented.

5.1. Some Open Questions

There are several *physiological parameters* like *noninvasive continuous blood pressure* or *continuous blood glucose concentration* that are crucial for

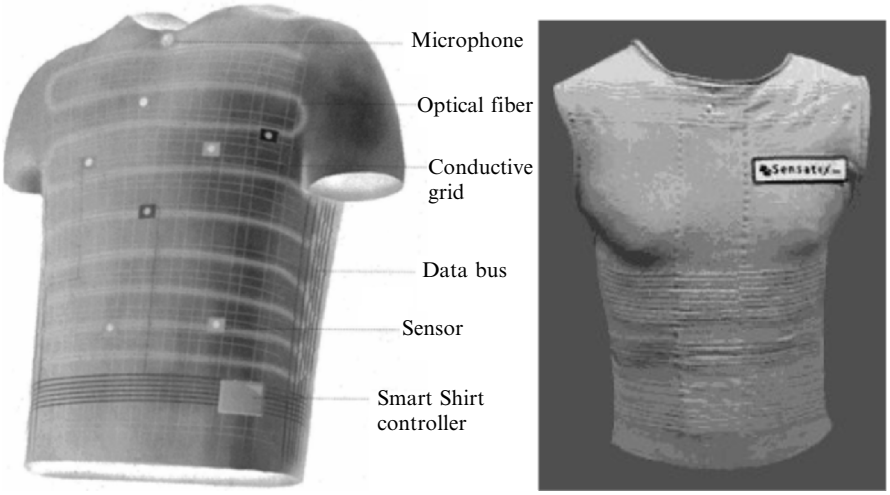


Figure 6.1-17. The concept of the SmartShirt (left) and a prototype. From [21]. (Copyright Sensatex Inc. Modified from [21], reprinted by permission.)

many applications and cannot be obtained yet. For continuous blood pressure monitoring, the pulse transit time (PTT) may be an option [22] as individual calibration is not a problem in an individualized personal healthcare scenario. First reports on accuracy in the ICU are promising [23–25].

Body area networks represent a key technology for personal healthcare to combine different sensors and produce new levels of information by *sensor fusion*. Figure 6.1-18 gives an example of a sportive body area network scenario used for monitoring of ECG, SpO2, tilt and motion.

Like in other distributed networks, *time synchronization* is an important issue, especially when it comes to distributed measurement tasks in the μs range.

Textile Integration is viewed as another key technology for personal healthcare devices [26]. Yet, there are also open questions regarding the application of this technology. One example is distributed measurements of potentials (e.g. from arm to foot as in bioimpedance spectroscopy, see [27]). Since these potentials are small *analogue signals*, one possible option would to integrate small shielded cables into the garment. While digital

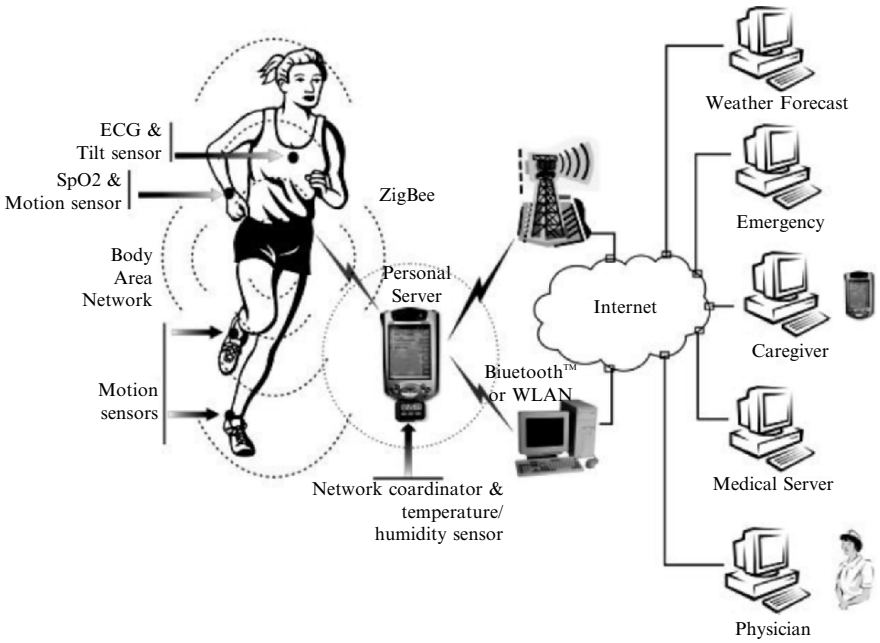


Figure 6.1-18. Example of a wireless body area network including communication links to the Internet. (Modified from [3], reprinted by permission.)

signal transmission through textiles has been successfully demonstrated (see e.g. [28]), analogue signal transmission through coaxial or better triaxial cables might be a problem. Another research question not yet answered is the *connection* between rigid electronic boards and a flexible textile garment. For many applications, electrodes, especially *textile electrodes*, are an important issue and there are first ideas on how to make them. However, replacing this resistive coupling by proper noncontact methods like capacitive, optical or magnetic coupling to the body would be preferable. Finally, hygienic aspects like *washability* need to be addressed.

For almost all wearable devices, *motion artefacts* are a key problem. There is a huge demand for robust sensing technologies and/or artefact-suppressing algorithms.

In addition, the question of an optimal *power supply* is open. Of course, one way to power wearable devices are batteries. However, as is well known from the hearing aid industry, changing these small batteries is a cumbersome procedure, especially for elderly people with reduced manual skills and vision. *Rechargeable batteries* may be a good approach and a smart power management strategy could be to design proper recharging stations (e.g. by integrating a means for inductive power transfer into the bed or the sofa). Another option would be to use *body energy sources* (like motion or heat production) to power personal health care devices, see [29] for reference. Also, using *flexible solar modules* as a power source for personal health care should be further investigated.

A difficult question in this context is *power distribution*, since electrical energy produced during walking by a piezoelectric element integrated into the shoes may be needed somewhere else (e.g. to support a wrist-worn device).

More general aspects of personal healthcare devices that need to be further addressed are *usability* and *ergonomics*. Regarding man–machine interaction, the needs of the elderly patients are certainly different from young computer users. Furthermore, *reliability* and *liability* become important issues as soon as everybody starts to rely on personal healthcare devices more than on other human senses.

5.2. Ongoing Research Programs

Several ongoing projects address issues related to personal healthcare. In the following, a few selected examples funded by the EU or the German Ministry of Research (BMBF) are given.

5.2.1. EU project sensation

SENSATION [30] aims to explore a wide range of micro- and nano-sensor technologies, with the aim to achieve unobtrusive, cost-effective,

and real-time monitoring, detection and prediction of human physiological states in relation to (e.g. wakefulness, fatigue, and stress) anytime, everywhere, and for everybody.

5.2.2. EU project MyHeart

With a budget of €33 million and 33 project partners from 11 countries, MyHeart [31] is another European project aiming at creating options for wearable smart electronic systems and associated services that empower users to take control over their health status. It is an effort of industrial research institutes, academics and hospitals to cover the whole value chain from textile research via fashion and electronic design to medical and home-based applications.

The project focuses on cardiovascular diseases (CVD), the leading cause of death in the western world. It is commonly accepted that a healthy and preventive lifestyle, as well as early diagnosis, can systematically combat the origin of CVD and save millions of life-years. MyHeart explores technologies to support people to adopt a more healthy and active lifestyle in order to reduce risk for developing CVD and limit the recurrence rate of earlier acute events. One of the first results of MyHeart is a prototype of a wearable, wireless monitoring system that measures and diagnoses body signals of the wearer to detect abnormal health conditions.

5.2.3. EU project WEALTHY

Another EU project focusing on textile personal healthcare is WEALTHY [32]. The main objective of WEALTHY is to set up a comfortable health monitoring system. This will be based on a ''wearable'' interface, implemented by integrating smart sensors (in fiber and yarn form), advanced signal processing techniques and modern telecommunication systems on a textile platform, and by developing a monitoring system for data management with local intelligence in the form of a decision support unit.

5.2.4. EU project TOPCARE

TOPCARE stands for ''Telematic Homecare Platform in Cooperatives Health Care Provider Networks.'' The overall objective of this project is to develop technical devices and telecommunication structures, and to lay the organizational groundwork for bringing cooperative health care services into the home of patients. A telematic homecare platform is being established and evaluated in European cooperative healthcare environments for home monitoring and treatment of patients needing:

- Infusion therapies,
- Controlled ventilatory support, and
- Monitored medication adjustment and adherence control when treated with anticoagulants.

Note that while the other EU projects presented mainly dealt with sensor technology and measurement tasks, this project includes therapy guidance and therapeutic devices.

5.2.5. BMBF project PHMon

PHMon stands for "Personal Health Monitoring System with Microsystem Sensor Technology" and is a research project conducted by the University of Karlsruhe together with industrial and scientific partners [34]. It is funded in part by the BMBF microsystem initiative. The aim of the project is to develop sensors and an integrated platform for continuous blood pressure monitoring [22–25], noninvasive blood glucose measurement through with infrared absorption and polarization in the eye, respiration evaluation by analyzing breathing sounds, and noninvasive measurement of intraocular pressure.

6. SUMMARY

Due to the demographic changes presented in this chapter (with a special focus, but not limited to the situation in Germany), medical technology has to and will change its face. Medical care and especially medical technology must adapt to the special needs of a growing elderly population (while not forgetting the young and middle-aged population). Due to a lower economic burden, more medical care will be delivered at home.

These developments will change the medical device market. In the future, the typical customer of medical technology will probably not be a hospital buying equipment for an intensive care unit, but it may very well be a private customer or a provider for nursery homes.

It can be expected that the public will discuss the economic questions connected to these demographic changes more vigorously in the near future. To the author, it seems rather likely that large parts of the financial burden connected to medical care, especially when dealing with not directly life-threatening disorders, will have to be shouldered by the patients.

Due to the already existing large patient populations, major target areas already are and will remain cardiovascular diseases and diabetes. Other areas, like neurological diseases, pulmonary diseases, and orthopaedic problems, will further grow as we become older.

REFERENCES

[1] Fa. Biotronik, www.biotronik.de last visited on May 11th, 2005.
[2] Park, S. and Hayaraman, S., 2003, Enhancing the quality of life through wearable technology, *IEEE Eng. Med. Biol. Mag*., **22**(3), 41–48.
[3] Jovanov, E., O'Donnel L. A., Raskovic, D., Cox, P. G., Adhami, R. and Andrasik, F., 2003, Stress monitoring using a distributed wireless intelligent sensor system, *IEEE Eng. Med. Biol. Mag*., **22**(3), 49–55.
[4] Jovanov, E., Milenkovic, A., Otto, C. and de Groen, P. 2005, A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation, Journal of NeuroEngineering and Rehabilitation, **2**(6), http://www. jneuroengrehab.com/content/2/1/6
[5] Statistisches Bundesamt (Federal German Office for Statistics), Bevölkerung Deutschlands bis 2050 (German population til 2050). www.destatis.de 2003.
[6] Cambridge Neurotechnology Ltd., Cambridge, UK. http://www. camntech .com last visited March 14th, 2005.
[7] Mini Mitter Company, Inc., Bend, OR, USA, www.minimitter.com last visited March 14th, 2005.
[8] BodyMedia, Inc., Pittsburgh, PA., www.bodymedia.com last visited on March 19th, 2005.
[9] HealthWear™ Product Center, Indianapolis, IN, USA. www.health-wear .com visited on June 11th, 2004.
[10] Lau, C.-P.,1993, Rate Adaptive Cardiac Pacing: Single and Dual Chamber, Futura Publishing, Inc., Mount Kisco, NY.
[11] Schhaldach, M., 1992, Electrotherapy of the Heart, Springer-Verlag, Berlin.
[12] Webster, J. G. (ed.), 1995, Design of Cardiac Pacemakers, IEEE Press, Piscataway, NJ.
[13] Polar Electro Oy, Kempele, Finnland, www.polar.fi last visited on March 19th, 2005.
[14] Polar USA, www.polarusa.com visited on June 6th, 2004.
[15] PicoMed GmbH, Überlingen, FRG, www.picomed.de last visited March 19th, 2005.
[16] Tapuz Medical Technology (T.M.T 2004) Ltd., www.tapuz.com last visited March 16th, 2005.
[17] Braun GmbH, Kronberg, FRG, www.braun.de last visited March 19th, 2005.
[18] Asada, H. H., Shaltis, P., Reisner, A., Rhee, S. and Hutchinson, R. C., 2003, Mobile monitoring with wearable photoplethysmographic sensors, *IEEE Eng. Med. Biol. Mag*., **22**(3), 28–40.
[19] Nellcor Puritan Bennett Inc., Pleasanton, CA, USA, www.nellcor.com last visited March 19th, 2005.
[20] *Vivometrics*. Vivometrics, Inc., http://www.vivometrics.com/responder/press_ima last visited March 16th, 2005.
[21] *Sensatex*. Sensatex Inc., www.sensatex.com last visited March 16th, 2005
[22] Elter, P., Methoden und Systeme zur nichtinvasiven, kontinuierlichen und belastungsfreien Blutdruckmessung, Ph.D. Thesis, Faculty for Electrical

Engineering and Information Technology, University of Karlsruhe, Germany, 2001.
[23] Vollmer, P., Boll, P., Groamann, U., Ottenbacher, J., Stork, W., Müller-Glaser, K. D. and Lutter, N., 2004, Entwicklung eines mobilen, kontinuierlichen und nicht-invasiven Blutdruckmessgerätes mit Bluetooth Datenübertragung, *Biomedizinische Technik*, Band 49, Ergänzungsband 2, Teil 1, pp. 240–241.
[24] Siebert, C., Kozma, E., Boll, P. and Lutter, N., 2004, Vergleich von Laser-Doppler-Flussmessung und Photoplethysmografie für die kontinuierliche, nichtinvasive Bestimmung des Blutdrucks, *Biomedizinische Technik*, Band 49, Ergänzungsband 2, Teil 1, pp. 430–431.
[25] Siebert, C., Löhner, M., Kozma, E., Boll, P. and Lutter, N., 2004, Antwortzeiten der kontinuierlichen, nichtinvasiven Blutdruckmessung bei raschen Blutdruckanstiegen und –abfällen, *Biomedizinische Technik*, Band 49, Ergänzungsband 2, Teil 1, pp. 432–433.
[26] Lymberis, A. and Olson, S., Intelligent biomedical clothing for personal health and disease management: state of the art and future vision, Telemedicine Journal and e-Health, **9**(4), 379–386.
[27] Vuorela, T., Kukkonen, K., Rantanen, J., Järvinen, T. and Vanhala, J., Bioimpedance measurement system for smart clothing, 7th IEEE International Symposium on Wearable Computers (*ISWC'03*), White Plains, NY, USA, Oct. 21–23rd, 2003.
[28] Marculescu, D., Marculescu, R., Zamora, N. H., Stanley-Marbell, P., Khosla, P. K., Park, S., Jayaraman, S., Jung, S., Lauterbach, C., Weber, W., Kirstein, T., Cottet, D., Grzyb, J., Tröster, G., Jones, M., Martin, T. and Nakad, Z., 2003, Electronic textiles: A platform for pervasive computing, *Proceedings of the IEEE*, **91**(12), 1995–2018.
[29] Starner, T., 1996, Human-powered wearable computing, IBM Systems Journal, **35**(3&4), 618–629.
[30] http://www.sensation-eu.org/ last visited March 20th, 2005.
[31] http://www.hitech-projects.com/euprojects/myheart last visited March 20th, 2005.
[32] http://www.wealthy-ist.com last visited March 20th, 2005.
[33] http://www.ibmt.fraunhofer.de/Produuktblaetter/MT_topcare_en.pdf last visited March 20th, 2005.
[34] http://www.phmon.de/englisch/ last visited March 20th, 2005.
[35] http://www.who.int/en/ last visited May 11th, 2005.

Chapter 6.2

CARBON NANOTUBE FIELD-EFFECT TRANSISTORS
The Importance of Being Small

Joachim Knoch
Institute for Thin Films and Interfaces, Forschungszentrum Jülich
j.knoch@fz-juelich.de

Joerg Appenzeller
IBM Research
joerga@us.ibm.com

Abstract In this chapter we elucidate the peculiarities of one-dimensional field-effect transistors by studying electronic transport in carbon nanotube field-effect transistors (CNFETs). It is shown that the "geometrical smallness"—meaning the extremely small diameter of carbon nanotubes—as well as the "electrical smallness," (i.e., the one-dimensionality of electronic transport in nanotubes), determine the electrical response of CNFETs. A model for the simulation of this electrical response of CNFETs is introduced and predictions based on this model are compared with analytical expressions as well as with experiments. It turns out that the particular behavior of CNFETs can be explained within a Schottky barrier transistor model with modified electrostatics and 1D transport. As an example of this behavior, the appearance of multimode transport in CNFETs is discussed in detail and the impact on the electrical characteristics is illuminated.

Keywords carbon nanotube field-effect transistors; multimode transport; one-dimensional transport; quantum capacitance; Schottky barrier

1. INTRODUCTION

Since approximately seven years, researchers have been exploring carbon nanotube field-effect transistors (CNFETs) as building blocks of a future nanoelectronics. The increasing interest in CNFETs stems from the fact that carbon nanotubes exhibit extraordinary electronic and structural

S. Mukherjee et al. (eds.), AmIware, 371–402.

properties. They hold promise for ballistic transport over distances as large as a few hundred nanometers at room temperature and are extremely small objects, ideally suited for aggressively scaled field-effect transistor (FET) applications. It is in this sense that they are potentially enabling a new level of functionality for *ambient intelligence*.

Figure 6.2-1 shows an SEM of a typical CNFET. A highly doped silicon substrate serves as a large area back gate. The nanotube is separated from the substrate by a gate oxide, 10 nm in thickness, grown on top of the silicon. Contacts made of metals, such as aluminum, titanium, or palladium, are used as source and drain. The electrical characteristics of such a CNFET very much resemble that of conventional MOSFETs. An example of typical output characteristics for a *p*-type device is shown in Figure 6.2-2. Consequently, the electronic transport has been interpreted in terms of conventional MOSFETs [1, 2]. However, the appearance of short-channel-like effects in electrostatically well-tempered devices [3] has stimulated further investigations aiming at a more profound understanding of transport in CNFETs. It is the extremely small geometry in combination with the "electrical smallness" (i.e., one dimensionality of carbon nanotubes) that leads to the different electrical response of CNFETs if compared to bulk-like silicon MOSFETs. For instance, a strong dependence of the electrical characteristics on the gate oxide thickness was found, which could consistently be explained with the presence of Schottky barriers at the contact interfaces [4, 5].

In this chapter we will elaborate in detail on the impact of the "geometrical" and "electrical smallness" on the device behavior of CNFETs using an analytical analysis supported by simulations. To this end a model for the quantum mechanical simulation of the electronic transport in CNFETs will be introduced in Section 2. Subsequently, particular features and benefits of the smallness of carbon nanotubes for the device performance will be discussed. Comparing analytical approximations with simu-

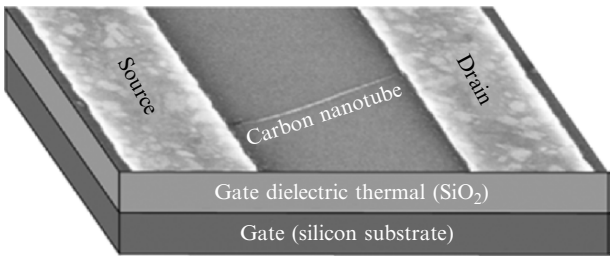


Figure 6.2-1. Electron micrograph of a CNFET. A large area back gate is separated from the tube by a gate dielectric.

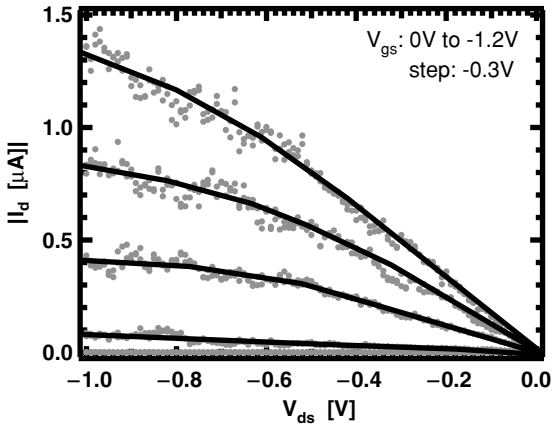


Figure 6.2-2. Typical output characteristics of a *p*-type CNFET. The lines are guides to the eye.

lations using the model presented in Section 2, we will explain the relevant phenomena.

2. MODELING TRANSISTOR ACTION

In this section, a general framework of a model for CNFETs and the quantum mechanical simulations of their electrical behavior will be introduced. Our model is rather general and applies to a variety of 1D FET structures. Nanotube-specific properties are accounted for by material characteristics, such as the effective masses, the energy gap, and diameter. Our model is able to describe:

- Ballistic as well as scattering limited transport in the nanotube channel;
- Single as well as multimode transport, including several conduction and valence bands with different effective masses in the conduction and valence bands;
- Tunneling phenomena, in particular, direct source to drain tunneling, tunneling through the gate oxide, and band-to-band tunneling; and
- Contact effects by attaching metallic source and drain reservoirs to the semiconducting tube channel.

The model will be used to discuss and explain peculiarities of the electronic transport in 1D FET devices that are related to their "geometrical" and "electrical smallness." The simulations are based on a self-consistent solution of the Schrödinger equation using the nonequilibrium Green's function formalism (NEGF), and the Poisson equation.

As was mentioned in the introduction, CNFETs behave like Schottky barrier FETs and consequently, in our simulations, we consider a model, where metallic source and drain electrodes are in direct contact with the nanotube and a Schottky barrier of varying height builds up at the contact interfaces. Although this model has been shown to have shortcomings [6], it is able to explain the transport phenomena relevant for the present analysis.

2.1. Electrostatics

Despite its smallness, a real CNFET is nevertheless a 3D object and a 3D calculation should be employed in order to simulate its electrical characteristics accurately. This, however, represents an enormous computational burden. Fortunately, in the case of a gate that is wrapped around the nanotube channel (as shown in Figure 6.2-3), it is possible to reduce the electrostatics to a 1D one using the approach by Auth and Plummer [7]. This approach captures all relevant aspects related to the scaling of the gate oxide thickness d_{ox}, the body (i.e., channel) thickness d_{nt}, and of course the appearance of short-channel effects in laterally scaled devices. The main ingredient is the approximation of the potential distribution in the channel perpendicular to the direction of current transport by a quadratic expansion, which leads in cylindrical coordinates to

$$\Phi(r,x) \approx c_0(x) + c_1(x)r + c_2(x)r^2, \tag{6.2-1}$$

where $r = 0$ is at the center of the nanotube. Due to the cylindrical symmetry, the following boundary conditions apply: $\partial\Phi/\partial r|_{r=0} = 0$ at

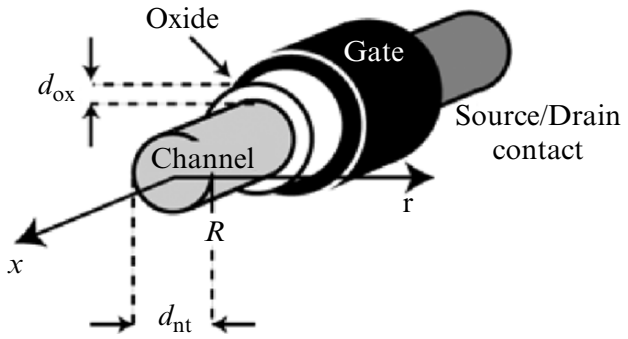


Figure 6.2-3. Schematics of a CNFET with a gate wrapped around the channel. The source/drain contacts can be either a metal or a doped nanotube segment.

the center and $\partial\Phi/\partial r|_{r=R} = \varepsilon_{ox}/\varepsilon_{nt} \cdot (\Phi_g - \Phi_f)/(R \ln(1 + d_{ox}/R))$ at the border of the channel [7]. Here, $R = d_{nt}/2$, $\Phi_f(x) = \Phi(r = R,x)$ is the surface potential and Φ_g is the gate potential. ε_{nt} and ε_{ox} are the relative dielectric constants of the nanotube and the gate oxide, respectively. With these boundary conditions, a reduced Poisson equation for the surface potential $\Phi_f = \Phi(r = R,x)$ can be derived. Inserting Equation (6.2-1) into the Poisson equation yields the constants following 1D modified Poisson equation for Φ_f [7]

$$\frac{d^2\Phi_f}{dx^2} - \frac{\Phi_f - \Phi_g + \Phi_{bi}}{\lambda^2} = -\frac{e(\rho \pm N)}{\varepsilon_0 \varepsilon_{nt}} \tag{6.2-2}$$

that will be used throughout this article [1]. Here, Φ_{bi} is the built-in potential, ρ is the density of mobile carriers, and N a constant charge background due to a doping of the nanotube with either donors (“+”-sign) or acceptors (“−”sign). The length λ is given by $\lambda = \sqrt{\varepsilon_{nt} d_{nt}^2 \ln(1 + 2d_{ox}/d_{nt})/8\varepsilon_{ox}}$ [7]. This modified Poisson equation describes well the electrostatics of a fully depleted FET and has been successfully applied to investigate silicon-on-insulator MOSFETs [9, 10]. A CNFET is the ideal fully depleted device, since in a nanotube, there is no “bulk”-part giving rise to a depletion capacitance.

Equation (6.2-2) allows easy access to a number of insights into the nanotube device characteristics. For instance, considering the limit $\rho \approx 0$, which-corresponds to the off-state of a CNFET, Equation (6.2-2) can be solved analytically leading to a solution of the form:

$$\Phi_f(x) \propto \exp\left(-\frac{x}{\lambda}\right). \tag{6.2-3}$$

This means that λ determines the length scale, on which potential variations are being screened. The screening becomes more effective if devices exhibit ultrathin bodies d_{nt} and ultrathin gate oxides d_{ox}. Scaling the channel length down improves the device performance but eventually can lead to a loss of the electrostatic gate control of the channel. Equation (6.2-2) states that electrostatic integrity is preserved as long as λ is smaller than the channel length L. Since in nanotubes—inherently small objects—λ can be made very small without the introduction of dopants, they seem to be ideally suited for

[1] Note, that the quadratic approximation is strictly valid if λ is much larger than $R = d_m/2$ [8]; in virtually all cases considered here $\lambda > d_m/2$ is fulfilled.

the realization of ultimately scaled devices from the electrostatics point of view. At the same time, the complete absence of dopants prevents variations in characteristics from device-to-device, which is known to be a major problem in today's aggressively scaled conventional MOSFETs.

While a wrapped-around gate as was discussed above is ideal for preserving electrostatic integrity [7] experimental devices to date usually have a single, planar gate as is shown in the electron micrograph (Figure 6.2-1). However, comparing our simulations with experimental characteristics, we consistently find that the electrostatics of such devices can be described reasonably well by the same Poisson equation (Equation 6.2-2), but with a larger screening length of the form $\lambda = \sqrt{d_{nt} d_{ox} \varepsilon_{nt} / \varepsilon_{ox}}$, which was originally derived for a planar gate, planar channel geometry, such as a single-gated MOSFET on SOI [9, 10]. Unless stated otherwise, the reduced Poisson equation with this larger λ will be used in the following.

2.2. Quantum Transport Equations

For the calculation of the charge in and current through the nanotube, we employ the nonequilibrium Green's function formalism. Together with the reduced Poisson equation (Equation 6.2-2), our model allows the self-consistent calculation of the electronic transport in CNFETs.

To numerically compute the Green's functions, we make use of Datta's approach [11, 12]. We consider a 1D finite difference scheme with lattice constant a and nearest neighbor hopping parameter $t^{\mp}$ as illustrated in Figure 6.2-4(a). A quadratic dispersion relation in the conduction and valence band (see Figure 6.2-4(b)) is used with effective masses being m^*_{con} and m^*_{val}, respectively. In order to describe the complex band structure in the band gap, we make use of Flietner's dispersion relation [13, 14]. At lattice point i this leads to

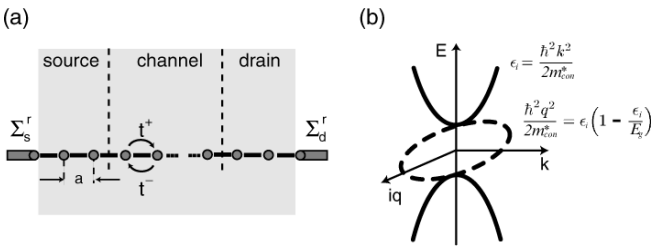


Figure 6.2-4. (a) Finite difference grid with lattice spacing a and (b) shows the conduction and valence bands and the complex band structure in the band gap for $\alpha = 0$.

$$\frac{\hbar^2 q^2}{2m^*_{con}} = \varepsilon_i \left(1 - \frac{\varepsilon_i}{E_g}\right)\left(1 - \alpha \frac{\varepsilon_i}{E_g}\right)^{-2}, \tag{6.2-4}$$

with $\alpha = 1 - \sqrt{m^*_{con}/m^*_{val}}$ and $\varepsilon_i = E - \Phi^i_f$ being the kinetic energy. For the specific case of a semiconducting nanotube, the effective masses in the conduction and valence bands are equal (i.e., $m^*_{con} = m^*_{val} = m^*$ so that $\alpha = 0$), and the charge neutrality or branching point E_{br} is at midgap. The discretized form of the Hamiltonian follows to be

$$H_{i,j} = \begin{cases} t^- + t^+ + \Phi^i_f & i = j \\ -t^- & i\text{ -1 } = j \\ -t^+ & i\text{ + 1 } = j \\ 0 & otherwise \end{cases} \tag{6.2-5}$$

where *i,j* are indices of different points of the finite difference grid. $t^{\mp} = [1/2t_i + 1/2t_{i\mp1}]^{-1}$ and the hopping parameters t_i equal $\hbar^2/2m^*a^2$ in the conduction/valence band and $t_i = \hbar^2/(2m^*a^2(1 - \varepsilon_i/E_g))$ within the band gap of magnitude E_g. The retarded Green's function G^r in matrix notation (discrete form) is given by the expression

$$G^r_{i,j} = \frac{1}{a}[(E \times \underline{1} - H - \Sigma^r)^{-1}]_{i,j}, \tag{6.2-6}$$

where $\underline{1}$ is the unity matrix. The self-energy function Σ^r is a sum of contributions that account for: (1) the coupling of the channel to semi-infinite source and drain contacts and (2) scattering in the channel. In principle, gate leakage can be incorporated via Σ^r as well [15], but has been omitted here for simplicity. In the following, we consider only metallic source and drain contacts with a Schottky barrier of height Φ_{SB} with respect to the conduction/valence band as is illustrated in Figure 6.2-5, where the gray line represents the conduction band (the valence band has been omitted in the illustration for clarity), and the small circles refer to points of the finite difference grid. Scattering in the nanotube channel is accounted for by attaching Buettiker probes via appropriate self-energy functions to each site of the finite difference grid [16], which is also shown in Figure 6.2-5. Each Buettiker probe has its own Fermi level E^j_f such that the dotted line in Figure 6.2-5 represents the position dependent quasi-Fermi level in the channel. In the case considered here, each electrode at site *j* (either Buettiker probe or source/drain contact) contributes to Σ^r, a matrix, which has zero entries everywhere except at (*j,j*) such that Σ^r

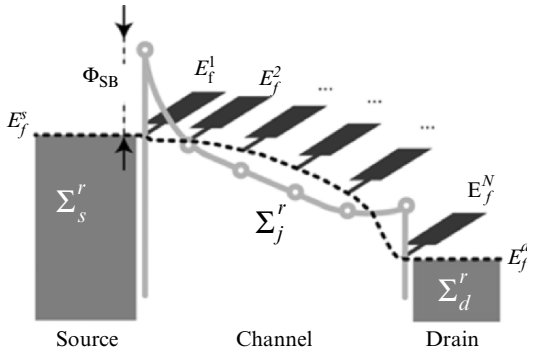


Figure 6.2-5. Incorporation of scattering by attaching floating Buettiker probes at each grid point. The resulting Fermi levels in the probes constitute the quasi-Fermi-level distribution along the current transport direction.

becomes a diagonal matrix. To be specific, the diagonal elements of Σ^r are given by

$$\Sigma_{s,d}^r = -\frac{(t^-)^2}{t_{met}} \exp\left(ia\sqrt{\frac{2m_0(E + E_f^{s,d})}{\hbar^2}} \right);$$
$$\Sigma_j^r = -\gamma t \exp\left(ia\sqrt{\frac{2m^*(E - \Phi_f^j)}{\hbar^2}} \right), \tag{6.2-7}$$

where $j = 2(N-1)$ runs over all sites with Buettiker probes and N being the dimension of the finite difference grid; $\Sigma_{s,d}^r$ are (1,1) and (N,N) entries, respectively, accounting for the coupling to the source/drain contacts. The coupling parameter $t_{met} = \hbar^2/2m_0a^2$, and the parameter γ describe the coupling of the probes to the channel. The difference between the source/drain contacts and the probes is that the Fermi energy in the Buettiker probes is not fixed by a certain terminal voltage, but individually floats to a value so that the total current flowing into and out of each Buettiker probe sums up to zero. During the self-consistent calculations, the appropriate Fermi energies in the probes are found by an iterative procedure; for details, see Ref. [16]. Since carriers can enter one probe at a particular energy and leave it at another energy, each Buettiker probe represents a dissipative scattering site, whose effectiveness is mediated by the coupling γ of the probe to the channel. As was shown by Venugopal (see appendix of Ref. 16), one can associate the mean free path of carriers in the channel l_{scat} with the parameter γ:

$$l_{scat} = 2a \times \frac{1}{\gamma}. \tag{6.2-8}$$

This means, if γ is large (i.e., if the Buettiker probe is tightly bond to the channel), the mean distance between scattering becomes short and vice versa.

Having calculated G^r, the local density of states (LDOS) can be determined. Figure 6.2-6 shows a gray scale image of the LDOS in a ballistic SB-CNFET (i.e., with $\gamma = 0$). Dark areas refer to regions with a low DOS, whereas lighter areas indicate a high DOS. The standing wave pattern in the channel of the SB-CNFET is a result of the Fermi velocity mismatch between metal and nanotube. The nonzero DOS at the contact interfaces is due to metal wave-functions penetrating the semiconductor gap, where they get exponentially damped.

In order to determine the charge density ρ—needed in the Poisson equation to compute the potential distribution—one has to calculate the electron and hole correlation functions $G^n = G^r \cdot \Sigma^{in} \cdot G^a$ and $G^p = G^r \cdot \Sigma^{out} \cdot G^a$ (we use Datta's notation everywhere [11]) with $G^a = (G^r)^\dagger$; the "$\cdot$" stands for matrix multiplication. The matrices $\Sigma^{in,\,out}$ are related to the retarded self-energy function(matrix), and thus are diagonal matrices as well with their diagonal entries given by $\Sigma^{in}_{j,\,j} = f(E_f^j)\Gamma_{j,\,j}$ and $\Sigma^{out}_{j,\,j} = (1 - f(E_f^j))\Gamma_{j,\,j}$. Here, $f(E_f^j)$ is the equilibrium Fermi distribution with Fermi energy E_f^j of the jth electrode (i.e., source and drain if $j = 1$ or $j = N$, and a Buettiker probe otherwise), and $\Gamma_{j,\,j} = -2\mathrm{Im}(\Sigma^r_{j,\,j})$. The charge density is then given by

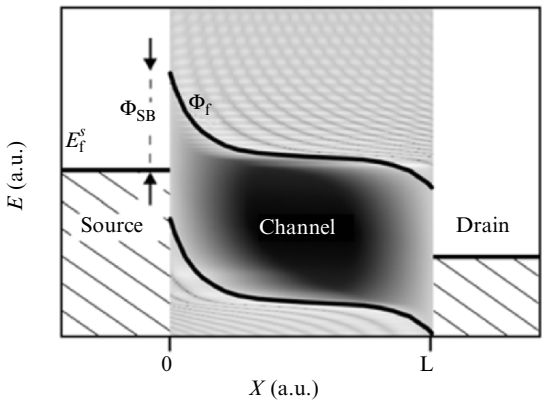


Figure 6.2-6. Local density of states for a CNFET with Schottky contacts.

$$\rho_j = \frac{2}{\pi}\int_{E_{br}(x)}^{\infty} dE G^n_{j,\,j}(E) - \frac{2}{\pi}\int_{-\infty}^{E_{br}(x)} dE G^p_{j,\,j}(E), \qquad (6.2\text{-}9)$$

where the first term on the right hand side determines the electron contribution and the second term the hole contribution to the charge density; the factor of 2 in the expression for ρ_j is due to the 2 conduction channels in a nanotube. Poisson's equation (Equations (6.2-2) and (6.2-9)) are solved together iteratively until a self-consistent solution is found [2]. Finally, the current flowing from source to drain is computed using the Fisher–Lee relation [16, 18]

$$I_d = \frac{4e}{h}\int dE \sum_j T^{N,\,j}(E)\,[f(E - E_f^d) - f(E - E_f^j)], \qquad (6.2\text{-}10)$$

where the transmission coefficient between the drain contact at site N and another electrode at site j is $T^{N,\,j}(E) = \Gamma_{N,N}|G^r_{N,j}|^2\Gamma_{j,j}$. Equation (6.2-10) is the Landauer result, where $4e/h$ is the current one mode can carry and the integral gives the number of modes that contribute to the current.

3. TRANSPORT IN CNFETS

After having introduced a model for the simulation of CNFETs, we will now use this model to study the electronic transport in these devices. We will first focus on the impact of the ''geometrical'' and then on the impact of the ''electrical smallness'' of carbon nanotubes on the performance of CNFETs. The ''geometrical smallness'' plays the dominant role for the electrostatics, whereas the ''electrical smallness'' (i.e., the 1D of carbon nanotubes) is reflected in a distinctly different density of states. As is discussed below, this makes the so-called quantum capacitance becoming a relevant entity. The magnitude of this capacitance compared to the oxide capacitance C_{ox} strongly influences the performance of CNFETs.

[2] A Newton–Raphson method is used to determine the update of Φ_f at each iteration step [17]. This method leads to self-consistency within a few iterations.

3.1. Impact of the Small Geometry

It has been discussed in Section 2.1, that the "geometrical smallness" (i.e., the small body thickness d_{nt} of nanotubes) leads to an exponential screening of potential variations. This is also true for the Schottky barriers at the contact channel interfaces and implies that the Schottky barriers can be made very "thin" by letting $\lambda \to 0$. (Note that this is distinctly different from a large area metal semiconductor contact. In this case, the potential inside the semiconductor falls off (approximately) proportional to the square instead of exponentially.) Consequently, one can expect the electrical behavior of CNFETs to strongly depend on λ. Here we will study the impact of λ on the electrical characteristics of CNFETs. We will concentrate first on the off-state and then briefly discuss the on-state. As it turns out, both, the off-state as well as the on-state strongly depend on oxide and body thickness.

3.1.1. Off-state

In a fully depleted conventional long channel device, the inverse subthreshold slope $S = (\partial \log I_d / \partial V_{gs})^{-1}$ should always be 60 mV/dec, independent of the actual oxide thickness d_{ox} (as long as short-channel effects are absent and no traps at the oxide-channel interface are present). This fact can already be deduced from a closer look at Equation (6.2-2). Far away from the source-channel interface, the curvature of $\Phi_f \approx 0$ and, because $\rho \approx 0$ in the off-state, Equation (6.2-2) reduces to [3]

$$-\frac{\Phi_f^0 - \Phi_g + \Phi_{bi}}{\lambda^2} = 0, \tag{6.2-11}$$

where Φ_f^0 is the potential in the channel far away from the contact interface. As illustrated in Figure 6.2-7(a), this potential barrier determines the current flow through the channel: only electrons in source that are thermally excited over Φ_f^0 can contribute to the current. From the equation above, it becomes apparent that a change in gate voltage results in the same change of the surface potential Φ_f^0 (i.e., $\delta\Phi_g = \delta\Phi_f^0$) leading to an $S = 60\,\text{mV/dec}$ [19].

In Schottky barrier devices, the situation is different: $\delta\Phi_g = \delta\Phi_f^0$ also holds far away from the Schottky barrier (see Figure 6.2-7(b)), but unlike the case above it is not Φ_f^0 anymore that determines the current, but the

[3] A nonzero doping density N only changes the built-in potential Φ_{bi} and hence will have no effect on the off-state of the CNFET.

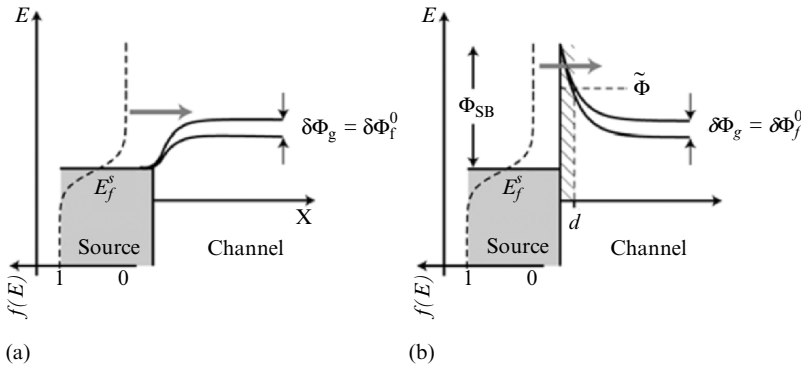


Figure 6.2-7. Change of the potential Φ_f at the source end of the channel for two different Φ_g for (a) a conventional-type FET and (b) an SB-CNFET.

Schottky barriers at the channel-contact interfaces. This situation is shown in Figure 6.2-7(b): a change in Φ_g yields the same change in Φ_f only far away from the contact interface (black line). The current injection, however, is determined by the change of the Schottky barrier.

As was discussed above (see Section 2.1), λ is the relevant length scale of potential variations and hence determines the shape of the Schottky barrier. In order to quantify the influence of λ, we derive a first-order expression for the inverse subthreshold slope. In the off-state, the density of mobile carriers can be neglected and the potential at the source Schottky diode is given by

$$\Phi_f(x) = (\Phi_{SB} - \Phi_f^0)e^{-\frac{x}{\lambda}} + \Phi_f^0. \tag{6.2-12}$$

For not too large oxide thicknesses and if $\Phi_{SB} \gg k_B T$ the current is mainly determined by the tunneling component of carriers originating from an energy range between $\leq \Phi_{SB}$ and $\tilde{\Phi}$ with $\tilde{\Phi} < \Phi_{SB}$. The reason for this is that below $\tilde{\Phi}$ the tunneling probability is exponentially suppressed and we can take $\tilde{\Phi}$ as a lower cutoff energy. Above Φ_{SB}, the Fermi function has become negligibly small, because $\Phi_{SB} \gg k_B T$. We now make the crude approximation that the tunneling probability is unity between $\tilde{\Phi}$ and Φ_{SB} and zero below $\tilde{\Phi}$. This is justified because of the exponential shape of the Schottky barrier (cf. Equation 6.2-3), which is a direct consequence of the small geometric dimensions of nanotubes. The numerical value of S defines the change of gate voltage needed to alter the current I_d by one order of magnitude. For V_{ds} large enough, the contribution of the drain contact to I_d can be neglected and since we

approximated the transmission probability to be 1 above $\tilde{\Phi}$, S can be computed by requiring the ratio of the drain currents for two different gate potentials (voltages) to be equal to 10:

$$10 \stackrel{!}{=} \left(\int_{\tilde{\Phi}(\Phi_{g1})}^{\infty} dE \exp\left(-\frac{E - E_f^s}{kT} \right) \right) \Bigg/ \left(\int_{\tilde{\Phi}(\Phi_{g2})}^{\infty} dE \exp\left(-\frac{E - E_f^s}{kT} \right) \right), \quad (6.2\text{-}13)$$

where $f(E - E_f^s)$ has been approximated by an exponential. Note, that we put the upper limit of the integration to infinity. The error, however, is negligible, if $\Phi_{SB} \gg k_B T$. As a result

$$\tilde{\Phi}_1 - \tilde{\Phi}_2 = k_B T \ln(10). \quad (6.2\text{-}14)$$

One can associate a tunneling distance d with the lower cutoff $\tilde{\Phi}$ (i.e., $\tilde{\Phi} = \Phi_f(d)$), as illustrated in Figure 6.2-7(b); d is the thickness of the Schottky barrier, for which the tunneling contribution becomes negligible. This means, that for a certain set of parameters of the CNFET (such as effective mass, Schottky barrier height, etc.), d is constant but does, of course, depend on those parameters, in particular on the effective mass (see below). Inserting this tunneling distance d into Equation (6.2-12) yields $\tilde{\Phi}_{1,2} = (\Phi_{SB} - \Phi^0_{f_{1,2}}) \exp(-d/\lambda) + \Phi^0_{f_{1,2}}$ for the two lower cutoff energies and we get

$$\Phi^0_{f_1} - \Phi^0_{f_2} = \frac{k_B T \ln(10)}{1 - e^{-d/\lambda}}. \quad (6.2\text{-}15)$$

As long as $d < \lambda$, which is the case for realistic gate oxide and body thicknesses, we can expand the exponent to first order in d/λ. Decreasing the effective mass requires a larger d in order to get the same (low) tunneling probability at the lower cutoff energy $\tilde{\Phi}$ and vice versa. If to first order the Schottky barrier is approximated by a barrier of thickness d and constant height, the WKB approximation [20] yields a tunneling probability of $\exp(-const \cdot \sqrt{m^*} d)$. It is then easy to see that for the tunneling probability to be constant, d scales as $1/\sqrt{m^*}$. Finally, noting that $\delta\Phi_f^0 = \delta\Phi_g$ in the off-state, we obtain:

$$S \propto \sqrt{m^*} \sqrt{\frac{\varepsilon_{nt}}{\varepsilon_{ox}} d_{nt} d_{ox}} k_B T \ln(10). \quad (6.2\text{-}16)$$

Hence, S is proportional to $\sqrt{d_{ox}}$ as was already reported by Heinze and coworkers [21], and experimentally verified by Appenzeller and coworkers [4]. In addition, S depends in the same way on the body thickness and effective carrier mass of the channel material. The above dependence of S on the various parameters of the SB-CNFET has been confirmed using the self-consistent simulation discussed earlier and is shown in Figure 6.2-8. Here, S is plotted versus $\sqrt{d_{ox}}$ and constant $d_{nt} = 1\,\text{nm}$. As can be seen, the S values lie on a straight line as predicted by the analytical approximation. Only for the smallest $d_{ox} = 0.5\,\text{nm}$ S deviates from the behavior, but for such a thin oxide λ becomes smaller than d so that the exponent in Equation (6.2-15) cannot be expanded anymore. The same graph follows, if $d_{ox} = 1\,\text{nm}$ and d_{nt} varies since λ depends on d_{ox} and d_{nt} in the same way; the square-root dependence of S on the effective mass has been confirmed by simulations too (not shown here). For comparison, $S = 60\,\text{mV/dec}$, independent of d_{ox} as expected for a conventional well-behaved MOSFET is shown in Figure 6.2-8 as well (gray line). Finally, note that for the derivation of Equation (6.2-16), we did not have to assume any nanotube specific properties and thus the above expression for S is quite generally valid for ultrathin body FETs. As a result, a good off-state in SB-FETs can be expected, if d_{ox}, d_{nt}, and m^* are made as small as possible [22].

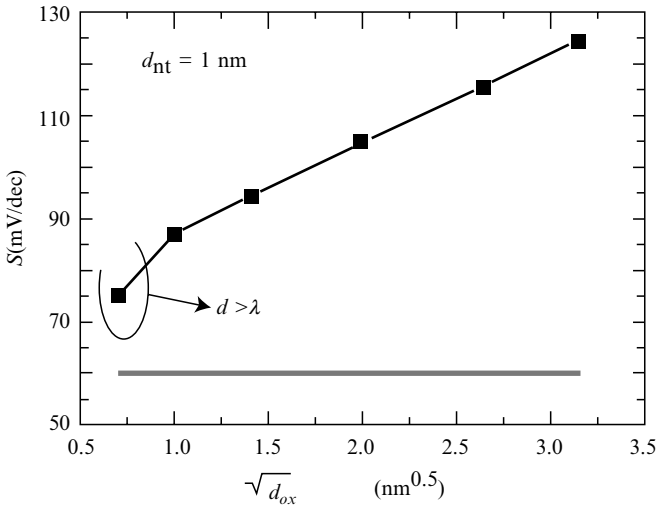


Figure 6.2-8. Inverse subthreshold slope S as function of $\sqrt{d_{ox}}$ (black line squares). For not too thin oxides, a linear relationship is obtained. Gray line is S of a conventional-type MOSFET.

3.1.2. On-state

To illuminate the influence of the device geometry on the on-state, we approximate the current I_d for small bias V_{ds} as follows

$$I_d = \frac{4e}{h} \int dET(E)\,(f_s - f_d) \approx \frac{4e^2}{h} V_{ds} \int dET(E)\left(-\frac{\partial f}{\partial E}\right), \qquad (6.2\text{-}17)$$

where $T(E)$ is the transmission probability at energy E. Here, we consider the case of a ballistic SB-FET such that $T(E_f^s)$ is due to the presence of the Schottky barriers at the channel contact interfaces only. Since $-\partial f/\partial E$ is peaked around E_f^s, the integration yields approximately

$$I_d \approx \frac{4e^2}{h} V_{ds} T_{SB}(E_f^s). \qquad (6.2\text{-}18)$$

This expression is the Landauer formula (see Refs. [11, 23], for instance) for transport in mesoscopic systems multiplied by the tunneling probability through the Schottky barriers. The transconductance g_m follows to be

$$g_m = \frac{4e^2}{h} V_{ds} \frac{\partial T_{SB}(E_f^s)}{\partial V_{gs}}. \qquad (6.2\text{-}19)$$

Usually an SBFET exhibits a performance inferior to a conventional type device because of the relatively large Schottky barriers. However, due to the strong variation of the potential distribution at the Schottky barriers and thus due to the strong dependence of $T_{SB}(E_f^s)$ on λ, excellent transconductances can be expected for CNFETs. In fact, recently SB-CNFETs have been realized with thin gate oxides that show an excellent on-state with g_m of up to 4.3 μS [3, 24]. Despite a relatively high Schottky barrier of around 0.3 eV, the g_m of these devices surpasses those of conventional Si-MOSFETs. Again, the reason for this is the extremely small diameter d_{nt} of carbon nanotubes that leads in combination with thin gate dielectrics to a very tight gate control of the Schottky barriers at the contact interfaces. This fact has a number of unusual implications: For instance, it was found that extracting the Schottky barrier height from experimental characteristics using the thermionic emission theory yields a drastic underestimation of the true barrier height; in principle even negative barriers can result from an analysis using thermionic emission [25]. This means that the barrier in the on-state becomes highly transparent with tunneling probabilities close to one allowing for excellent on-states.

3.2. Electrical Smallness

Besides their extreme "geometric smallness," the most prominent feature of carbon nanotubes is their "electrical smallness": carbon nanotubes are objects, where each subband represents a 1D channel for conduction leading to a different DOS in the channel if compared to conventional Si-MOSFETs, for instance. It was mentioned earlier that in such low-dimensional systems, the so-called quantum capacitance C_q becomes a relevant entity. Due to its importance for the electronic transport in 1D FETs, we will introduce the concept of the quantum capacitance and discuss its dependence on the dimensionality of the device under consideration. In addition, the influence of Schottky barriers at the channel contact interfaces on C_q will be studied. Finally, the impact on the device characteristics will be discussed.

3.2.1. The quantum capacitance

From a conventional MOS capacitor, it is known that well above the threshold voltage $V_{th}(V_{gs} > V_{th})$, the surface potential Φ_f^s (i.e., the potential at the gate oxide-channel interface), hardly moves anymore with increasing gate voltage and the MOS system then behaves like a parallel plate capacitor [19]. In a CNFET, however, the surface potential can under certain circumstances be influenced by the gate, even if $V_{gs} > V_{th}$. The reason for this completely different response of the bands to the gate potential is the decreased charge carrying capability of nanotubes due to their 1D density of states. The relationship between charge and response of the bands to a gate potential gives rise to a capacitance, the aforementioned quantum capacitance C_q [26–28]. The ratio of the quantum capacitance C_q and oxide capacitance C_{ox} is key to the understanding of (e.g., the appearance of multimode transport) CNFETs as is discussed later.

Consider an *n*-type, long-channel CNFET in the on-state such that $\rho \neq 0$; for simplicity, the doping density N is set to zero. At distances far away from the contact-channel interfaces, the curvature of the surface potential is approximately zero. As a result, Equation (6.2-2) can be rewritten in the form

$$\frac{1}{e}\left(\Phi_f^0 - \Phi_g + \Phi_{bi}\right) = \lambda^2 \frac{e\rho}{\varepsilon_0 \varepsilon_{nt}} = \frac{Q_{tot}}{C_{ox}}, \tag{6.2-20}$$

where Q_{tot} and C_{ox} are the charge and oxide capacitance per unit area in the single-gate planar case and the charge and capacitance per unit length in the gate-wrapped-around case (cf. Section 2.1). (Note that Equation (6.2-20) points out an alternative way of how to calculate the modified

Poisson equation (Equation (6.2-2)) for different device geometries, such as planar double gate and gate-all-around, for instance: simply insert C_{ox} appropriate for the device under consideration and identify λ^2 from the above expression). Taking the derivative of this equation with respect to Φ_f^0 one gets

$$\delta\Phi_f^0 = \frac{C_{ox}}{C_{ox} + C_q}\delta\Phi_g \tag{6.2-21}$$

with the quantum capacitance C_q given by

$$C_q = e\frac{\partial Q_{tot}}{\partial\Phi_f^0}. \tag{6.2-22}$$

This means that in an FET, the oxide capacitance and the quantum capacitance are connected in series as is shown in Figure 6.2-9, so that the total gate capacitance becomes [4]

$$C_{tot} = e\frac{\partial Q_{tot}}{\partial\Phi_g} = \frac{C_{ox}C_q}{C_{ox} + C_q}. \tag{6.2-23}$$

The following conclusions can be drawn from Equation (6.2-21): as long as $C_q \leq C_{ox}$, a change of gate voltage always results in a change of Φ_f^0. If on the other hand $C_q \gg C_{ox}$, a change of gate voltage yields $\delta\Phi_f^0 \to 0$ meaning that the movement of the surface potential stops as is the case in a conventional MOSFET.

For a long-channel conventional type FET (see Figure 6.2-9(a)) with small applied bias, we can determine the quantum capacitance as follows. Let $D^{3D}(E - \Phi_f^0)$ be the 3D density of states; Φ_f^0 again is the potential far away from the contact interface. Then, neglecting the contact regions, the charge in the channel can be written as

$$Q_{tot} \propto ed_{nt}WL\int dED^{3D}\left(E - \Phi_f^0\right)f(E - E_f^s), \tag{6.2-24}$$

where d_{nt} is the thickness, W the width, and L the length of the channel. According to Equation (6.2-22), the quantum capacitance becomes

[4] Expression (6.2-23) is only valid for a long channel device. In fact, $\delta Q_{tot} = \partial Q_{tot}/\partial V_{gs} \cdot \delta V_{gs} + \partial Q_{tot}/\partial V_{ds} \cdot \delta V_{ds} = C_q C_{ox}/(C_q + C_{ox})\delta V_{gs} + C_d\delta V_{ds}$, where C_d is the drain capacitance. In electrostatically well-behaved devices, C_d is negligible.

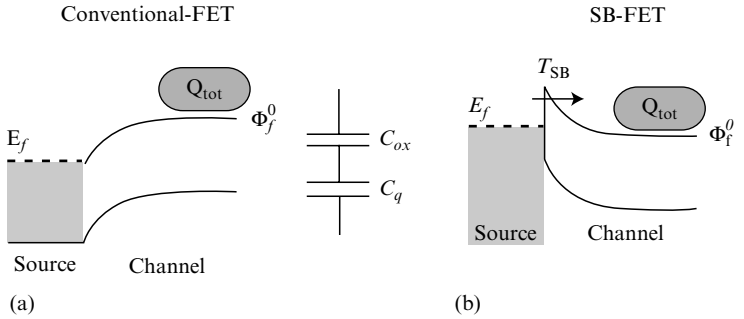


Figure 6.2-9. Total charge in the channel far away from the contact interface. The total gate capacitance is determined by the oxide and quantum capacitance.

$$C_q^{3D} = e\frac{\partial Q_{tot}}{\partial \Phi_f^0} \propto -e^2 d_{nt} WL \int dE D^{3D}(E) \frac{\partial f(E - E_f^s)}{\partial E}$$
$$\approx e^2 d_{nt} WL D^{3D}(E_f^s - \Phi_f^0), \tag{6.2-25}$$

since the derivative of the Fermi function is peaked around E_f^s. This means that in 3D, $C_q^{3D} \propto \sqrt{E_f^s - \Phi_f^0}$ and thus C_q^{3D} increases when the surface potential changes, whereas $C_{ox} = \text{const}$. Therefore, in 3D systems, like MOSFETs, $C_q \gg C_{ox}$ is always valid at large enough gate voltage, [5] and consequently the surface potential hardly moves anymore in the on-state while V_{gs} increases [29]. However, the situation is completely different in a 1D system, such as a CNFET [28]. Here, C_q^{1D} is proportional to the 1D density of states (i.e., $\propto 1/\sqrt{E_f^s - \Phi_f^0}$), such that in principle $C_q^{1D} \leq C_{ox}$ can be achieved in 1D systems. Calculating C_q^{1D} and C_{ox} for 1D, it becomes apparent that rather thin gate oxides and/or high-k dielectrics are needed in order to obtain $C_q^{1D} < C_{ox}$, though [30].

In case of a 1D system with Schottky barriers, such as an SB–CNFET, a more detailed consideration is needed and things can be quite different. If the Schottky barrier height $\Phi_{SB} \gg k_B T$, then tunneling is the dominant process of carrier injection into the channel. In this case, Equation (6.2–24) for the charge in the channel has to be modified to

$$Q_{tot} \propto eL \int dE T_{SB}(E - \Phi_f^0) D^{1D}(E - \Phi_f^0) f(E - E_f^s), \tag{6.2-26}$$

[5] For instance, in a MOSFET with d_{nt} and $d_{ox} = 2\,\text{nm}$, the oxide capacitance is approximately $2 \cdot 10^{-6}\,\text{F/cm}^{-2}$, whereas for $E_f^s - \Phi_f^0 = 0.1\,\text{eV}$ C_q becomes $4 \cdot 10^{-4}\,\text{F/cm}^{-2}$.

where T_{SB} is the tunneling probability for carriers to be injected into the channel through the source/drain Schottky barriers. Consequently, the quantum capacitance becomes

$$C_q^{1D} \approx e^2 L T_{SB}\left(E_f^s - \Phi_f^0\right) D^{1D}\left(E_f^s - \Phi_f^0\right) + O(\partial T_{SB}/\partial \Phi_f^0), \qquad (6.2\text{-}27)$$

where the higher order corrections proportional to $\partial T_{SB}/\partial \Phi_f^0$ (i.e., proportional to the change of the transmission probability through the Schottky barrier with change of surface potential), are neglected here. The presence of the Schottky barrier influences C_q in two ways: first, the particular potential landscape of the Schottky barrier prefers carrier injection at high energies (relative to the conduction band), where the density of states is low, and second, as is known from the analysis above (see Section 3.1), that the potential distribution and therefore the tunneling probability $T_{SB}(E_f^s)$ strongly depends on d_{ox} and can be significantly smaller than 1. Hence C_q in SB-CNFETs will be smaller than in a conventional-type CNFET, and moreover, both, C_{ox} *and* C_q decrease when making d_{ox} larger and vice versa. This means that even for larger d_{ox}, both capacitances can still be of the same order—a critical finding that has not been previously noticed.

The exact value of the ratio between C_q and C_{ox} depends of course also on the Schottky barrier height, which provides an additional degree of freedom independent of the oxide thickness that allows to modify the quantum capacitance. Increasing the Schottky barrier height decreases the transmission probability T_{SB} and hence decreases C_q, while C_{ox} remains unchanged. The relation between C_{ox} and C_q and its dependence on the oxide thickness and Schottky barrier height has important implications for the performance of CNFETs and 1D FET structures in general [6]:

First, if $C_{ox} > C_q$, the total gate capacitance C_{tot} is determined rather by C_q than C_{ox} (cf. Equation (6.2-23)). C_q on the other hand can be lowered by increasing the Schottky barrier height. With the right choice of barrier and very thin gate oxides, it is possible to decrease C_q without a substantial loss of drive current such that the device delay $\tau = C_{tot} V_{dd}/I_d$ improves. This means in 1D FETs, the presence of Schottky barriers can be beneficial for the high frequency performance (details will be published elsewhere).

[6] Most of our statements can be applied to small nanowire structures or molecular devices as well.

Second, since $C_{ox} \geq C_q$ in CNFETs with Schottky barriers, the surface potential will respond to a change of V_{gs} even well in the on-state (cf. Equation (6.2-21)) such that higher 1D subbands or modes—hundreds of meV separated from the first conduction/valence band—can participate in the current [7]. The appearance of this multimode transport in CNFETs, its observability, and the impact on the performance of CNFETs will be explored in the following section.

3.2.2. Appearance of multimode transport

Before we look into the issue of multimode transport more closely by using simulations, the question arises, how one can tell whether or not multiple modes participate in the current when looking at the $I_d - V_{gs}$ characteristics? The answer is that very often one cannot! To be able to observe multimode transport, it is necessary that current in one mode saturates with increasing gate voltage before a subsequent mode significantly contributes to the current. In this case, a step-like increase of I_d follows. Figure 6.2-10 shows this scenario: in (a), the subband separation is sufficiently large so that the contributions of different modes can be distinguished; (b) shows the transconductance contributions of the first three modes. If their overlap is small (i.e., if subsequent modes are energetically well separated from each other as in the present example), multimode transport is observable and manifests itself in the step-like increase of I_d. However, if the subband separation is small so that the transconductance contributions overlap (see Figure 6.2-10(d)), the steps in I_d vanish as illustrated in Figure 6.2-10(c) and the $I_d - V_{gs}$ curves look like if transport was due to a single mode only. In this case multimode transport occurs but cannot be observed. We discuss later that this is exactly what happens in Schottky barrier CNFETs.

In higher dimensional systems, such as conventional MOSFETs, the subbands are very close to each other so that a saturation ideally never occurs. The saturation in the $I_d - V_{gs}$ curves of CNFETs is thus a direct consequence of the one-dimensionality of nanotubes, where the current through one subband saturates when it reaches the current level associated with the quantized conductance $4e^2/h$.

In order to further investigate under what circumstances multimode transport appears, we make use of the approximation for the current I_d at small bias (Equation (6.2-18)). We will first consider the case of a ballistic

[7] Note that as more and more subbands contribute, C_q will eventually dominate and the movement of bands will slow down and even stop. This, however, crucially depends on the energetic separation of the subbands.

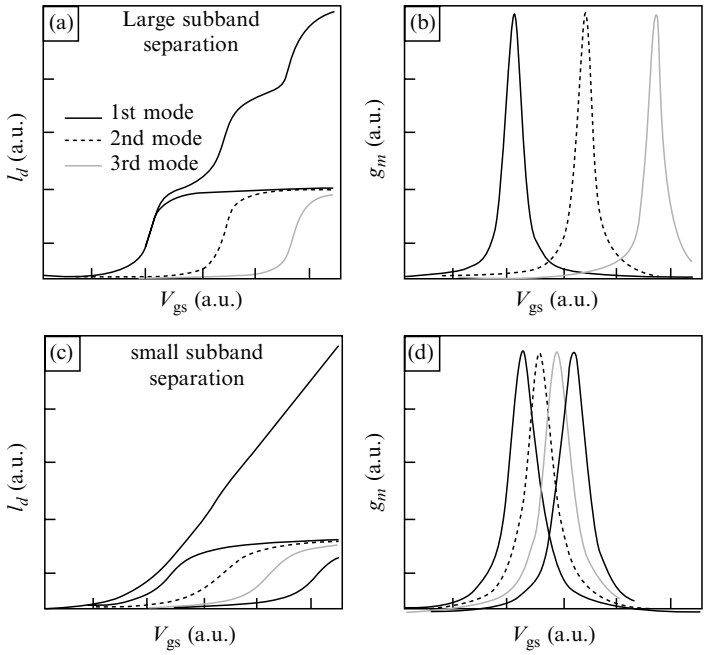


Figure 6.2-10. $I_d - V_{gs}$ curves in the case of multimode transport: (a) subbands have large energetic separation leading to a step-like increase of I_d, (b) transconductance contributions of the first three modes, (c) closely spaced subbands lead to a linear increase of current above V_{th}, and (d) shows the transconductance contributions in case of a small energetic separation between subbands.

CNFET with Schottky barriers at the contact interfaces. The role of scattering in the channel will be illuminated later.

3.2.3. Ballistic transport

Let T^i_{SB} be the transmission probability through the source and drain Schottky barriers for the mode i. Then the total transmission probability T in Equation (6.2-18) can be written as $T^i = T^i_{SB}/(2 - T^i_{SB})$ (see e.g., Ref. [11], p. 64) and Equation (6.2-18) becomes

$$I_d^{1D} = \frac{4e^2}{h} V_{ds} \sum_i \frac{T^i_{SB}}{2 - T^i_{SB}}. \tag{6.2-28}$$

For simplicity, we summed transmission probabilities, which is equivalent to an incoherent combination of the two Schottky barriers at source and drain. In Equation (6.2-28), the current contribution of each mode saturates when the transmission through the Schottky barrier at the Fermi energy $T^i_{SB}(E_f)$ approaches unity and the current level corresponds to the quantized conductance $g = 4e^2/h$.

In general, to observe multimode transport in ballistic 1D structures, both V_{ds} and k_BT have to be smaller than the subband separation. In SB-CNFETs, however, this is by no means a guarantee that multimode transport is observable. On the contrary, as will become clear below, it is very unlikely in SB-CNFETs that multimode can be *observed* although it is likely to *occur*. The presence of the Schottky barriers, more precisely the particular dependence of the tunneling probability through the barrier on the gate potential is what determines I_d and prohibits the observation of multimode transport. To see this we calculate the transconductance g_m

$$g_m = \frac{4e^2}{h} V_{ds} \frac{\partial \Phi_f^0}{\partial \Phi_g} \cdot e \frac{\partial}{\partial \Phi_f^0} \sum_i \frac{T^i_{SB}}{2 - T^i_{SB}}, \tag{6.2-29}$$

where Φ_f^0 and Φ_g are the surface potential far away from the contact interfaces and the gate potential, respectively. Using expression (Equation 6.2-21), this can be rewritten in the form

$$g_m = \frac{4e^2}{h} V_{ds} \cdot \underbrace{\frac{C_{ox}}{C_{ox} + C_q}}_{I} \cdot \sum \underbrace{\frac{2e}{(2 - T^i_{SB})^2}}_{II} \frac{\partial T^i_{SB}}{\partial \Phi_f^0}. \tag{6.2-30}$$

As was discussed above (and will be supported by simulations shown below), if a significant Schottky barrier is present at the contact interfaces, it is easy to obtain $C_{ox} \geq C_q$ in a CNFET, and therefore the term I in Equation (6.2-30) varies only between $\sim 1/2$ and 1. The term II is also only weakly dependent on Φ_f^0, since T^i_{SB} is a number between 0 and 1, and therefore this term gives rise to a factor between 1/2 and 2. Hence, g_m is dominated by $\partial T^i_{SB}/\partial \Phi_f^0$. A closed expression for g_m can be found if the Schottky barrier is approximated by a triangular shaped potential leading to the Fowler–Nordheim approximation for T^i_{SB} [19].

In order to be able to distinguish contributions of different subbands, the peaks of the different g_m contributions have to be separated enough as was mentioned earlier. This separation increases if the energetic separation of the subbands is made larger (i.e., by decreasing the nanotube diameter).

On the other hand, the effective Schottky barrier height for this particular subband increases simultaneously at the same amount as the subbands get separated and therefore the peak transconductance of this subband is significantly lowered. However, since for very thin oxides together with the ultrathin body of a nanotube T_{SB} can be highly transparent, one might expect that in this case multimode transport should become observable. Figure 6.2-11 shows the contributions of the first three subbands to the transconductance for $d_{nt} = 1\,\text{nm}$, $d_{ox} = 2\,\text{nm}$ and a Schottky barrier height of $\Phi_{SB} = 0.3\,\text{eV}$ and an energetic subband separation of 0.2 eV calculated using the Fowler–Nordheim approximation mentioned above. It can be seen that even for such a thin gate oxide and tube diameter, a subband separation of 0.2 eV is too large as to be able to observe multimode transport, because the g_m peaks of higher subbands do not exceed the contribution of the first subband. The reason for this is twofold: (1) at first, even for very thin gate oxides and ultrathin bodies, the transmission probability T_{SB} increases quickly only over a certain gate voltage range, but then as it gets closer to unity the change of T_{SB} with altering V_{gs} slows down considerably. This fact can be understood if one considers the Schottky barrier as having a triangular shape—as was done in the Fowler–Nordheim approximation above—with the tip of the triangle (at the top of the actual barrier) having an angle φ. If the angle is large (close to 90°), it changes quickly when altering V_{gs} and consequently, T_{SB} rapidly changes. On the other hand, making φ go to zero (then $T_{SB} \rightarrow 1$), requires very large gate voltages with the change in φ, and hence in T_{SB} becoming

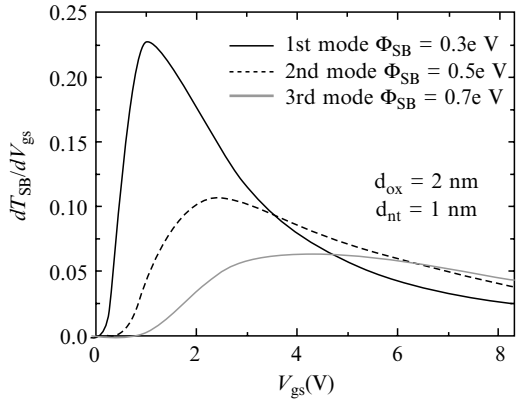


Figure 6.2-11. Transconductance contributions of the first three subands in a ballistic SB-CNFET with $\Phi_{SB} = 0.3\,\text{eV}$, $d_{nt} = 1\,\text{nm}$, and $d_{ox} = 2\,\text{nm}$; the subband separation is 0.2 eV.

very small. As a result, for large gate voltages beyond the peak transconductance g_m drops rather slowly as is shown in Figure 6.2-11. Secondly, the peaks of g_m get smaller for higher subbands since the Schottky barrier is larger and therefore the maximum change of current with gate voltage i.e., the peak transconductance is lower. In essence, the gate voltage range over which the transmission probability through the Schottky barrier is made unity (in this case I_d saturates) is so large that it overlaps with the gate voltage, where subsequent subbands contribute most to the transconductance. Therefore, multimode transport is usually not *observed* in SB-CNFETs. However, for a significant Schottky barrier (i.e., $\Phi_{SB} \gg k_B T$ as frequently encountered in experiments), one can always expect multimode transport to occur. The magnitude of the contributions of higher subbands crucially depends on the subband separation as will become clear below.

The reason why multimode transport always occurs was already mentioned above, but is worth rephrasing here. Due to the presence of the Schottky barrier, C_q is on the order of or smaller than C_{ox} over a large range of gate oxide thicknesses. This is expected for thin gate oxides, where C_{ox} becomes large but is also true for large gate oxides, since the carrier injection into the channel is strongly reduced for large d_{ox} and thus C_q decreases (cf. Equation (6.2-27)) accordingly. Therefore, even well in the on-state, the bands will respond to the gate potential and thus higher subbands play a role for current transport. This is an important finding since although multimode transport is hardly observable in SB-CNFETs it does occur and contributes to the current, particularly in larger diameter nanotubes. As a result, the current in experimental devices is often attributed to a single mode alone and hence can lead to a misinterpretation of effects related to the metal-nanotube contact. In particular, Schottky barrier heights are likely to be underestimated due to an increased current level. Next, we will concentrate on a quantitative analysis using simulations regarding the appearance and the magnitude of multimode transport in CNFETs. We have included in our model higher subbands by adding additional pairs of conduction and valence bands (with larger E_g) and doing the same calculations as in the single-band case. In order to keep the numerical burden as small as possible, we have restricted the computations to three subbands. m^*_{con} in the different subbands are all assumed equal for simplicity. The total charge contributed by all three subbands is added and used for the iterative solution of the Poisson equation. Finally, the current carried by all subbands is computed using the Fisher–Lee relation, Equation (6.2-10).

We have simulated the $I_d - V_{gs}$ curves for a long channel CNFET with two different oxide thicknesses of $d_{ox} = 2\,\text{nm}$ and $d_{ox} = 20\,\text{nm}$. All other

parameters were kept the same for both devices. To be specific, we considered devices with a Schottky barrier height of $\Phi_{SB} = 0.3\,\text{eV}$ and $E_g = 0.6\,\text{eV}$, and a subband separation of 0.05 eV. [8]

Figure 6.2-12(a) shows the transfer characteristic for the device with $d_{ox} = 2\,\text{nm}$. The black straight line refers to the total drain current I_d, the light gray dotted, the gray dashed, and dark gray straight lines denoted with (i) to (iii), refer to the current contributions of each subband. While the curves do not appear perfectly smooth due to the discrete gate voltage step size, there is clearly no sign of multimode transport in the total device current although higher subbands (mainly the second) substantially ($\sim$ 20%) contribute according to our simulations.

The inset in Figure 6.2-12(a) shows how the surface potential of the first subband changes with gate voltage. For small gate voltages, the slope of this curve is -1 (gray dashed line), because the device is in its off-state and a change in V_{gs} results in the same change of surface potential. For increasing gate voltages, the slope decreases but does not go to zero. The gray area in the inset illustrates the region of possible $\Phi_f^0(V_{gs})$ values, where the black dashed line describes the limit $C_q \gg C_{ox}$ for a conventional MOSFET. The gray dashed line on the other hand illustrates the quantum capacitance limit (i.e., $C_q \ll C_{ox}$), and the slope remains -1 even well in the on-state. It is apparent that the device considered here is rather in the quantum capacitance limit, which was expected since higher subbands contribute significantly to the total drain current. We have discussed above that for a Schottky barrier CNFET, a device with large d_{ox} can still be in the quantum capacitance limit, because the carrier injection into the channel is affected by a larger d_{ox}, thus decreasing C_q. This behavior can be seen in Figure 6.2-12(b) in case of the device with $d_{ox} = 20\,\text{nm}$: The inset shows the same qualitative dependence of Φ_f^0 on V_{gs} as was observed for the device with the thin oxide. Consequently, we expect that higher subbands contribute to the current, which is indeed the case: they contribute about 14% to the total current.

3.2.4. On the role of scattering

As we have seen in the previous paragraph, multimode transport does occur in CNFETs based upon nanotubes with larger diameter (i.e., small subband separation), but can hardly be observed in the electrical characteristics. The transconductance contributions of subsequent modes have to be separated enough from each other and the peak transconductance has

[8] This set of parameters does not correspond to a particular nanotube diameter, but is chosen to illustrate the critical dependencies.

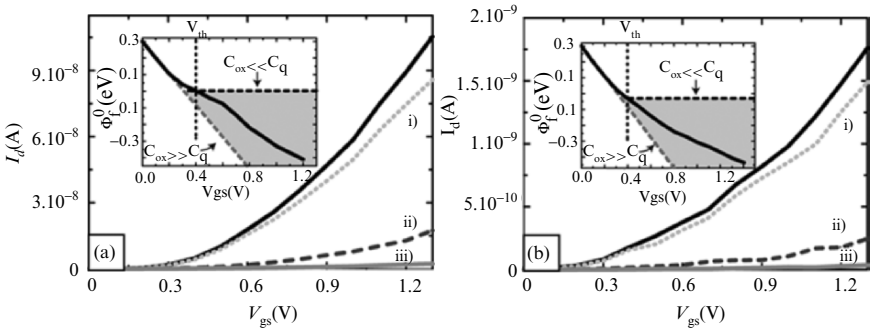


Figure 6.2-12. Transfer characteristics at $V_{ds} = 0.03\,\mathrm{V}$ for devices with (a) $d_{ox} = 2\,\mathrm{nm}$ and (b) $d_{ox} = 20\,\mathrm{nm}$ and otherwise equal parameters (see main text). Contributions of the first three subbands are denoted by (i), (ii), and (iii). The inset shows Φ_f^0 versus V_{gs}.

to be large enough in order to observe a step-like increase of I_d, conditions that are generally not met in SB-CNFETs. For a large subband separation, the peak transconductances of higher modes are always less than the contribution of the first mode, and for small subband separation, the gate voltage range needed for one mode to saturate is too large so that the transconductance contributions cannot be resolved. A way to achieve current saturation in a much smaller gate voltage range is to deliberately introduce strong scattering in the nanotube channel, thereby limiting the current carrying capability of each subband. Such scattering can be realized by heavily doping the nanotube with potassium, for instance [31, 32]. The reason why scattering changes the situation and makes multimode transport observable becomes apparent, if scattering is included in the analysis above. In this case the total transmission through the CNFET is determined by the transmission through the source and drain Schottky barriers T_{SB} and by the transmission through the channel $T_{scat} \sim 1/(1 + L/l_{scat})$ with l_{scat} being the mean free path for scattering. Hence, the overall transmission function for subband i is [11]

$$T^i = \frac{T^i_{SB} T_{scat}}{2T_{scat} - 2T^i_{SB} T_{scat} + T^i_{SB}}. \tag{6.2-31}$$

Note, that the expression for the transmission function only applies for calculating the current. Scattering has only little effect on the quantum capacitance and hence in Equation (6.2-27), the transmission function is only determined by the carrier injection into the channel (i.e., T is determined by the Schottky barriers at the contact channel interfaces alone).

For simplicity, we assume a gate voltage independent T_{scat}, which in addition is the same in each subband. Inserting this in Equation (6.2-18) and calculating the transconductance yields

$$g_m = \frac{4e^2}{h} V_{ds} \cdot \frac{C_{ox}}{C_{ox} + C_q} \cdot \sum \underbrace{\frac{2}{(2 - 2T^i_{SB} + T^i_{SB}/T_{scat})^2}}_{II} \frac{\partial T^i_{SB}}{\partial \Phi^0_f} \qquad (6.2\text{-}32)$$

In Equation (6.2-32), the second term $C_{ox}/(C_{ox} + C_q)$ again is only weakly dependent on Φ^0_f and varies between 1/2 and 1. But in contrast to Equation (6.2-30), the term *II* now plays an important role because for strong scattering (i.e., $l_{scat}/L \to 0$), it tends to go to zero as $T^i_{SB} \to 1$. As a result, the transconductance is limited to a much smaller Φ_g range as compared to the ballistic case. Figure 6.2-13 shows the transconductance contribution of the first three subbands as discussed earlier for the ballistic case; but here scattering is taken into account and l_{scat}/L chosen to be approximately 1/50. In this case, the three contributions of the different subbands can be distinguished since the peaks of g_m are separated from each other and the transconductance level of each subband is large enough. As long as the scattering is not that large as to destroy the appearance of distinct modes, increasing the scattering leads to a better resolution of the different contributions.

It is worth rephrasing the present situation: in the case of SB-CNFETs, scattering is needed in order to be able to distinguish between the different modes in the electrical characteristics, since scattering diminishes the

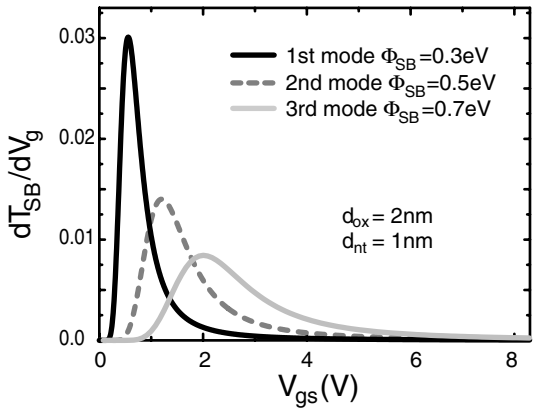


Figure 6.2-13. Transconductance contributions of the first three subbands with scattering l_{scat}/L is chosen to be 1/50.

impact of the Schottky barriers on the transfer characteristics. This is rather different if compared to other low dimensional structures, like quantum point contacts, for instance. Here, scattering destroys the appearance of multimode transport. The reason for this is the rather small energetic separation of different subbands that is on the order of 1–10 meV [33]. In carbon nanotubes on the other hand, the subband separation can be 200–300 meV, and hence very strong scattering is needed to destroy the appearance of different modes.

In a recent study, we have verified the above discussion about the impact of scattering on the observability of multimode transport, both, experimentally as well as with simulations [31]. Using potassium doping of the nanotube channel [32], we introduced strong scattering in the CNFET and measured the $I_d - V_{gs}$ curves for moderate bias (for details Ref. [31]). Figure 6.2-14(a) shows the result of this experiment. One can clearly see a step in the characteristics at around $V_{gs} \approx -1.1$ V. On a larger gate voltage range as shown in the inset of Figure 6.2-14(a), it is possible to identify several steps (marked by arrows) that we attribute to the onset of higher subbands. In addition, we simulated the $I_d - V_{gs}$ characteristics of the experimental devices using the model presented in Section 2 with Buettiker probes taking care of scattering in the nanotube channel. A Schottky barrier $\Phi_{SB} = 0.3$ eV and a subband spacing between the first and second subband of 300 meV were assumed in agreement with predictions based upon tight binding calculation for the nanotubes used in the experiment. The ratio L/l_{scat} was set to 30 and all other parameters for the simulations were chosen in accordance to the experimental device under consideration (i.e., $d_{nt} = 1.4$ nm, $d_{ox} = 5$ nm, and $E_g = 0.6$ eV); all effective masses were taken to be $0.1\,m_0$. The result of the simulation is shown in Figure 6.2-14(b) for two different V_{ds}. Although the current level of the simulated curve is higher, since we did not try to find the ratio L/l_{scat} that fits best to the experiment, the qualitative agreement between theory and experiment is apparent, reinforcing our physical interpretation of the role of scattering and the general picture of multimode transport in CNFETs.

4. CONCLUSIONS

Carbon nanotube field effect transistors represent a relatively new class of field-effect transistor devices exhibiting very promising electrical characteristics. At first sight, one may interpret CNFETs in terms of conventional bulk-like MOSFETs. However, a closer look reveals new, interesting aspects, and a distinctly different behavior of CNFET devices.

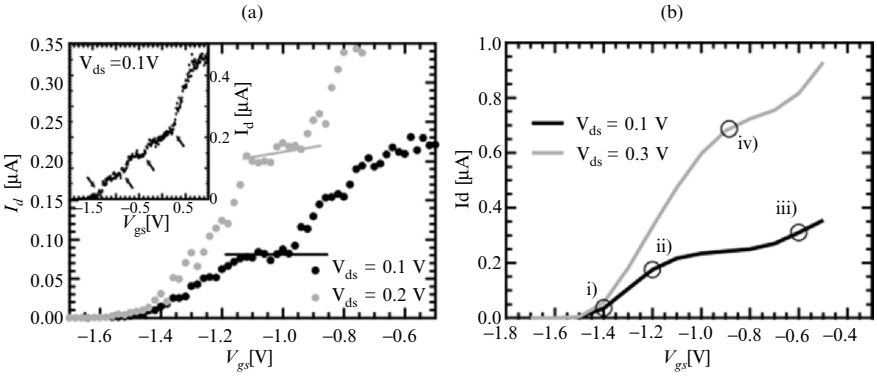


Figure 6.2-14. (a) Experimental $I_d - V_{gs}$ curves for a heavily *K*-doped CNFET. The inset shows a larger V_{gs} range; arrows mark the onset of higher subbands. (b) Simulations using the model with Buettiker probes to account for scattering. (Reprinted figure with permission from J. Appenzeller et al., *Phys. Rev. Lett.*, **92,** 226802, 2004.)

In this article, we have addressed some of these aspects that are related to the unique "geometrical" and "electrical smallness" of carbon nanotubes.

The small diameter and the cylindrical geometry of carbon nanotubes make them ideal objects for ultimately scaled devices. In combination with thin gate oxides, a very tight gate control can be achieved effectively suppressing short-channel effects even for very small channel lengths. It is this gate control that allows for excellent electrical characteristics of CNFETs despite the presence of Schottky barriers at the contact interfaces. In effect, steep inverse subthreshold slopes and excellent on-states surpassing those of conventional silicon MOSFETs can be realized.

The one-dimensionality of carbon nanotubes has a strong impact on the electrical response of CNFETs to an applied gate voltage. In particular the interplay between 1D density of states and the presence of Schottky barriers yields a strongly reduced quantum capacitance if compared to higher dimensional systems. As a result, C_q is on the order of, or smaller than the oxide capacitance C_{ox} over a large range of oxide thicknesses. Therefore, the gate potential is able to manipulate the surface potential and thus multimode transport always occurs in SB-CNFETs, particularly in devices based upon nanotubes with larger diameters, where the contribution of higher subbands is substantial. Although the presence of the Schottky barriers is responsible for the appearance of multimode transport, they prohibit at the same time its observability. However, we have shown that by introducing strong scattering in the channel of a CNFET, multimode transport becomes observable proving that multimode transport indeed occurs and has to be taken into account in a quantitative analysis.

Concluding, note that the present analysis is not restricted to CNFETs. CNFETs were investigated since they represent to date the best and most prominent example of 1D FETs. Therefore, our findings are rather generally valid and also apply to other 1D FET structures exhibiting similar "geometrical" and "electrical smallness."

REFERENCES

[1] Zhou, Ch., Kong, J. and Dai, H., 2000, Electrical measurements of individual semiconducting single-walled carbon nanotubes of various diameters, *Appl. Phys. Lett*., **76,** 1597–1599.
[2] Tans, S. J. and Dekker, C., 2000, Potential modulations along carbon nanotubes, *Nature*, **404,** 834–835.
[3] Appenzeller, J., Knoch, J., Martel, R., Derycke, V., Wind, S. and Avouris, Ph., 2002, Short-channel like effects in Schottky barrier carbon nanotube field-effect transistors, *Internat. Electron Dev. Meeting 2002, Technical Digest*, 285–288.
[4] Appenzeller, J., Knoch, J., Derycke, V., Wind, S. and Avouris, Ph., 2002, Field-modulated carrier transport in carbon nanotube transistors, *Phys. Rev. Lett*., **89,** 126801.
[5] Heinze, S., Tersoff, J., Martel, R., Derycke, V., Appenzeller, J. and Avouris, Ph., 2002, Carbon nanotubes as Schottky barrier transistors, *Phys. Rev. Lett*., **89,** 106801.
[6] Knoch, J., Mantl, S., Lin, Y. -M., Chen, Z., Avouris, Ph. and Appenzeller, J., 2004, An extended model for carbon nanotube field-effect transistors, *Device Research Conf. 2004, Conference Digest*, pp. 135–136.
[7] Auth, Ch. and Plummer, J. D., Scaling theory for cylindrical, fully depleted, surrounding gate MOSFET's, *IEEE Electron Dev. Lett*., **18,** 74–76.
[8] Pikus, F. G. and Likharev, K. K., 1997, Nanoscale field-effect transistors: an ultimative size analysis, *Appl. Phys. Lett*., **71,** 3661–3663.
[9] Yan, R. -H., Ourmazd, A. and Lee, K. F., 1992, Scaling the Si MOSFET: From bulk to SOI to bulk, *IEEE Trans. Electron Dev*., **39,** 1704–1710.
[10] Young, K. K., 1989, Short-channel effect in fully-depleted SOI MOSFET's, *IEEE Trans. Electron Dev*., **36,** 399–402.
[11] Datta, S., 1995, *Electronic Transport in Mesoscopic Systems*, Cambridge University Press.
[12] Datta, S., 2000, Nanoscale device modeling: The Green's function method, *Superlatt. Microstructures*, **28,** 253–278.
[13] Flietner, H., 1972, E(k) relation for 2-band scheme of semiconductors and application to metal-semiconductor contact, *Phys. Stat. Solidi*, **54,** 201–208.
[14] Hatta, E., Nagao, J. and Mukasa, K., 1996, Tunneling through a narrow-gap semiconductor with different conduction and valence band effective masses, *J. Appl. Phys*., **79,** 1511–1514.

[15] Knoch, J., Lengeler, B. and Appenzeller, J., 2002, Quantum simulations of an ultrashort channel single-gated n-MOSFET on SOI, *IEEE Trans. Electron Dev.*, **49,** 1212–1218.
[16] Venugopal, R., Paulsson, M., Goasguen, S., Datta, S. and Lundstrom, M. S., 2003, A simple quantum mechanical treatment of scattering in nanoscale transistors, *J. Appl. Phys.*, **93,** 5613–5625.
[17] Lake, R., Klimeck, G., Bowen, R. Ch. and Jovanovic, D., 1997, Single and multiband modeling of quantum electron transport through layered semiconductor devices, *J. Appl. Phys.*, **81,** 7845–7869.
[18] Fisher, D. S. and Lee, P. A., 1981, Relation between conductivity and transmission matrix, *Phys. Rev. B*, **23,** 6851–6854.
[19] Sze, S. M., 1981, *Physics of Semiconductor Device*, John Wiley & Sons, Inc.
[20] Landau, L. D. and Lifshitz, E. M., 1977, *Quantum Mechanics*, Pergamon Press, Oxford.
[21] Heinze, S., Radosavljevic, M., Tersoff, J. and Avouris, Ph., 2003, Unexpected scaling of performance of carbon nanotube Schottky-barrier transistors, *Phys. Rev. B*, **68,** 235418.
[22] Knoch, J. and Appenzeller, J., 2002, Impact of the channel thickness on the performance of Schottky barrier metal-oxide-semiconductor field-effect transistors, *Appl. Phys. Lett.*, **81,** 3082–3084.
[23] Lundstrom, M., 2000, *Fundamentals of Carrier Transport*, Cambridge University Press.
[24] Lin, Y. -M., Appenzeller, J. and Avouris, Ph., 2004, Novel structures enabling bulk switching in carbon nanotube FETs, *Device Research Conf., Conference Digest*, 133–134.
[25] Appenzeller, J., Radosavljevic, M., Knoch, J. and Avouris, Ph., 2004, Tunneling versus thermionic emission in one-dimensional semiconductors, *Phys. Rev. Lett.*, **92,** 048301.
[26] Luryi, S., 1988, Quantum capacitance devices, *Appl. Phys. Lett.*, **52,** 501–503.
[27] John, D. L., Castro, L. C. and Pulfrey, D. L., 2004, Quantum capacitance in nanoscale device modeling, *J. Appl. Phys.*, **96,** 5180–5184.
[28] Rahman, A., Guo, J., Datta, S. and Lundstrom, M. S., 2003, Theory of ballistic nanotransistors, *IEEE Trans. Electron Dev.*, **50,** 1853–1864.
[29] Taur, Y. and Ning, T. H., 1998, *Fundamentals of Modern VLSI Devices*, Cambridge University Press.
[30] Guo, J., Datta, S. and Lundstrom, M. S., 2002, Assessment of silicon MOS and carbon nanotube FET performance limits using a general theory of ballistic transistors, *Internat. Electron Dev. Meeting 2002, Technical Digest*, 711–714.
[31] Appenzeller, J., Knoch, J., Radosavljevic, M. and Avouris, Ph., 2004, Multimode transport in Schottky barrier carbon nanotube field-effect transistors, *Phys. Rev. Lett.*, **92,** 226802.
[32] Radosavljevic, M., Appenzeller, J., Avouris, Ph. and Knoch, J., 2004, High performance of potassium n-doped carbon nanotube field-effect transistors, *Appl. Phys. Lett.*, **84,** 3693–3695.

[33] van Wees, B. J., Kouvenhoven, L. P., Willems, E. M. M., Harmans, C. J. P. M., Mooji, J. E., van Houten, H., Beenakker, C. W. J., Williamson, J. G. and Foxon, C. T., 1991, Quantum ballistic and adiabatic electron transport studied with quantum point contacts, *Phys. Rev. B*, **43,** 12431–12453.

Chapter 6.3

HARDWARE FOR AMBIENT SOUND REPRODUCTION

Ronald M. Aarts
Philips Research Eindhoven
ronald.m.aarts@philips.com

Abstract Today's and tomorrow's audio and video applications put increasing demands on sound reproduction techniques, particularly because of the advent of ambient intelligence (AmI). A good sound reproduction system is generally in conflict with the boundary conditions posed by AmI, both by size as well as by setup flexibility. Hence, improving the sound quality within these conditions is important, because the traditional means have difficulties with these constraints. Various new and old means for sound reproduction are discussed as possible candidates, including "singing display," and "Bary-Bass": the former uses a display as a sound generator; the latter a system, which maps the low frequency region (20–120 Hz) onto a single tone, and uses an extremely efficient transducer at that particular tone. Apart from the transducers, various other options are discussed to relax the boundary conditions of traditional sound reproduction setups required by AmI.

Keywords barybass; driver; force factor; headphones; incredible surround; loudspeaker; phantom source; sound reproduction; ultrabass

1. INTRODUCTION

Before and after the "birth" of the classical electrodynamic loudspeaker in 1925, various other concepts appeared [1–6], and some of them have left the scene, to mention a few:

- laser loudspeakers using the photo acoustic effect [7];
- loudspeaker arrays, consisting of various drivers;

S. Mukherjee et al. (eds.), AmIware, 403–419.

- "audio spotlight" using interfering ultrasonic sound beams [8–9];
- "flame loudspeaker" and "Ionophone," using pyroacoustic transduction;
- vibrating panels;
- (digital) sound projector (see Figure 6.3-5);
- headphones;
- neck-sets (see Figure 6.3-7);
- electromagnetic loudspeakers [1];
- piezo loudspeakers [1];
- electrostatic loudspeakers [1];
- vibrating (LC) Displays ("Singing Display," based on electrostatic forces) [6]; and
- BaryBass, a resonant loudspeaker (see Figure 6.3-10–6.3-12) [5].

Some of these systems will be discussed later, for the others, one is referred to the bibliography section. For ordinary living room applications, classic sound reproduction by ordinary loudspeakers will do for most of the time, however, for ambient audio, we might need some of the above-mentioned alternatives. The reasons can be due to size, privacy, but also whether it is to be produced locally, or it is sound for everybody, or perhaps even sound, which follows you in any room of the house.

In the following we will show some special loudspeaker systems, and then we will present various techniques to overcome several problems with traditional sound reproduction.

1.1. "Flat-Pack"

SoundpaX loudspeakers (from NXT) are "flat-pack," corrugated cardboard loudspeakers. They are very light and easy to foldaway. An example is shown in Figure 6.3-1.

Figure 6.3-1. SoundpaX loudspeakers are "flat-pack," corrugated cardboard loudspeakers.

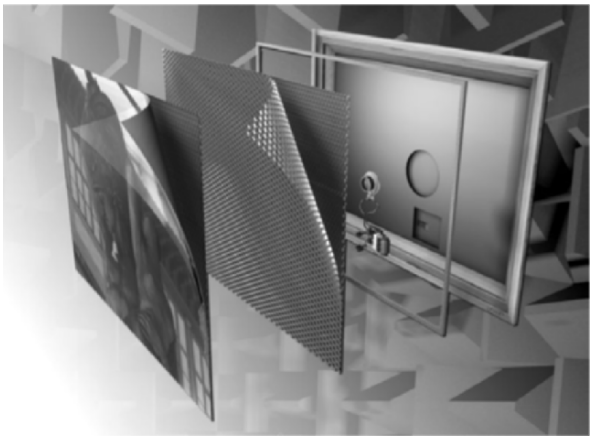


Figure 6.3-2. New generation speaker which can blend invisibly into your room, as the detachable frames allow you to insert your favorite prints.

1.2. Picture Frame

Another example is given in Figure 6.3-2; this new generation speaker by NXT's technology may be successfully applied to a wide variety of applications: multimedia, plasma TVs, home stereos, architectural acoustics, and consumer electronics. They are lightweight and flexible speakers that can reproduce high- to midrange frequencies. The panels blend invisibly into your room, as the detachable frames allow you to insert your favorite prints.

1.3. "Singing" Display

Singing Display [6] has as aim to generate sound from the display itself to save component costs and miniaturize audio-visual products.

1.3.1. Background

As displays become more pervasive, many products are becoming audio-visual. A case in point is the GSM telephone, where the display has become indispensable. Indeed, the GSM display is becoming so large to allow for games, Internet, video, etc., whilst the phone itself is becoming so small, that there is little room to accommodate the loudspeaker or microphone, which is still required for the primary GSM function (i.e., making a phone call).

1.3.2. "Singing" display

It is proposed to avoid this issue by making use of the display itself to produce (loudspeaker) or detect (microphone) the audio signals. GSM

products are shown in Figure 6.3-3. In the GSM in the middle, the "singing display" has taken over the role of the loudspeaker, whilst on the right hand side, the super "singing display" has taken over the role of both loudspeaker and microphone and, optionally, the keyboard.

It is clear that this approach not only saves space and weight in the product, but also reduces the component count and could hence make the product cheaper. Whilst the idea is illustrated with a GSM embodiment, its scope of application is all possible audio-visual products due to the cost savings, particularly for portable applications (where space/weight saving becomes essential). The most common display used in portable applications is the LCD. LCDs come in many modes (TN, STN, MVA, etc.), and types (passive or active matrix), but have a common feature that they comprise a thin layer of electrooptic material sandwiched between two substrates and are driven using an electric field.

1.3.3. Layout

The proposal is to exploit the specific geometry of the LCD to induce an acoustic output. The geometry is shown schematically in Figure 6.3-4.

The LCD is driven by applying an (AC) voltage to the electrodes, which are either transparent (ITO) or reflective (Al). The observation is that under certain conditions, it is possible to use the applied voltage to cause the LCD to vibrate and create an acoustic output. The vibration is

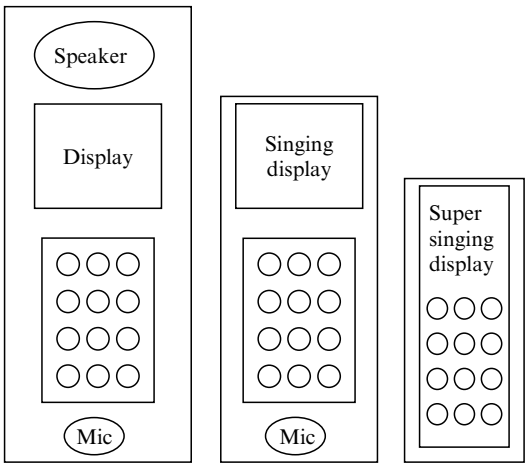


Figure 6.3-3. Traditional GSM telephone: GSM with "Singing Display" (no loudspeaker) and with super "Singing Display" (no loudspeaker, keyboard, or microphone).

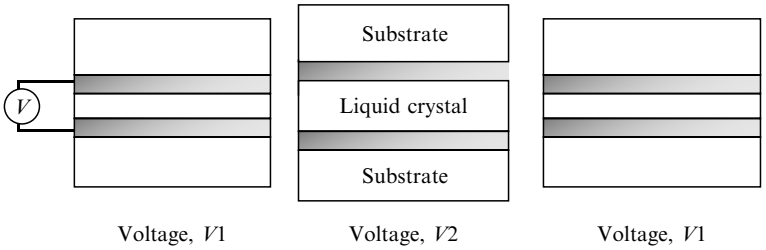


Figure 6.3-4. An LCD consists of two substrates with electrodes (grey area). The liquid crystal is situated between the substrates. Application of a variable voltage causes the cell to vibrate.

caused by electrostatic forces across the liquid crystal layer. The frequency and magnitude of the output is tunable both in frequency and amplitude, depending upon details of the applied voltages. By correct application of the applied voltages, it will be possible to use the display as a loudspeaker. The singing display concept has been proven, however, the acoustic output of the singing display is currently too low to get a sufficient sound pressure level.

1.4. Sound projector

1Limited's Digital Sound Projector is a single slim panel that connects directly to a DVD or CD player. By producing tight, focusable beams of sound, the sound projector beams the separate sound channels around the listener's room. By reflecting off walls and other surfaces in the room, these beams finally come to the listener from left and right, front and rear; see Figure 6.3-5. This single unit replaces a more conventional five loudspeaker setup for surround sound reproduction.

2. WEARABLE AUDIO MODULES

In September 2000, Philips and Levi's launched wearable electronics products. The product range is branded industrial clothing design (ICD+), and consists of four different jackets. Each of the four styles contains a simple body area network using wires integrated into the jacket design. This network allows the synchronous control of the Philips Xenium GSM mobile and Philips Rush MP3 player through the use of a unified remote control. A multidisciplinary team of textile designers, electronic engineers, and product designers have been working together on wearable electronics at Philips Research in Redhill, UK. An example is shown in Figure 6.3-6.

Figure 6.3-5. 1Limited's Digital Sound Projector is a single slim panel producing tight, focusable beams of sound, beaming the separate sound channels around the listener's room.

Another example is by Infineon, which has developed a prototype audio module for the integration in clothes. In addition to ensuring the functionality—for example, as an MP3 player—special attention was paid to a robust and textile-ready design. The components are designed so that the electronics and the interconnections between the textile structures do not interfere with a comfortable wear, allow for easy and convenient use, and allow the clothing to be washed without the need to remove the electronics. A flat keyboard is built with metallized films on an electrically conductive fabric strip. The metal films are attached with an adhesive that is commonly used in the clothing industry. A tiny sensor module is connected to the metal films and registers when the pads are "pressed." The earplug microphone set is also connected to the audio module through the fabric strip.

3. MULTICHANNEL AUDIO

The presence of digital versatile disk (DVD) and super audio CD (SACD) has made multichannel audio popular in sound systems for consumer use today. Here, a method is presented, which converts

Figure 6.3-6 An example of Philips' and Levi's wearable audio devices.

two-channel stereo to multichannel sound reproduction using a three-dimensional (3D) representation [10] (hereafter referred to as ''space mapping''). Although many have introduced multichannel sound systems with a large number of channels, we restrict ourselves to a home cinema setup, for which investigations have shown that five channels are sufficient. This setting is adopted from multichannel configuration with three loudspeakers placed in front of the listener, and the other two at the back. By use of principal component analysis, we developed an algorithm that produces a vector, which indicates the direction of both dominant signal and remaining signal. These two signals are then used as basis signals in the matrix decoding. It offers two improvements above existing multichannel techniques. Firstly, a problem associated with channel cross talk is reduced, and therefore better sound localization is achieved. The latter gives more space to the listener to enjoy the offered program rather than the restricting listening area referred to as the ''sweet spot.'' Secondly, a better sound distribution to the surround channels is achieved by using a cross-correlation technique, while maintaining energy preservation. So, it remains backward and forward compatible with ordinary stereo.

3.1. The Center Channel

We consider the three-channel approach in particular. It is known that the sound quality of stereo sound reproduction can be improved by adding an additional loudspeaker between each adjacent pair of loudspeakers. This additional center loudspeaker can be fed with the sum signal of the left and right channel. A major drawback of this approach is that cross talk with left and right channels is inevitable, and resulting in a narrowing of the stereo image. However, we derived a center channel's gain using the direction of a stereo image, which is time varying. It automatically tracks the main direction of the dominant signal.

4. POSITION INDEPENDENT STEREO

Another method to achieve correct localization for stereophonic sound reproduction in a wide listening area is to use a loudspeaker array and so-called time-intensity trading, a mechanism of the human auditory system determined via psychoacoustic experiments within a wide listening area [11]. The use of two spatially separated loudspeakers imposes restrictions on the ability of stereophony to reconstruct the correct acoustic field so that a sharp image can be perceived. Such a system can provide a well-defined image for a centrally located listener mainly at low frequencies, depending on the geometrical displacement of the speakers relative to the listener.

The basis of stereophony is the ability to create phantom sources. It is known that the brain locates a monophonic signal originated from a single source by comparing the differences in the arrival time and intensity of that signal at each ear. If the same monophonic signal is played through two loudspeakers on either side of the listener, then the sound seems to appear from midway between the two loudspeakers, since the traveling time of the signal arriving at each ear is the same. This is called a phantom (or virtual) source. We will discuss how to enlarge the region, within which the image remains reasonably. In general, it can be stated that correct localization within a wide listening area is beneficial for all applications, where a good stereophonic sound is required. The idea of achieving an enlargement of the sweet spot area in a stereophonic setup has been introduced and studied at the Philips Research Labs, Eindhoven, and the stereo sound system has been called ''Position Independent'' (PI). The main idea is that the directivity pattern of a loudspeaker array should have a well-defined shape so that a good stereo sound reproduction is achieved in a large listening area. Optimal digital filters are then

designed and applied to individual drivers of linear loudspeaker arrays in order to obtain a directivity pattern of a specific shape. This shape has then to be adapted to the time intensity trading mechanism of the human auditory system. The goal here is to derive an optimal directivity for the PI-stereo system, which is based on parameterized time intensity trading data, and then to find, by means of an optimization process, the corresponding FIR filter coefficients that achieve this optimal directivity pattern. It has been proven that an optimal directivity pattern for a loudspeaker can be realized by using an array of drivers positioned at a specific distance from each other. In our case, a practical design to achieve PI-stereo sound reproduction is a pair of loudspeaker cabinets, each cabinet equipped with a pair of drivers for the high frequency range with a separation suitable for this frequency range and for the mid frequencies with a corresponding separation, so as to obtain the desired optimal directivity pattern.

5. ULTRA BASS

In many sound reproduction applications, it is not possible to use large loudspeakers, due to size and or cost constraints. Typical applications are ambient audio, but also portable audio, multimedia, and TV. In these applications, the devices are often of small size, and therefore the transducers are inherently small as well. Needless to say, the competitive market also dictates the highest possible audio quality of these products. However, probably the most well-known characteristic of small loudspeakers is a poor low frequency (bass) response. In practice, this means that a significant portion of the audio signal may not be reproduced (sufficiently) by the loudspeaker. For loudspeakers used in applications as mentioned above, reproduction below 100 Hz is usually negligible, while in some applications, this lower limit can easily be as high as several hundred Hertz. The bass portion of an audio signal contributes significantly to the sound "impact," and depending on the bass quality, the overall sound quality will shift up or down. Therefore, a good low-frequency reproduction is essential.

A traditional and conceptually very simple method to increase the perceived sound level in the lower part of the audible spectrum (below the loudspeaker's resonance frequency, which is usually the lower limit) is to amplify the low frequency part of the audio spectrum, by a fixed or dynamic (depending on signal amplitude and or reproduction level) amount. For very low frequencies, the mechanical limits of the loudspeaker will limit the stroke the cone can make, leading to distortion and

possibly, loudspeaker overload. Thus, physically increasing the radiated sound pressure level means forcing the loudspeaker to radiate sound in a frequency range, for which it is not equipped. It may be better to prevent this completely by methods outlined below. In the process we shall discover several advantages of these methods. Now, from psychoacoustic theory, we know that a pitch perception can occur at a frequency that is not contained in the audio signal. This is possible through nonlinearities in the cochlea (difference tones), or a higher-level neural effect in the auditory system (virtual pitch). These two effects, appear to be very suitable effects for our purpose of enhancing bass perception using small loudspeakers. These effects can be utilized by some simple nonlinear (but controlled) processing, replacing very low frequencies in the audio signal by higher frequencies [12, 13]. These will still have the same perceived pitch as the original, using the psychoacoustic effects previously mentioned. Such effects also occurred in transistor radios, where undesired nonlinearities gave rise to a distorted sound. However, the method that we now propose uses nonlinearities in a controlled manner, and restricted to only the lowest frequencies, such that the effect is to our benefit. Without any information about the signal processing employed, we can immediately infer a number of advantages that such a scheme shall provide:

A higher radiated sound pressure level from a given loudspeaker, because of increased efficiency and decreased cone excursion. Furthermore, at higher frequencies the auditory system is more sensitive, which will also contribute to increased loudness;

Less power consumption, because of increased efficiency. This can be very important for portable applications; and,

Fewer disturbances in neighboring areas, because of the fact that the lowest frequencies are not physically present, while the added higher frequencies are absorbed more efficiently than the low frequencies.

6. INCREDIBLE SURROUND SOUND

It is virtually impossible to imagine sound reproduction today without stereophonic techniques, and it is to the credit of both the technology and human binaural hearing capabilities that a single pair of loudspeakers can evoke auditory perspectives so convincingly. Incredible Sound is a convincing stereo base-widening system, developed to improve the sound reproduction in applications with closely separated loudspeakers [14]. The aim of incredible surround sound is to offer a practical solution, replacing the traditional approach generally used. A filter is derived, using a simple model, where ideal loudspeakers and an acoustically transparent

subject's head are assumed. This system appeared to be very practical to implement and tolerant against head movements.

7. NOMADIC RADIO: WEARABLE AUDIO MESSAGING AND AWARENESS

Nomadic Radio developed at MIT Media Laboratory is as a unified messaging system that utilizes spatialized audio, speech synthesis, and recognition on a wearable audio platform. A client–server-based messaging infrastructure is already in place, and support is added for communication and location awareness. Messages such as hourly news broadcasts, voice mail, and email are automatically downloaded to the device throughout the day. The current system operates primarily as a wearable audio-only interface, although a visual interface is used for development purposes. A combination of speech and button inputs are used to control the interface. Textual messages such as email, calendar reminders, weather forecasts, and stock reports are delivered via synthesized speech. Users can select a category, such as news or email, browse messages sequentially, and save or delete them on the server. As the system gains the capability to determine its location, a scenario is envisioned, where the listener's location context enables the system to provide relevant messages as needed. For example, as the user moves close to a particular room, she may hear a voice message left by a colleague, or more importantly she is reminded of a meeting if she is not in a desired location at a specific time.

7.1. Design of the Wearable Audio Platform

Audio output on wearables requires use of speakers worn as headphones or appropriately placed on the listener's body. Headphones are not entirely suitable in urban environments, where users need to hear other sound sources, such as traffic or in offices, where their use is considered antisocial as people communicate frequently. In these situations, speakers worn on the body could instead provide directional sound to the user (without covering the ear), yet they must be designed to be easily worn and least audible to others. The *Soundbeam Neckset* (shown in Figure 6.3-7), worn around the neck, has been modified for audio I/O from the wearable. The *Neckset* is a patented research prototype originally developed by Andre Van Schyndel at *Nortel* for use in hands-free telephony. It consists of two directional speakers, mounted on the user's shoulders, and a directional microphone placed on the chest. A button on the *Neckset*

Figure 6.3-7. MIT's Soundbeam Neckset.

will activate speech recognition, or deactivate it in noisy environments. Spatialized audio is rendered in real-time and delivered to the *Neckset*.

8. THREE-DIMENSIONAL (3D) HEADPHONES

Headphone virtualizers that are commercially available today are optimized for a head other than that of the listener. This results in large localization errors for most listeners. At Philips Research, a system is introduced that includes a calibration procedure, which can be carried out conveniently by the listener [15, 16]. This system consists of ordinary headphones, into which microphones have been mounted. The sound reproduction using headphones then gives the same listening experience to the user as if the reference multichannel loudspeaker system was being used. Besides the usual computational requirement for a headphone virtualizer, this system needs in addition two low-cost microphones.

8.1. Technology Background

The way in which sound propagates from the loudspeaker towards the ear drums of the listener depends on the loudspeaker, the room, and the physical properties of the listener (e.g., the shape of the head, ears, and torso).

The physical properties of the head and outer ears of the listener modify the sound as it travels from the source to the eardrums. The transfer functions describing this sound propagation from multiple sound sources to both ears are known as head-related transfer functions

(HRTFs). Multichannel audio can be filtered with the HRTFs of the listener prior to headphone sound reproduction. If loudspeaker reproduction is emulated using headphones, compensation for the sound reproduction characteristics of the headphones is required. In this way the multichannel loudspeaker system can be emulated very accurately. When audio is filtered with HRTFs that are measured from another person, there are large errors in the vertical and front back localization. Therefore, a system is introduced that is personalized to the listener.

The system at hand consists of headphones with integrated microphones and a digital signal processing unit (DSP), to which the headphones are connected. During the calibration, the DSP is connected to a multichannel loudspeaker setup. A noise signal is played through each of the loudspeakers and is registered by the microphones. The DSP then computes how the sounds should be processed prior to headphone reproduction, such that exactly the same sound is generated at the position of the microphones, which are very close to the ears. When the calibration is completed, the listener can manually choose between loudspeaker or headphone sound reproduction, showing the capabilities of the system.

8.2. Hexaphone

A dedicated implementation of the work has been realized which is codenamed ''Hexaphone.'' Two examples are shown in Figures 6.3-8 and 6.3-9.

Figure 6.3-8. Prototype of headphones with integrated microphones.

Figure 6.3-9. Prototype of earphones with integrated microphones.

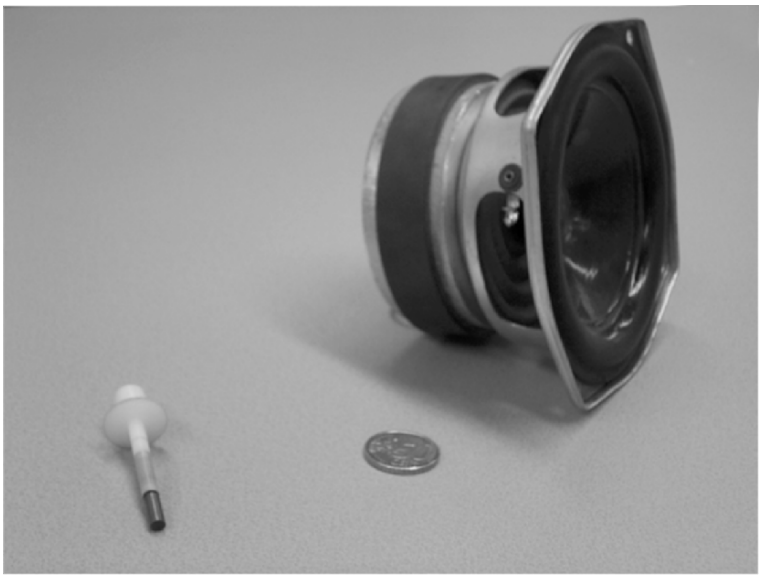


Figure 6.3-10. Left: the magnet system of the BaryBass transducer; right: a normal medium-sized bass loudspeaker. A 50 euro cents coin is shown for size comparison; the actual price is much less.

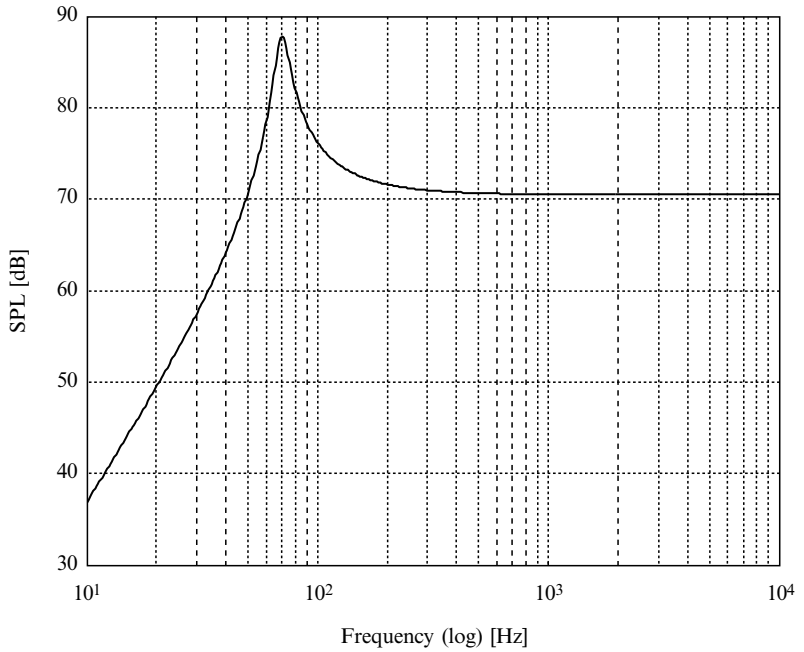


Figure 6.3-11. The response (Sound Pressure Level (SPL) [dB] versus frequency) of the BaryBass driver (log/log plot).

9. BARYBASS

Direct-radiator loudspeakers typically have a very low efficiency, since the acoustic load on the diaphragm or cone is relatively low compared to the mechanical load. On the one hand, the efficiency is inversely proportional to the moving mass, while on the other hand, it is proportional to the square of the product of the cone area and the force factor (determined by the magnet system and the voice coil). Furthermore, in order to get a sufficiently low resonance frequency, the moving mass must be high enough, and the cabinet volume—which acts as an air spring—must be large enough. However, for many consumer applications, the cone size should be small. In addition, the driving mechanism of a voice coil is quite inefficient in converting electrical energy into mechanical motion. These conflicting conditions cannot be met with a classical loudspeaker. Low frequency drivers (woofers) have a magnetic structure (see Figure 6.3-10, right side) that is rather large, so that the typical frequency response is flat enough and the efficiency is high enough. The solution consists of two steps [5]. First, we relax the requirement that the frequency response must be flat. By making

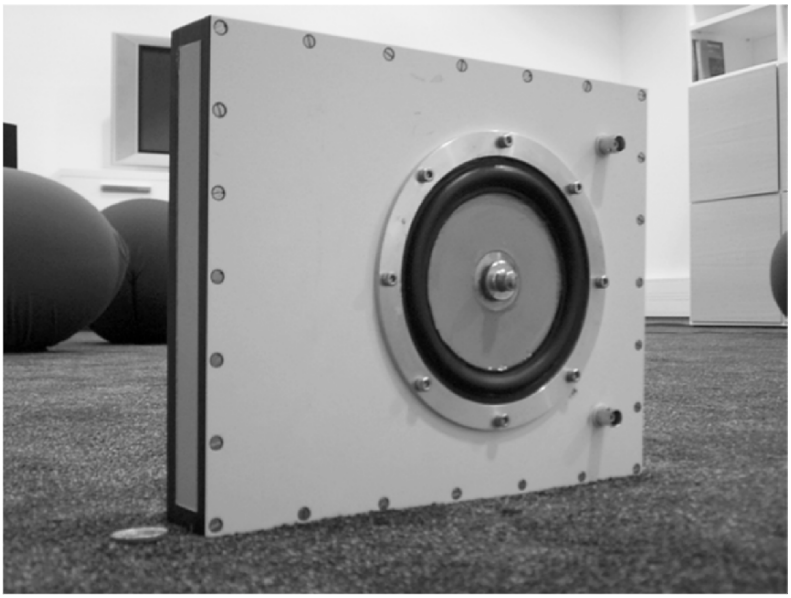


Figure 6.3-12. This new loudspeaker can be mounted in a very small volume and thin cabinets and blend invisibly into your room.

the magnet considerably smaller (see Figure 6.3-10, left side), a large peak in the sound pressure level (SPL) curve (see Figure 6.3-11) will appear. At the resonance frequency, the efficiency can be a factor 10 higher than that of a normal loudspeaker. In this case, we have at the resonance frequency of about 70 Hz—a high level of almost 90 dB @ 1 Watt input power, using only a small cabinet. Since it is operating in resonance mode only, the moving mass can be enlarged, without degrading the efficiency of the system. Due to the large peak, the normal operating range of the driver decreases considerably, however. This makes the driver not suitable for normal use. To overcome this, a second measure is applied. We map the low frequency content of the music signal, say, 20–120 Hz, to a slowly amplitude modulated tone, whose frequency equals the resonance frequency of the transducer. The modulation is chosen so that the coarse structure (the envelope) of the music signal after the mapping is the same as before the mapping. The required electronics is implemented both in the digital and in the analog domain, the latter one requiring less than a dozen transistors and a few RC components. An example of a BaryBass driver mounted in a flat enclosure with a volume of less than 1 l is given in Figure 6.3-12. The resonance frequency is about 50 Hz, which is probably, currently, the world's smallest subwoofer with such a low-resonance frequency, while the sound power efficiency is very high.

10. CONCLUSIONS

We have shown that using either special loudspeakers or signal processing, it is possible to go beyond the limits, which are usually dictated by physics. The reason that this is possible indeed is to utilize psychoacoustic phenomena, which relax these limits. This gives new opportunities for sound reproduction in general, but in particular in the ambient intelligence context.

REFERENCES

[1] Hunt, F. V., 1954, *Electroacoustics*, John Wiley.
[2] Olson, H. F., 1957, *Acoustical Engineering*, Van Nostrand.
[3] 1998, Commemorative issue: 50 years of contributions to audio engineering', *J. Audio Eng. Soc.* **46**(1/2).
[4] Vanderkooy, J., Boers, P. M. and Aarts, R. M., 2003, Direct-radiator loudspeaker systems with high Bl, *J. Audio Eng. Soc.*, **51**(7/8), 625–634.
[5] Aarts, R. M., 2005, High-efficiency low-B1 loudspeakers, *J. Audio Eng. Soc.*, 53(7), 579–592.
[6] Aarts, R. M. and Johnson, M. T., 2002, Sound and vision system, European Patent EP1386519, Filed 20 March 2002.
[7] Westervelt, P. J. and Larson, R. S., 1973, Laser-excited broadside array, *J. Acoust. Soc. Am.*, **54**(1), 121–122.
[8] Yoneyama, M., et al., 1983, The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design, *J. Acoust. Soc. Am.*, **73**(5), 1532–1536.
[9] Pompei, F. J., 1999, The use of airborne ultrasonics for generating audible sound beams, *J. Audio Eng. Soc.* **47**(9), 726–731.
[10] Irwan, R. and Aarts, R. M., 2002, Two-to-five channel sound processing, *J. Audio Eng. Soc.*, **50**(11), 914–926.
[11] Rodenas, J., Aarts, R. M. and Janssen, A. J. E. M., 2003, Derivation of an optimal directivity pattern for sweet spot widening in stereo sound reproduction, *J. Acoust. Soc. Am.*, **113**(1), 267–278.
[12] Larsen, E. and Aarts, R. M., 2004, Audio bandwidth extension, *Application of Psychoacoustics, Signal Processing and Loudspeaker Design*, John Wiley.
[13] Larsen, E. and Aarts, R. M., 2002, Reproducing low-pitched signals through small loudspeakers, *J. Audio Eng. Soc.*, **50**(3), 147–164.
[14] Aarts, R. M., 2000, Phantom sources applied to stereo-base widening, *J. Audio Eng. Soc.*, **48**(3), 181–189.
[15] Aarts, R. M., 2004, Headphones with integrated microphones, US patent 6,829,361, December 2004.
[16] Schobben, D. and Aarts, R. M., 2005, Personalized multi-channel headphone sound reproduction based on active noise cancellation. *Acta Acoustica*, 91(3), 440–450.

Chapter 6.4

SECRET KEY GENERATION FROM CLASSICAL PHYSICS
Physical Uncloneable Functions

Pim Tuyls and Boris Škorić
Philips Research Eindhoven
pim.tuyls@philips.com; boris.skoric@philips.com

Abstract In this chapter we explain why security is important in an ambient intelligent (AmI) environment. In order to achieve trust and security in this environment not only cryptographic algorithms are needed, but also secure methods for generation and storage of secret keys. Physical uncloneable functions (PUFs) can be used to this end, because they have built in security properties, such as uncloneability and tamper evidence. We show that cryptographic keys can be extracted securely from PUFs. This enables one to go beyond simple identification applications. In particular, more advanced protocols, such as authentication, key exchange, certified execution, and proofs of execution, can be based on PUFs.

Keywords authentication; encryption; key extraction; tamper resistance

1. INTRODUCTION

As electronic components are becoming cheaper, smaller and more powerful will become omnipresent, and influence our daily life in many aspects. The electronics will be *networked* and disappear into the background, becoming invisible. Devices will react in an *adaptive*, *personalized,* and *anticipatory* way to human beings. In [1], a list of enabling technologies for ambient intelligence (AmI) is given. It ranges from user-centered design and ambient lighting to ubiquitous computing and trust. In this chapter, we will mainly discuss this last item; explain why it is important,

S. Mukherjee et al. (eds.), AmIware, 421–447.

and investigate how physical principles can be used to protect cryptographic keys.

The very defining features of AmI imply a number of serious security risks. In *networked* systems, many devices communicate with each other, which allows for many points of attack. The *personalized* nature of the interactions requires the storage of customization parameters, which may be privacy sensitive. The *context awareness* requires numerous sensors, giving attackers a chance to eavesdrop on our private affairs. Finally, *adaptive* and *anticipatory* devices have to be programmable, which introduces all the well-known risks of viruses, spyware, worms, and Trojan horses inherent in programmable environments.

In order to work well, *trust* and *confidence* in the surrounding electronics in an AmI world is of crucial importance. The realization of these two properties depends on the use of security technologies. One of the key components of security technology is cryptography. Its three traditional objectives are: *confidentiality*, *authenticity,* and *non-repudiation*. By using encryption algorithms, people can send secret information to specific recipients. This provides confidentiality. Authentication is a technique that allows checking, whether information is indeed coming from the stated sender and, whether it has been tampered with. Non-repudiation, finally, is a property of a security system that prevents people from denying previous actions or commitments.

All those algorithms depend on *secret keys*. The algorithms are always assumed to be public and only the keys are considered to be secret. The security of cryptographic algorithms depends therefore on the secrecy of the keys.

In order to keep the keys secret, cryptographic algorithms are designed such that, given that an attacker knows the algorithm and captured some messages (*plaintexts*) and their encryptions (called *ciphertexts*), it remains very difficult for him to extract the secret key from these plaintext–ciphertext pairs. The security properties of currently used symmetric key crypto systems, like AES and DES., and public key systems, like RSA, El Gamal, etc., are well-investigated and understood. It turns out that the best-known attacks against those crypto systems are still too complex to be practical.

Another way to find the secret key is to attack the physical device, where the key is stored. This kind of attack is called a *hardware attack*. This attack turns out to be much easier and hence much more successful [2, 3]. Therefore, there is a need to protect keys against hardware attacks. This is the topic of this chapter. We extend a recently proposed method to protect cryptographic keys. Instead of protecting keys by appropriate defensive measures (like coatings, packaging, etc.), we derive keys from

complex physical systems that are inherently uncloneable. The idea is that many keys can be derived from those systems and that they are automatically destroyed when the physical system is tampered with.

1.1. Motivation

Hardware components for security applications (memory, discs, processors, etc.) have to satisfy very tough requirements. As explained above, it is important that they keep critical secrets, such as cryptographic keys hidden, often while performing computations that employ these secrets. To this end, the hardware must be made resistant to various types of passive and active attacks, such as probing, etching, electron beams, ion beams, machining methods, fault induction, electromagnetic analysis (differential) power analysis, etc. The usual way of making chips opaque to scrutiny is to cover them with a protective coating, or to package them in a strong barrier. In some applications, it is even mandatory for security components to detect tampering attacks while they are underway, or at least to show evidence of tampering afterwards (*tamper evidence*). Sometimes, it is required that they erase their critical secrets as a response to an attack. Current detection methods make use of wires under constant surveillance or various sensors that measure voltage, temperature, or mechanical stress. A final constraint is that all these hardware measures have to be kept as cheap as possible. Clearly, this puts high pressure on their design and often results in compromises at the cost of security. When tamper proofing fails and secrets do leak out of a device, hackers can impersonate the device by making "clones." Well-known examples of clones are: copied SIM-cards, credit cards, pay-TV smartcards, bank passes (fuel station attack[1]). Losses due to cellular phone cloning are estimated at hundreds of millions of dollars per year.

Physical uncloneable functions (PUFs) have been proposed as a cost-effective way to produce uncloneable tokens for identification [4, 5]. The identification information is contained in an inexpensive, randomly produced (i.e., consisting of many random components), highly complicated piece of material. The secret identifiers are read out by performing measurements on the physical system and performing some additional

[1] In the *fuel station attack*, an attacker places a small (unnoticeable) coil in the bankpass reader at the fuel station and attaches a small camera to the roof. When your bankpass is read by the machine, the attacker reads the critical information on the magnetic strip of your bankpass and records your pincode with the camera. Using this information, he produces a cloned bankpass. Using this cloned bankpass and your pincode, he can now withdraw or spend money in your name.

computations on the measurement results. The advantage of PUFs over electronic identifiers lies in the following facts:

(1) Since PUFs consist of many random components, it is very hard to make a clone, either a physical copy or a computer model;
(2) PUFs provide inherent tamper evidence due to their sensitivity to changes in measurement conditions; and
(3) Data erasure is automatic, if a PUF is damaged by a probe, since the output strongly depends on many random components.

The outline of this chapter is as follows. In Section 1.2, we present a general introduction to PUFs and in Section 1.3, we give some intuition for their applications. Then in Section 1.4, we list a number of physical systems, on which PUFs can be based. The security model and the possible ways of attacking PUFs are discussed in Section 1.5. In Section 2, we explain how key extraction from noisy measurements is performed. We introduce a new algorithm using *helper data* and we demonstrate its applicability for optical and coating PUFs. Finally, in Section 3, we give protocols for PUF-based authentication, renewal, copy protection, certified execution, and proof of execution.

1.2. General Introduction to Physical Uncloneable Functions (PUFs)

A PUF is a function that is realized by a physical system, such that the function is easy to evaluate, but the physical system is hard to characterize and hard to clone [4, 6, 7]. The PUFs were introduced by Pappu [4] as a cost-effective way of identifying physical tokens. Since a PUF cannot be copied or modeled, a device equipped with a PUF becomes uncloneable. This makes PUFs attractive as a protective measure against attacks based on copying of key material (fuel-station attack), and for digital rights management (DRM) systems.

A PUF is a physical system designed, such that it interacts in a complicated way with stimuli (challenges), and leads to unique and unpredictable responses. A PUF *challenge* and the corresponding *response* are together called a *challenge–reponse-pair* (CRP). Since several challenges can be applied that lead to unique responses, a PUF is modeled as a function mapping challenges to responses. From a security perspective, a PUF is similar to a keyed hash function. The physical system consisting of many *random* components is equivalent to the key. Given the input (challenge) and the key (the system), the output is easily generated; but given the output, it is difficult to obtain information about the key. Furthermore, the system does not allow efficient extraction of the relevant

properties of its interacting components by investigation of a small number of CRPs.

Physical systems that are produced by an uncontrolled production process (i.e., one that contains some randomness), turn out to be good candidates for PUFs. Because of this randomness, it is hard to produce a physical copy of the PUF. Furthermore, if the physical function is based on many complex interactions, then mathematical modeling is also very hard. These two properties together are defined as *uncloneability*. The most well-known examples of PUFs are: *Optical PUFs*, *Acoustic PUFs*, *Coating PUFs,* and *Silicon PUFs,* though many more are possible.

In order to use PUFs, we need hardware components for challenging the PUF and a detector for measuring its responses. Additionally, processing power is needed to turn the outcomes into secret bit strings that can be used for cryptographic purposes. The challenging components, the detector and the processor executing the algorithm can be located on the device with the embedded PUF, or they can be inside a separate reader device. In the first case we speak of an *integrated PUF,* while the latter one is a *stand-alone PUF*. An advantage of an integrated PUF is that it allows adding *control* to it. A *controlled PUF (CPUF)* is a PUF that is bound to a processor, which completely governs the input and output. The chip prevents frequent challenging of the PUF and forbids certain classes of challenges. It scrambles incoming challenges, so that an attacker cannot systematically probe the device. Furthermore, it hides the physical output of the PUF, revealing to the outside world only indirect information derived from it (e.g., by encrypting or hashing the bit-string that is extracted from the analog output). Hence this control layer substantially strengthens the security.

1.3. Applications

From a security perspective, the uniqueness of the responses and uncloneability of the PUF are very useful properties. Because of these properties, PUFs can be used as unique identifiers, means of tamper-detection and/or as a cost-effective source for key generation (common randomness) between two parties. By embedding a PUF inseparably into a device, the device becomes uniquely identifiable and uncloneable. Here "inseparable" means that any attempt to remove the PUF will, with very high probability, damage the PUF and destroy the key material it contains. A wide range of devices can be equipped in this way (e.g., smart-cards, credit cards, RFID tags, value paper, chips, security cameras, etc. This makes it possible to identify these devices and check the trustworthiness of their output. We discuss two kinds of applications: one for stand-alone PUFs and a more advanced one for integrated PUFs.

The usage of a PUF consists of two phases: *enrolment* and *authentication*. During the enrolment phase, the Verifier produces the PUF and stores an initial, small set of CRPs securely in his database. Then the PUF is embedded in a device and given to a user. The *authentication* phase starts when the user presents his device to a terminal. The Verifier sends a randomly chosen PUF challenge from his database to the user. If the Verifier receives the correct response from the device, the device is *identified*. Then, this CRP is removed from the database and will never be used again for identification purposes.

If, additionally, one needs to exchange secret messages between the device and the Verifier (e.g., for a financial transaction), a secure authenticated channel is set up between the Verifier and the device, using a session key based on the PUF response.

The CPUFs allow for new applications, such as *"certified execution"* and *"certified measurement"* [6]. A CPUF can generate a signed digital certificate, stating that a certain kind of processing has occurred inside the CPUF, such as program execution or some measurement, and listing the results of this processing. The certificate can be verified by the original Verifier as well as by third parties. A powerful example is a CPUF bound to a security camera. Footage from such a camera can serve as strong evidence in court, since the certificate allows the judge to verify the identity of the camera and to ascertain that the pictures have not been tampered with.

1.4. Physical Realizations

Several physical systems are known on which PUFs can be based. The main types are Optical PUFs [4], Coating PUFs, Silicon PUFs [6, 7], and Acoustic PUFs. In this chapter, we mainly discuss Optical PUFs and Coating PUFs.

1.4.1. Optical PUFs

Optical PUFs consist of a transparent material (e.g., glass) containing randomly distributed light scattering particles. They exploit the uniqueness of speckle patterns that result from multiple scattering of laser light in a disordered optical medium. The response (output) is a speckle pattern. It is a function of the internal structure of the PUF, the wavelength of the laser, its angle of incidence, focal distance, and other characteristics of the wavefront.

Optical probing of the PUF is difficult because the light diffusion obscures the locations of the scatterers. At this moment, the best physical techniques can probe diffusive materials up to a depth of approximately 10 scattering lengths [8].

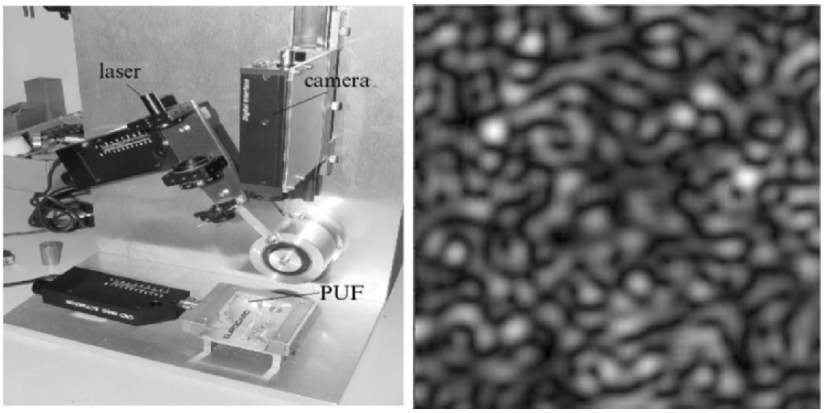


Figure 6.4-1. Left: Experimental apparatus for the read-out of an Optical PUF embedded in a credit card. Right: Speckle pattern that is obtained when an Optical PUF is irradiated with a laser beam.

Even if an attacker learns the positions of all the scatterers, this knowledge is of limited use to him. If he tries to make a physical copy of the PUF, he runs into the problem that precise positioning of a large number of scatterers is an arduous process. It would seem easier to make an "electronic" clone (i.e., a device that simply looks up the correct responses to all challenges from electronic memory) without bothering with the physics. However, even this turns out to be very hard. It requires optical modeling of multiple coherent scattering. More precisely, the attacker has to solve the "forward problem," which is a very complex task [9].

1.4.2. Coating PUFs

Coating PUFs are PUFs that are integrated with an IC, and hence belong to the class of integrated PUFs. More precisely, the IC is covered with a coating consisting of (e.g., aluminophosphate), which is doped with random dielectric particles (e.g., TiO_2, $SrTiO_3$, or $BaTiO_3$). By random dielectric particles, we mean several kinds of particles of random size and shape with a relative dielectric constant ε_r differing from the dielectric constant of the coating matrix. The PUF consists of the combination of the coating with the dielectric material.

In order to challenge the Coating PUF, an array of metal sensors (e.g., a comb structure), is laid down between the passivation layer and the coating. Sufficient randomness is only obtained if the dielectric particles are smaller than the distance between the sensor parts.

A challenge corresponds to a voltage of a certain frequency and amplitude applied to the sensors at a certain point of the sensor array. Because of the presence of the coating material with its random dielectric properties, the sensor plates behave as a capacitor with a random capacitance value. The capacitance value is then turned into a key.

Coating PUFs have the advantage of possessing a high degree of integration. The matrix containing the random particles can be part of the opaque coating. Thus, the tamper-resistance coating, which protects the secrets present in the electronics (stored and during computation), itself serves as a carrier of (inherently tamper-resistant) secrets. Coating PUFs also have the advantage that they can be easily turned into a Controlled PUF (CPUF). The control electronics can simply be put underneath the coating.

Probing the PUF from the outside gives insufficient information to the attacker. The outcomes of the capacitance measurements from inside are very sensitive to the precise locations of the dielectric particles. Even if the precise locations of the random particles are known, physical reproduction of the coating costs a prohibitive amount of effort because of the complexity.

If successful probing is possible, then *electronic* cloning may be feasible. Our current knowledge of Coating PUFs indicates that the amount of key material present in a Coating PUF is much smaller than for Optical PUFs.

1.4.3. Acoustic PUFs

In an Acoustic PUF, we measure the response of an object to an acoustic wave. An electrical signal is transformed to a mechanical vibration through a transducer. This vibration propagates as a sound wave through the object and scatters on the randomly distributed inhomogeneities in the object. The reflections of those waves are measured by another transducer, which converts the vibration back into an electric signal. It turns out that the reflections are unique for each token.

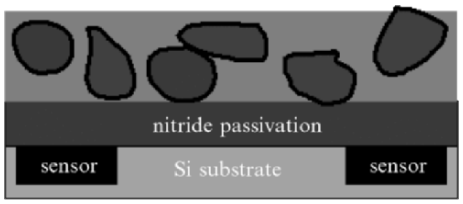


Figure 6.4-2. Structure of a Coating PUF.

1.4.4. Silicon PUFs

Silicon PUFs were introduced in [7]. They exploit the fact that manufacturing variations in a circuit cause substantial differences in circuit delays. These variations are caused by mask variations, but also temperature and pressure variations during manufacturing. The magnitude of the delay variations caused in this way is about 5%.

A challenge for the PUF is a digital input signal. The delay caused by the circuit between input and output is the response of the PUF. It turns out, however, that these PUFs are quite sensitive to temperature variations, and that compensation algorithms for this effect have to be implemented to make the system work properly. For more details about Silicon PUFs, see [7].

1.4.5. Attack models

Introduction of a new type of security hardware will of course invite new modes of attack. It is important to have a good understanding of the exact level of security offered by a PUF. Loosely speaking, we say that a PUF is compromised, or broken, as soon as an attacker can (a) physically reproduce the PUF or (b) predict responses with reasonable probability.[2] In case (b), the attacker cannot reproduce the physical token, but he can mimic its behavior.

We make the following assumptions. The enrolment takes place at a trusted authority (i.e., an authority that does not misuse CRPs or give them away to third parties). For the authentication phase, we distinguish between uncontrolled PUFs and CPUFs. CPUFs contain a trusted internal reader (i.e., responses are not visible in the clear). Hence, CPUFS can authenticate messages themselves, and the ''external'' reader has no function other than passing messages between the CPUF and the Verifier. For uncontrolled PUFs, we assume that the reader is *transparent*. By transparent, we mean that an attacker can see the information that is processed by the reader (including responses), but cannot modify this information or the actions of the reader. Furthermore, we assume that the reader is capable of authenticating messages to the Verifier with a secret key built into (nontransparent) tamper-resistant hardware.

Hence, the communication channel from the reader to the Verifier is authenticated for both cases. Consequently, an attacker can learn CRPs, but cannot insert his own messages to the Verifier into this channel. Furthermore, we assume that the two-way communication channel is

[2] Reasonable probability means a probability which is substantially better than random guessing.

public (i.e., susceptible to eavesdropping). Finally, the Verifier is assumed to be trustworthy.

Below, we list a number of attack methods (without pretending to be exhaustive) that target either the physical token itself, or those parts of the protocol that involve CRP handling. For each attack we discuss countermeasures.

1.4.6. CRP measurement attack

Attack

The attacker gains unnoticed access to the victim's PUF for a short period of time and quickly measures and stores as many CRPs as he can. We differentiate between

(1) Brute Force: The full set of CRPs is obtained and stored in a lookup table. The attacker uses this table as an electronic clone of the PUF.
(2) Intrapolation: Sufficiently, many CRPs are obtained for the creation of an electronic clone based on a physical modeling algorithm that can intrapolate between the measured CRPs.

Countermeasures

PUF read-out should be relatively slow [10]. In this way the attack will last too long to be practical. Further countermeasures are

(1) Brute Force: The PUF must support a huge number of CRPs, so that the attack takes too much time.
(2) Intrapolation: Idem. In addition, the complexity of the physics should reduce the effectiveness of the intrapolation algorithm.

Remarks

- The most trivial of all attack methods is the brute force attack, requiring the least amount of skill;
- A CPUF is resistant against the measurement attack, since it forbids challenges to take place in quick succession, and it does not disclose the responses in the clear (see Section 3.4); and
- In [10], a quantitative measure was introduced that indicates how resilient a PUF is against brute force attacks, namely, the number of different CRPs that can be used before a PUF's secrets are exhausted. An equivalent formulation of statement is: how many CRPs does an attacker have to measure before he can, information-theoretically speaking, predict responses to new challenges?

For Optical PUFs, the analysis of the security parameter was performed in [10]. If one models an Optical PUF as a waveguide with random scattering between the transverse modes, it can be shown that the security parameter is approximately given by

$$N_{\text{tot}} \approx \frac{A_{\text{PUF}}}{A} \frac{d^2}{\lambda l}, \tag{6.4-1}$$

where l is the mean free path, A the irradiated area of the PUF by the laser, A_{PUF} the total area of the PUF, λ the wavelength of the laser, and d is the thickness of the PUF. Note that the result in Equation (6.4-1) refers to an idealized model with perfect detection, where the only noise present is the quantization noise in the number of photons. Although the result should be viewed as a "worst case" answer, we expect that the scaling in the parameters d, λ, and l holds.

1.4.7. Eavesdropping on the channel between the prover and the verifier

Attack

Attackers eavesdrop on the communication channel between the PUF and the Verifier.

Countermeasures

Protocols have to be designed, such that an attacker does not gain useful information from the messages going over the communication channel. (Obviously responses are secret, while challenges might be public.) To achieve this, they can either use classical cryptographic techniques, like encryption algorithms, Zero-Knowledge protocols, etc. or exploit the huge number of CRPs present in the PUF.

1.4.8. Man-in-the-middle attack

A man-in-the-middle attack exists against PUF-based authentication schemes. By way of example, we investigate the case, where a smart-card equipped with a PUF is used for authentication. We assume that the attacker has captured one CRP.

The man-in-the-middle positions himself between the reader and the Verifier. When the smart-card initiates contact with the Verifier, the man-in-the-middle intervenes and impersonates the Verifier. He performs the following steps:

(1) He sends the challenge from the one CRP known to him to the smart-card;

(2) The smart-card responds correctly, and now the attacker has established a Secure Authenticated Channel (SAC) with the smart-card;
(3) The attacker contacts the Verifier. The Verifier starts the authentication protocol by sending a challenge. The attacker captures the challenge and initiates a renewal protocol (see Section 3.2) with the smart-card and forwards the Verifier's challenge;
(4) The smart-card sends its secret response to the attacker over the SAC; and
(5) Using this secret, the attacker constructs a correct response to the Verifier.

As a result, the Verifier and the smart-card think that they have established a SAC with each other, but in fact the attacker can eavesdrop on any message that passes between them (since he has a common secret with the smart-card as well as with the Verifier). Note that the attack applies to CPUFs as well as uncontrolled PUFs.

Several countermeasures are possible. One option is to store the Verifier's public key in the smart-card, which is used at the start of any protocol to check that the protocol was initiated by the legitimate Verifier. This requires some public key operations in the smart-card, which are rather expensive and hence should be avoided as much as possible.

A cheaper countermeasure makes use of the way, in which PUF challenges are processed. The underlying assumption needed for the man-in-the-middle attack to work is that, the Verifier's authentication challenge can be fed to the PUF when the smart-card is expecting a renewal challenge. Hence, the attack is thwarted by cleverly differentiating between the two input modes. In Section 3, a CPUF protocol is presented, where authentication challenges are fed directly to the PUF, but renewal challenges are first subjected to a one-way hash function by the smart-card. In this way, the attacker is forced to compute the inverse hash of the Verifier's authentication challenge before he can send it to the smart-card.

1.4.9. Active physical probing of the PUF

Attack

The attacker determines the exact characteristics of the random components of the PUF by physical probing. Then he creates a clone by

(1) Physical reproduction; and
(2) Physical modeling of the input–output behavior.

Countermeasures

Complexity of the PUF and of the physics underlying the challenge–response behavior. Ultimately, the PUF's random characteristics can always be obtained by a resourceful attacker. A complex PUF, however, will force the attacker to buy expensive equipment, or to spend a large amount of time probing. Furthermore, the creation of a clone is made difficult by

(1) a large number and high complexity of components; and
(2) complexity of the physics.

Remarks

- A wide variety of physical probing methods is available to the assailant: dissection, X-rays, electron microscopes, MRI, etc.
- It is very hard, if not impossible, to define a quantitative measure of resilience. As with all tamperproofing efforts, it is an arms race between the security engineers and the attackers.
- There is an important difference with ordinary, electronic storage of secrets. A successful characterization of the PUF is not sufficient to break the security. In addition, the attacker will either have to painstakingly produce a replica, or to spend a lot of computing power in an effort to correctly model the challenge–response physics.

2. KEY EXTRACTION

In the study by Pappu et al. [5], it was demonstrated that Optical PUFs can be used for identification. There, the Verifier's decision is based on the correlation between two speckle patterns: When the correlation is above a certain threshold value, the Verifier is convinced that the proving party possesses the PUF that was originally enrolled. If the correlation is lower, identification fails.

This correlation-based approach has two main drawbacks:

(1) The proving party has to send a speckle pattern to the Verifier, revealing information about the PUF to eavesdroppers; and
(2) At the end of a successful identification, the Prover and Verifier have *not* established a shared secret that can be used as a session key. Extraction of a key would allow for more advanced protocols.

In this section, we explain how a secure key is derived from a PUF. Because of noise, the responses to a challenge during enrolment and

verification differ slightly. Cryptographic functions (encryption functions, one-way functions, etc.) are very sensitive to small changes in the keys they use. Hence, responses to challenges cannot be used as cryptographic keys. Therefore, additional techniques have to be applied to eliminate the noise. The algorithm that we present for key extraction from PUFs is based on so-called *helper data* [11, 12] (or, equivalently, Fuzzy Extractors [13]).

2.1. The Principle of Key Extraction from Noisy Data

Generally speaking, a key extraction algorithm is built on a Secret Extraction Code (SEC) [11], or equivalently a fuzzy extractor [13]. For the sake of simplicity, we describe the algorithm in terms of a *shielding function* [12], which generates a special set of SEC [11], while having all the necessary properties. We denote the PUF response to a challenge C during the enrolment phase by $X \in R^n$, and during the verification phase by $Y \in R^n$. A function $G(.,.)$ is called δ-*contracting,* if for all Y that lie within a sphere of radius δ of $X(\| X - Y \| \leq \delta)$, and for fixed W,$G(Y,W) = G(X,W)$. We use δ-*contracting* functions to extract keys $S = G(X,W)$ from noisy data X using helper data W. A function $G(.,.)$ is called ε-*revealing,* if the helper data W leaks less than ε bits on S (in the information-theoretic sense) (i.e., $\mathbf{I}(W;S) \leq \varepsilon$). An ($\varepsilon$, δ)-*shielding* function G: $\mathbf{R}^n \times \{0,1\}^w \rightarrow \{0,1\}^k$ is a function that is δ-contracting and ε-revealing. It is used to extract a secret of length k from the PUF response as follows.

2.1.1. Enrolment phase

During the enrolment phase, the PUF is challenged with a challenge C and the response X according to this challenge is measured. Then, a random key S is chosen from $\{0,1\}^k$ and helper data W is computed by solving $G(X,W) = S$ for W. The quadruplet *(PUF_ID, C,W,S)* is then stored in a CRP database.

2.1.2. Verification phase

When the PUF is inserted into the reader, the PUF's identity is sent to the Verifier. The Verifier chooses a random challenge C from its database and sends it to the PUF together with the corresponding helper data W. Then the PUF is challenged with the challenge C and its response Y is measured. A key S' is then computed as $S' = G(Y,W)$. It follows from the δ-contracting property of the function G that S' equals S, if Y is sufficiently close to X. The key S can then be used securely, since the helper data does not reveal information on S.

2.2. Key Extraction from Optical PUFs

In practice, one faces several sources of noise when trying to extract keys from speckle patterns:

- stray light;
- scratches, dirt, and moisture on the PUF surface;
- misalignment of the laser with respect to the PUF; translations, and rotations; and
- misalignment of the detector with respect to the PUF.

We present the results of a number of experiments that we have done.

2.2.1. Experimental setup and key extraction method

We performed experiments with the following setup. The laser is a DBF laser with a wavelength of 785 nm (spectral width 1 nm). The beam diameter is 1 mm. We have used five scattering samples with a thickness of 0.4 mm. Pictures of the reflected speckle pattern are taken with a 1024 by 768 pixel CCD camera with a pixel pitch of 6.25 μm. The bitmap has 256 gray levels. The distance between the laser and the sample is 10 cm, and the distance from the sample to the camera is 13 cm.

Bit-strings were extracted from the speckle patterns as follows. Each bitmap is first down sampled by a factor of 8 (in both x and y direction) in order to reduce the redundancy present in the picture due to the speckle size, which in the geometry described above is larger than 8 pixels on average. Then a Gabor filter is applied [4]. Only the Gabor coefficients in the 45° and 135° directions are computed and quantized. This yields a bit-string of length 2400.

In the enrolment phase, helper data is constructed by selecting those Gabor coefficients that have an absolute value above a certain threshold. These are the "robust" components of the response. Positive values are mapped to "1," negative values to "0," which leads to a bit string Z. The threshold is chosen, such that averaged over speckle patterns, approximately 1000 Gabor coefficients exceed the threshold. Then, an *error correcting code* (ECC) is chosen and a secret code word s in ECC is randomly generated. Then the difference $W = z \oplus s$ is stored as part of the helper data.

The total set of helper data consists of a set of indices pointing at the locations of the robust Gabor coefficients and of the string W. The Verifier stores the secret s and the helper data.

During the authentication phase, a speckle pattern is measured and Gabor coefficients are derived in the same way as during the enrolment phase. Then, the helper data from the enrolment phase are used to select

the robust coefficients. This leads to a bit string Z'. Then, the second part of the helper data, W, is used to compute $Z' \oplus W = (Z' \oplus Z) \oplus s$. Clearly, when the number of errors is not too large, then the error correction code decodes this correctly into s.

2.2.2. Results

The effect of small misalignments is shown in Figures 6.4-3 and 6.4-4. The sensitivity to rotations was tested by varying the angle of incidence of the laser beam. The sensitivity to translations was tested by shifting the samples in one direction. All measurements were repeated 10 times (reinserting the samples each time), and averaged over these 10 instances. As a direct measure of the difference between two speckle patterns S_1, S_2, we use the correlation C between the bitmaps,

$$C = \frac{\langle S_1(\vec{x}_i) S_2(\vec{x}_i) \rangle_i - \langle S_1(\vec{x}_i) \rangle_i \langle S_2(\vec{x}_i) \rangle_i}{\sigma_1 \sigma_2};\ C \in [-1, 1],$$

where $<>_i$ denotes the spatial average and σ is the standard deviation in the gray level of the speckle pattern.

The graphs in Figures 6.4-3 and 6.4-4 show that the Gabor coefficients look completely independent (50% errors) for rotations larger than 0.7 mrad and shifts larger than 0.1 mm. The robust bits, however, are significantly more resilient; there, the error level of 50% is reached at much larger perturbations.

In order to correct the noise on the robust bits, we have taken a BCH code with parameters (1023, 56, and 191). This means that we have 2^{56} code words of length 1023 in the code that can correct 191 errors. The secret keys have a length of 56 bits.

2.3. Key Extraction for Coating PUFs

In order to give an idea of the identification capacity of Coating PUFs, we present the following experimental results in Figure 6.4-3. The difference between the intraclass and interclass variance is clearly very large. This allows us to derive unique keys from different ICs.

We describe the procedure for extracting key bits from a Coating PUF. This method is called *quantized key extraction,* and allows to extract keys from analog (continuous data). For the sake of simplicity, we show how 1 bit can be reliably derived from a single capacitance measurement. It is straightforward to extend this method to the extraction of more bits. The capacitance value is a zero mean Gaussian distributed random variable with variance σ_x^2. For simplicity, we adopt a model where the noise N is also a zero mean Gaussian random variable with variance σ_n^2. For

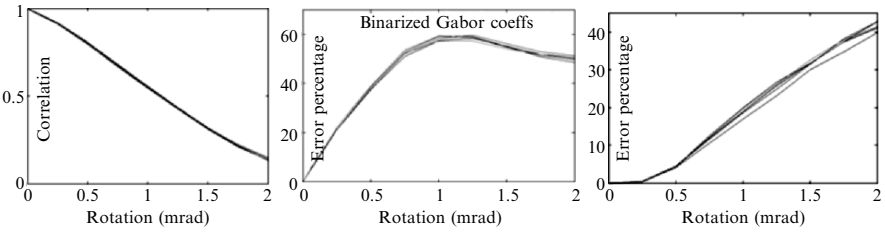


Figure 6.4-3. Effects of rotating the angle of incidence. Left: correlation between original and perturbed speckle pattern. Middle: error percentage in the binarized Gabor coefficients (2400 bit-string). Right: error percentage in the selected robust bit-string (≈1000 bits).

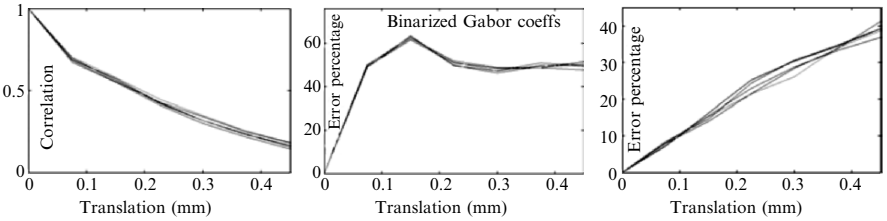


Figure 6.4-4. Effects of shifting the sample. Left: correlation between original and perturbed speckle pattern. Middle: error percentage in the binarized Gabor coefficients (2400 bit-string). Right: error percentage in the selected robust bit-string (≈1000 bits).

$Y = X + N$, helper data W and secret S, the δ-contracting function is given by

$$S = \begin{cases} 1 & \text{if } 2nq \leq Y + W < (2n+1)\,q, \quad \text{for any } n = \ldots, -1, 0, 1, \ldots \\ 0 & \text{if } (2n-1)\,q \leq Y + W < 2nq, \quad \text{for any } n = \ldots, -1, 0, 1, \ldots \end{cases}$$

with q a quantization step size. During enrolment, X is measured and the Trusted Authority will construct a W, such that $X + W$ lies in the middle of the nearest quantization interval.[3] The value of W will be

$$W = \begin{cases} (2n + \frac{1}{2})q - X & \text{if } \rightleftharpoons \mathrm{s_i} = 1 \\ (2n - \frac{1}{2})q - X & \text{if } \rightleftharpoons \mathrm{s_i} = 0 \end{cases}$$

where $n = \ldots, -1, 0, 1, 2, \ldots$, is chosen such that $-q < W < q$. The value of n is discarded, but the values of W are used as helper data.

[3] This can be interpreted as a watermark of Quantization Index Modulation [14].

Using a more sophisticated quantization technique, we achieve the following results. We performed experiments on a wafer consisting of 71 ICs containing 21 measurement points per IC. The mean of the capacitance distribution is $\mu_c = 7.85\ 10^{-13}$ F and its variance $\sigma_c = 7.32\ 10^{-14}$ F. The noise distribution has a variance of $\sigma_n = 7.28\ 10^{-17}$ F. Using the above-mentioned key-extraction method, we derive 11 bits at an equal error rate of 0%.

2.4. Key Extraction for Acoustic PUFs

Key extraction of acoustic PUFs can be performed using the quantized key extraction method, explained in the Section 2.3. This leads to keys with an effective entropy of 20 bits. For more details we refer to [15].

3. PROTOCOLS

3.1. Identification and Authentication

The goal of an *identification* protocol is to check whether a specific PUF is present at the reader. As an example, one might think of a credit card or smart-card (equipped with a PUF) that is put into the reader by a user and has to be identified by the Verifier. The Verifier verifies the presence of the PUF in the card by checking the responses to one, or more challenges that were previously measured during enrolment.

We present a protocol for identification over an insecure channel.

3.1.1. Identification protocol

- **User:** Puts his card with PUF in the reader and claims its ID.
- **Verifier:** Randomly chooses a challenge C from his CRP database and sends it to the User together with the corresponding helper data W.
- **Reader:** Challenges the PUF with the Challenge C, measures the Response Y, and uses the helper data W to compute S'. Finally S' is sent back to the Verifier.
- **Verifier:** Checks whether S' equals the key S stored in his database. Then, he removes the triplet (C,W,S) from his database and never again uses these data for identification. Next time, another CRP is randomly chosen from his database.

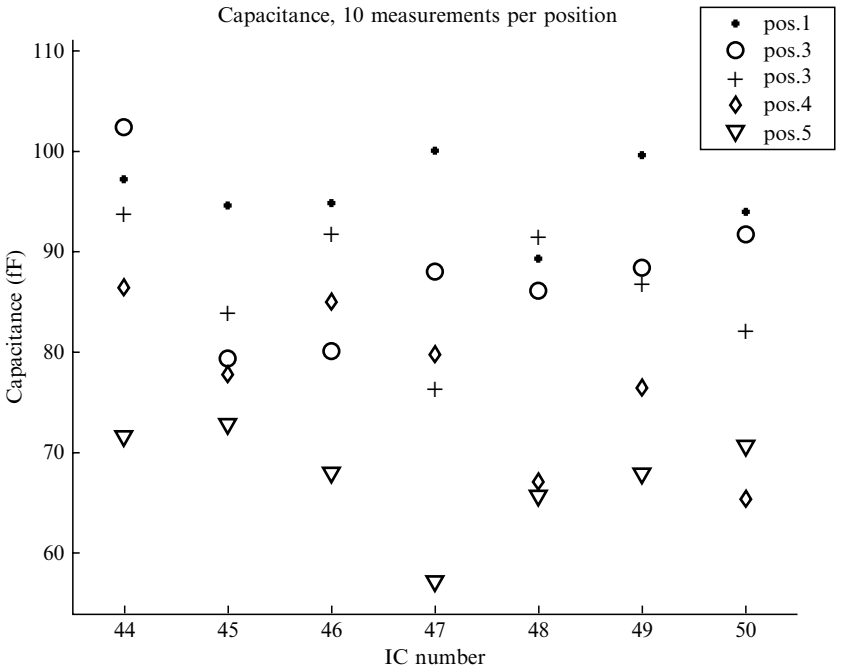


Figure 6.4-5. Measured capacitances on five positions on seven different ICs. Horizontal axis: IC number. Vertical axis: capacitance values.

We note that the security of this protocol relies on the fact that an attacker, who has seen (C_1, W_1, S_1) cannot predict the key S_2 that corresponds to the data (C_2,W_2), and on the fact that the PUF supports a large number of CRPs.

The goal of an *authentication* protocol is to ensure that the received messages originate from the stated sender, in particular from the device containing the PUF. The objective is to make the PUF generate the same key as the one that is stored at the Verifier's database. This key is used to MAC[4] (sign) messages.

3.1.2. Authentication protocol

- **User:** Puts his card with PUF in the reader and claims his ID.
- **Verifier:** Randomly chooses a challenge C from his CRP database and sends it to the User together with the corresponding helper data W and a random nonce m.

[4] MAC stands for message authentication code. It is used for signing a message, such that the receiver of the message can check whether the messages is coming from the legitimate person, and not from an imposter.

- **Reader:** Challenges the PUF with the Challenge C, measures the Response Y, and uses the helper data W to compute S'. Finally $M_{S'}(m)$ is sent back to the Verifier (where $M_{S'}(m)$ denotes a MAC on m, using the key S').
- **Verifier:** Computes $M_S(m)$ with the key S stored in his database and compares it with $M_{S'}(m)$. If both are equal, then $S = S'$ with very high probability. The key S is then used to MAC and/or encrypt all further messages.

The security of this scheme depends on the fact that (when the key S is unknown) the MAC $M_S(b)$ to a message b is unpredictable given that the attacker has seen the MAC to a message $a \neq b$ [16].

3.2. Renewal

During enrolment, a database of CRPs (with appropriate helper data) is created and stored by a Verifier. Since a PUF typically has many CRPs, the database might become very large. For a system with many users, the size of the database would become unacceptable. Therefore, it is desirable to start from a small number of CRPs and update those later on, when necessary. This is done with a *renewal protocol,* which is performed as (discussed later), before the Verifier has completely run out of CRPs.

3.2.1. Renewal

- **User and Verifier:** Run an authentication protocol, resulting in a secure authenticated channel (SAC) with session key S based on one of the last CRPs in the Verifier's database.
- **Verifier:** Randomly generates "*prechallenges*" $P_1, \ldots, P_n$. Then, he computes challenges $C_i = h(P_i)$, where h is a one-way function. If one of these newly generated challenges has been used before, it is discarded and a new one is generated. The prechallenges $P_1, \ldots, P_n$ are sent to the User over the SAC.
- **Reader:** Performs the one-way function h on each prechallenge, obtaining challenges $C_1, \ldots, C_n$, measures the responses $X_1, \ldots, X_n$, and sends them to the Verifier over the SAC.
- **Verifier:** Derives helper data $W_1, \ldots, W_n$ and keys $S_1, \ldots S_n$ from the responses and stores $\{h(P_i), W_i, S_i\}$.

The security of this protocol follows from the security of the session key that is agreed on during authentication. The challenges and responses

stay hidden to eavesdroppers. Note that the inclusion of the one-way function at the side of the reader is an effective measure against man-in-the-middle attacks (see Section 1.5.3).

3.3. Copy Protection

By embedding a PUF into a smart-card, credit card, RFID-Tag, value paper, etc., these objects become uncloneable, and their identity is checked by running one of the protocols mentioned above.

A PUF can also be used for copy protection purposes by embedding it into a data carrier, such as an optical disc. The attack that we prevent is the making of bit-by-bit copies of the optical disc. On the disc we store a table containing a list of $(C,W,h(S))$-triplets, where C denotes a challenge, W the helper data, S the key derived from the response to the challenge C, and h a one-way hash function. Furthermore, this table is signed with the *private key* of the Authority that is responsible for PUF enrolment. In order to make signature verification possible, each playback device is equipped with the Authority's Public Key.

When the disc is put into the player, the player checks the authenticity of the disc by running the following protocol:

(1) The player checks the signature on the list of $(C,W,h(S))$-triplets;
(2) The player randomly chooses a number of challenges from the list of triplets, measures the responses, uses the helper data W to compute S', and checks whether $h(S)=h(S')$; and
(3) If the check is ok, the player decides that the disc is authentic and plays the content. Otherwise, the player does not play the content stored on the disc.

In the first step, the player verifies that the list was created by the legitimate authority, by checking the signature. If the signature were not there, any attacker could produce a disc, embed his own PUF, and create the corresponding table. Since the attacker does not have the Authority's Private Key to sign the list properly, he is not able to put a valid list on the disc. Steps 2 and 3 then verify whether the list indeed corresponds to the embedded PUF. By using the same kind of technique, the content can also be encrypted with a key derived from the PUF embedded in the disc.

3.4. Special Protocols for CPUFs

The protocols described below are executed with Controlled PUFs. In [6], a CPUF was introduced, such that it can only be accessed by

programs and only by using two primitives called GetResponse and GetSecret. Let *F* denote the physical action of the PUF and *h* a one-way hash function, then, these primitives are defined as follows:

```
GetResponse(PreChallenge) =
  F(h(h(Program),PreChallenge))
GetSecret(Challenge) =
  h(h(Program),F(Challenge))
```

In these primitives, Program denotes the code of the program that will be executed in an authentic way. Right before starting the program, the CPUF computes h(Program) and later uses this value when the primitives GetResponse and GetSecret are invoked. The name "Response" refers to a direct PUF response, while "Secret" is a derived quantity.

3.4.1. Bootstrapping

In order to start any useful protocol with a CPUF, a user needs to have at least one CRP. The GetResponse and GetSecret primitives given above do not allow a user to obtain CRPs, if he does not already possess one. To allow for enrolment, a separate CPUF mode must exist; the "bootstrapping" mode, which is only executed at the Certification Authority (C.A.). In this mode, the CPUF outputs PUF responses in the clear, without applying one-way functions or encryptions. The existence of the bootstrapping mode is a potential security hazard. Ideally, the CPUF cannot be forced to reenter into this mode after enrolment has finished. This could be achieved (e.g., by burning through some components).

3.4.2. Certified execution

The motivation for *Certified Execution* is as follows:

Alice wants to run a computationally expensive program on Bob's computer. When she gets the result, she wants to have a proof that the computation was indeed performed on Bob's computer and that the result has not been faked. It will be explained how this can be achieved with CPUFs [6].

Definition

An e-certificate for program "Program" with input "Input" run on a processor P is a string that is efficiently generated by Program(Input) executed on P such that the user of P can efficiently check whether the output was indeed generated by Program(Input) on P.

Let *(C,X)* be a valid CRP for Bob's PUF. Alice sends a program (called "CProgram") to the CPUF, with the following two arguments

$$C, E \& M_K (Program, Input)$$

where $E \& M_K$ stands for encryption and MAC-ing with the key K,

$$K=h(h(CProgram), X).$$

The program Cprogram looks as follows:

```
(C, E&M) =Inputs;
  K=GetSecret (C);
  (Program, Input) = D & M_K (E&M);
Abort if the MAC does not match
Result = Program(Input);
Certificate = M_K (Result);
Output(Result, Certificate);
```

The fact that the result of the computation is MAC-ed with the shared secret K convinces Alice that the output has indeed been created by Bob's CPUF (Alice can compute the key K herself). Note, however, that the *e-certificate* has limited use. From the fact that it is based on a CRP shared between the CPUF and Alice, it follows that Alice can produce an *e-certificate* herself. This means that she cannot use an *e-certificate* as a proof towards third parties.

3.4.3. Proof of execution

An *e-proof* convinces third parties that a certain program was executed on a certain processor.

Definition: An e-proof "Eproof" for program "Program" with input "Input" on a processor P is a string efficiently generated by Program(Input) on P, such that there exists a protocol A between the processor P and any arbiter, with inputs Eproof and Program, and possibly some auxiliary information, which can efficiently:

(1) Decide correctly whether *e-proof* was generated by Program(Input) on P or not; and
(2) If correctly generated, decrypt the results "Result," which were generated together with e-proof by program(input) on P.

We present a CPUF-based solution that realizes the above-mentioned definition of an *e-proof*. To this end, we change a Program into one

having an *execution* and an *arbitration* mode. During the execution mode, the program generates not only the results of the computation but also the *e-proof*. The user (Alice) combines it with the *Certified Execution* technique to be sure that the program was correctly executed on Bob's CPUF. An arbiter can then check the *e-proof* by running the program in arbitration mode. There are two key ingredients. The first one is the fact that the `GetResponse` primitive depends on the hash of the complete program that contains both modes. The second one is that the encryption/MAC-key is a key that cannot be generated by Alice herself.

Execution mode

We consider the following setting: Alice wants to execute `Program(Input)` on Bob's PUF and wants to get an *e-proof* for the computation. Alice first modifies the program

```
Program(Input) into Eprogram(Inputs),
```

where

```
Inputs=(Program, Input, PC),Mode
```

with mode equal to "`execution mode`" and PC a random prechallenge. Eprogram is given by

```
begin program
 (X,Mode)=Inputs;
 If Mode is execution mode:
begin
 (Program, Input, PC)=X;
 Result=Program(Input);
 KE=GetResponse(PC);
 EMResult=E&M(Result,KE);
 EProof=(PC,EMResult);
 Results=(Result,EProof);
end
If Mode is arbitration mode:
begin
 /* See Section Arbitration */
 end
 Output(Results);
end program
```

Arbitration Mode

An arbiter who has received a proof of Execution Eproof from Alice prepares the input

```
Inputs=(EProof,Mode)
```

where Mode is equal to "arbitration mode" and executes the following Eprogram on Bob's CPUF.

```
begin program
  (X,Mode)=Inputs;
  If Mode is execution mode:
  begin
    /* See Section Execution Mode */
  end
  If Mode is arbitration mode:
  begin
    EProof=X;
    (PC,EMResult)=EProof;
    KA=GetResponse(PC);
    Result= D & M_KA(EMResult);
    CheckBit=(MAC of EMResult matches);
    Results=(Result,CheckBit);
  end
  Output(Results);
end program
```

The arbiter uses this technique in combination with *Certified Execution* to execute `Eprogram(Inputs)` on Bob's CPUF. The arbiter checks the *e-certificate* to verify the authenticity of the results that he gets back from the CPUF. The arbiter verifies whether the `Checkbit=True` (i.e., whether the MAC of `EMResult` matches). If so, the arbiter decides that `Program(Input)` on Bob's PUF has computed `Eproof` and `Result` in execution mode.

The above protocol is called an *arbiter protocol*. We note that an arbiter needs a valid CRP to be able to run the protocol. This can be taken care of by running a CRP management protocol with a Trusted Third Party, where the ownership of the PUF is registered and who distributes CRPs over a Secure Authenticated Channel.

4. CONCLUSIONS

In this chapter we have explained why security is important in an AmI environment. In order to achieve Trust and Security not only cryptographic algorithms are needed, but also secure methods for generation and storage of secret keys. Physical Uncloneable Functions can be used to this end because they have built in security properties, such as uncloneability and tamper evidence. We have given a generic algorithm to extract secure and reliable cryptographic keys from PUFs. The feasibility of this algorithm was shown in the case of Optical and Coating PUFs. Key generation from PUFs enables one to go beyond simple identification applications. In particular more advanced protocols, such as authentication, key exchange, certified execution, and Proofs of execution, become feasible.

ACKNOWLEDGMENTS

This work is the result from a pleasant and fruitful collaboration with the following people at Philips Research: Wil Ophey, Sjoerd Stallinga, Rob Wolters, Nynke Verhaegh, Arnold Gruijthuijsen, Petra de Jongh, Marten van Dijk, Geert-Jan Schrijen and with Blaise Gassend, Dwaine Clarke, and Srinivas Devadas at MIT.

REFERENCES

[1] Aarts, E. and Marzano, S., 2003, *The New Everyday: Views on Ambient Intelligence*, 010 Publishers, Rotterdam.
[2] Anderson, R. and Kuhn, M., 1996, Tamper resistance a cautionary note, *Proc. 2nd USENIX Workshop on Electronic Commerce*.
[3] Anderson, R. and Kuhn, M., Low cost attacks on tamper resistant devices, *in* M. Lomas (ed), *Security Protocols: 5th International Workshop*, Paris, France, LNCS, 1361, pp. 125–136.
[4] Pappu, R., Physical one-way functions, *Ph.D. thesis*, MIT, 2001.
[5] Pappu, R., Recht, B., Taylor, J. and Gershenfeld, N., 2002, Physical one-way functions, *Science*, **297,** 2026–2030.
[6] Gassend, B., et al., 2002, Controlled physical random functions, *Proc. 18th Annual Computer and Communications Security Applications Conf.*, December 2002.
[7] Gassend, B., et al., 2002, Silicon physical unkown functions, *Proc, 9th ACM Conf on Computer and Communications Security*, November 2002.

[8] Magnor, M., Dorn, P. and Rudolph, W., 2001, Simulation of confocal microscopy through scattering media with and without time gating, *J. Opt. Soc. Am. B*, **19**(11), 1695–1700.
[9] de Boer, J. F., Optical fluctuations on the transmission and reflections of mesoscopic systems, *Ph.D. thesis*, Amsterdam, 1995.
[10] Tuyls, P., Skoric, B., Ophey, W., Stallinga, S., Akkermans, A. H. M. *Information-Theoretic Security Analysis of Physical Uncloneable Functions*, In Patrick, A. P. and Yung, M., editors, Proceedings of Conference on Financial Cryptography and Data Security 2005, volume 3570 of Lecture notes in Computer Sciences pages 141–155, Springer–verlag, 2005.
[11] Tuyls, P. and Goseling, J., 2004, Capacity and examples of template protecting biometric authentication systems, *Biometric Authentication Workshop (BioAW 2004)*, LNCS 3087,Prague, pp. 158–170.
[12] Linnartz, J.-P. and Tuyls, P., 2003, New shielding functions to enhance privacy and prevent misuse of biometric templates, *4th International Conference on Audio- and Video-Based Biometric Person Authentication*.
[13] Dodis, Y., Reyzin, L. and Smith, A., 2004, Fuzzy extractors: How to generate strong secret keys from biometrics and other noisy data, *Advances in Cryptology – Eurocrypt'04*, LNCS 3027, pp. 523–540.
[14] Chen, B. and Wornell, G. W., 1998, Digital watermarking and information embedding using dither modulation, *IEEE Workshop on Multimedia Signal Processing*, Redondo Beach, CA.
[15] Vrijaldenhoven, S., *Acoustical Physical Uncloneable Functions*, Philips internal publication PR-TN-2004-300300.
[16] Gemmell, P. and Naor, M., 1994, Codes for interactive authentication, *Advances in Cryptology*, Crypto 93, LNCS 773.

Section 7

Conclusions

Chapter 7.1

CONCLUSIONS AND KEY CHALLENGES

Satyen Mukherjee
Philips Research North America
satyen.mukherjee@philips.com

Abstract Ambient intelligence (AmI) is seen as the next wave in the world of consumer electronics, where electronic systems in our living environment will cater to our needs and desires in a natural way. Many technologies come into play in providing the required solutions and many challenges are evident both in the hardware and software domains. Looking to the future, some of the key hardware challenges have been addressed in this book. This chapter builds on that and indicates 10 hardware related elements where major research challenges are foreseen.

Keywords AmI; extreme low power electronics; microsensors; microactuators; ubiquitous power

1. INTRODUCTION

From a long-term perspective, ambient intelligence (AmI) is about systems embedded in our environment, sensitive to the user's needs and emotions, that perform actions without explicit instructions from the user. The goal is to improve our quality of life and attend to our overall well-being. Many technologies come into play to achieve these objectives: sensing, storage, signal processing, communications (wired and wireless), displays, lighting, energy harvesting, and personal healthcare, among others. Digital technologies (both hardware and software), driven by the computer or data processing application domains, and made commercially viable by Moore's law, will continue to play a key role in this evolution. The crucial requirements of power efficiency in some areas and sophisticated sensing in others call for solutions that go beyond the domain of Moore's law

S. Mukherjee et al. (eds.), AmIware, 451–457.

(''more than Moore'' involving multiple technologies combined in system in packages).

In this new era of consumer electronics, advances in the underlying hardware technologies will propel the growth in AmI applications. Although much progress has been made, continual research efforts will be needed to address the ongoing challenges in 10 key areas:

(1) Microsensors,
(2) Ubiquitous Power Source,
(3) Extreme Low Power Electronics,
(4) Highly Efficient and Robust Wireless Communication Systems,
(5) Nonobtrusive and Flexible Displays,
(6) Self-configuring Networking Devices,
(7) Robust and Noise Immune Integrated Electronics,
(8) Embedded Processors,
(9) Microactuators, and
(10) Low Cost Electronic Solutions.

The following sections elaborate the listed challenging areas.

1.1. Microsensors

Sensors will play a critical role in many AmI applications. Starting with the sensing of ambient conditions, such as light level, temperature, humidity level, carbon monoxide and odor, presence detection at various levels of sophistication (image, well-being, mood, etc.) will be required. With the growing awareness of bioterrorism, radiation and chemical hazards, detection of gaseous chemical, and biological agents will become essential. In addition to basic functionality, these sensors will require high levels of accuracy to avoid false indications. Furthermore, the sensors should blend seamlessly with the surroundings and consume minimal power to allow unlimited lifetime.

Advanced MEMS technologies are expected to provide many solutions to a wide range of applications. MEMS devices make use of mechanical properties of the structures in conjunction with electronic sensing, processing, and control to accurately and reliably sense physical parameters, such as acceleration in automobile airbags. Microfluidic devices using MEMS structures are being developed to detect minute quantities of biological agents. These devices, when embedded in silicon, can be controlled by onboard electronics to provide local processing and communications to surrounding devices. Challenges in these applications lie in ensuring ease of use and long lifetime by means of self-cleaning and self-monitoring features. Furthermore, developing optimized embedded MEMS devices in mainstream CMOS is essential for low cost and

practical use. Therefore, for intelligent microsensors, more of Moore (System on Chip) as well as more than Moore (System in Package) will be required in developing complete solutions.

1.2. Ubiquitous Power Source

Powering the devices for AmI remains a major challenge as applications grow. Whether it is a smart sensor, an embedded processor, or an actuator, the meaningful solution should not require, for example, periodic maintenance or replacement of batteries. It will be essential to develop devices and techniques that allow harnessing energy from various available sources (light, temperature gradient, mechanical motion or vibration, air flow, pressure variation, etc.). The devices can be used either stand-alone, when sufficient power is available, or coupled with energy storage devices, such as rechargeable batteries or high-density capacitors. Additional long-life energy storage devices being investigated in the industry include microfuel cells, and even atomic energy storage devices. Coupled with energy scavenger and storage elements, high-efficiency power conversion circuits will be necessary to optimally drive the devices in question. Although much progress has been made in these areas, the maximum available power is still limited to a mW/cm^3, thereby limiting application options.

1.3. Extreme Low Power Electronics

Since providing adequate power to the devices is a challenge, it is obviously prudent to minimize power consumption in the embedded electronics using technologies that can operate with extreme low power. With mainstream scaled CMOS, power supply scaling and subthreshold circuit operations are being developed for both digital and analog and mixed signal circuits. There are many challenges here since scaling CMOS below a certain point (180 nm node) results in exponentially rising power loss due to leakage through the transistor gate insulators as well as through the short channels. Lowering supply voltage is the obvious remedy to this leakage related loss. This is appropriate in many AmI systems where performance (in terms of speed or data rate) can be traded for low power. In addition to design or circuit techniques, optimized process technologies, employing silicon on insulator (SOI) for instance, can allow an additional degree of freedom in low power solutions. Successful application of these techniques will require additional design tools and flows for reliable and reproducible designs. For large systems, modeling at different abstraction levels will be essential for practical utilization of these

techniques. Furthermore, for large circuits single chip solutions will be often less power efficient compared to 3D integrated multidie solutions (System in Package) due to losses in long interconnect.

1.4. Highly Efficient and Robust Wireless Communication Systems

As the number of devices in the environment grows, so do the communication needs and data size. With wireless communications using radio frequencies it is important to avoid interference among the devices. In residential applications, data rates range from tens of kilobits per second for status or control signals to a few Gbits per second for streaming high definition TV (HDTV) signals. The resulting interference among devices, as well as with neighboring residential locations, can pose major challenges. Communication architectures aimed at resolving some of these challenges include a judicious combination of wired backbone between regions (e.g., rooms in a residence) coupled with wireless within a region and adequate storage devices. Furthermore, allocation of specific frequency bands for different applications, along with standardization, help to mitigate the problem. At low data rates and moderate distances (20–250 Kbits/s), the Zigbee standard operating at 2.4 GHz is being pursued in the industry. For high data rates going up to 480 Mbits/s, ultrawide band (UWB) at 3–10 GHz is being proposed. The established WLAN employing 802.11a/b/g covering intermediate data rates and operating frequencies are prevalent. For multi-Gbits/s data rates (envisioned for WPAN), the 60 GHz frequency band is being explored. For many AmI applications, the challenge is compounded by the additional requirement of low power consumption in the wireless devices.

Implementation of these wireless communications systems will increasingly require multiple technologies for performance reasons, and will therefore require system in package in addition to system on chip solutions.

1.5. Nonobtrusive and Flexible Displays

Displays play a vital role in our interactions with the environment. Therefore displays will be a driver for many AmI applications where they can blend with the surroundings without compromising performance (image resolution, contrast, and brightness). In addition to the mobile flat and thin displays prevalent today, many applications will require large displays that are flat and flexible. Some applications will require rollable displays for flexibility and space saving reasons. The 3D displays will be particularly useful in providing a natural touch to the ambient.

Interactivity with the display will also be required in many applications. These interactions can be visual, audio, or even tactile. Displays and lighting will merge in many situations and compatible solutions for both will be necessary to enhance the impact of one on the other. Ambilight is an example of such a product where colored solid-state lighting around a flat display is controlled to match the images in the display, thereby enhancing viewer experience. Developments in polymer LEDs, Organic LEDs, and technologies, such as electrophoretic (e-ink) and electrowetting along with LCDs, will be helpful in finding many of the solutions. A multidisciplinary approach will be necessary to find ways to combine technologies for large area electronics and varied display effects. Large area electronic technologies will include polymer or plastic electronics along with organic semiconductor devices. These technologies will also be combined with well-established silicon technologies or, in many cases, amorphous and polysilicon thin film transistors. Reliability and lifetime of the devices remain the main challenges today.

Displays are typically implemented using technologies, termed broadly as large area electronics, which can be seen as an extension of Moore's law in 2D. In other words, techniques from the conventional Moore's law domain (lithography driven) will be deployed to a large extent. However, for cost and efficiency reasons other approaches, such as screen-printing, will become increasingly important.

1.6. Self-Configuring Networking Devices

With a myriad of devices or nodes in the ambient, it will be essential to develop techniques that allow easy setup of the AmI system. The nodes may be stationary or nomadic, depending upon the function and implementation flexibility requirement. The entire network formed of wired and wireless interconnections will set constraints on the devices at both hardware and software levels. Depending upon the application setting, the nodes may require peer-to-peer communication for low power consumption or communication via a few centralized master nodes or a combination of the two. The choice will in turn define the requirements in hardware architecture and building blocks. The introduction of a new device in the environment should be seamless. Self-configuration, with minimal user intervention, will be essential for practical applications. Run-time adaptive design concepts will be required in many situations. Wireless technologies and regulatory standards will generate additional constraints. From a marketing perspective, globalization of the communications standards will help promote economy of scale with new products.

1.7. Robust and Noise Immune Integrated Electronics

Many embedded devices in the AmI systems will face conflicting requirements: low power consumption on the one hand and robustness and reliability on the other. The classic trade-off between power consumption and signal to noise ratio (SNR) may not suffice in many situations where only limited power is available and a long lifetime is required, without compromising accuracy or precision. Therefore, new low power design techniques will be required to address error correction techniques in a noisy environment. Very low voltage and low power sensor-interface analog electronics, coupled with digital electronics to correct the errors created by the low SNR will need to be investigated to allow optimization of the power and SNR trade-off.

Packaging techniques such as 3D integration with appropriate partitioning of the different building blocks (digital and mixed signal or analog) can be used effectively to achieve improved noise immunity in the overall system in many implementations.

1.8. Embedded Processors

The proliferation of smart sensors and communication networks required in the AmI systems will drive the need for embedded processors with a wide range of performance requirements. A key question to be addressed is whether a platform approach or a particular point solution is appropriate in a given application. Performance, power, and cost will be some of the parameters that will need to be considered. Flexibility as well as time to market will play a role in choosing the proper solution. A wide range of communications standards will need to be introduced where software defined radio (SDR) will bring the most flexibility. In certain applications, such as face tracking, the processors will need to operate at very high operations per second requiring parallel processing in some form (Single Instruction and Multiple Data—SIMD or Multiple Instruction and Multiple Data—MIMD). Vector processors can provide the flexibility required to address challenges in low power implementation, compilation, scalability, and ability to handle a wide range of algorithms.

1.9. Microactuators

While sensing and gathering information from the ambient is a key element, providing information to the user is another vital element in most AmI applications. The information feedback can be visual, audio, sensory or olfactory, or even mental. In each case, devices need to be developed that blend into the surroundings. Furthermore microactuators in many cases can perform a function for certain types of sensing applications—for

example, microfluidic devices for biosensing and analysis. MEMs-based actuators, such as micromotors, can play a major role in AmI systems and affect our surroundings in unprecedented ways. Similarly, microrobots are candidates for a variety of applications in the manipulation of material. Microactuation can be based on a number of physical phenomena, such as electromagnetic, electrostatic, thermomechanical, phase change, piezoelectric, shape memory, magnetostrictive, electrorheological, electrohydrodynamic, and diamagnetism. Challenges lie in bringing some of these techniques up to manufacturing levels and transforming others from the lab demonstration phase to functional models that perform effectively in an AmI setting.

1.10. Low Cost Electronic Solutions

To make a lasting impact on our lives, AmI applications will need to be widespread like the lightbulb or cell phone. This is possible only if the solutions are low enough in cost to make applications affordable for large segments of the population.

While Moore's law has provided the roadmap for cost reduction in the digital IC world (mainstream CMOS technologies), which is clearly a major component in AmI solutions, many applications require electronic breakthroughs that go beyond the domains of Moore's law. For instance, wireless technologies require front-end electronic solutions that are typically analog in nature and very often employ passive components beyond the capabilities of mainstream CMOS technologies. The industry is pursuing two approaches to address this issue. One is the enhancement of the mainstream CMOS processes to include additional components in the back end (MEMS and other similar devices), thereby allowing single system on chip (SoC) solutions and hopefully lower cost. The other approach, also termed "More than Moore," involves novel packaging techniques resulting in system in package (SiP). There are several approaches to system in package including multichip modules, sophisticated 3D integration, passive integration, and inclusion of biological elements. In all of these, cost reduction is a major driving factor. Continued research is required in these areas to develop new materials (perhaps nanoparticle-based), devices and techniques to allow low cost packaging solutions while preserving functionality and performance.

With advances in each of the above areas the implementation gap between the AmI system concepts and hardware devices shown in Figure–1 (Preface) will be bridged effectively. From the breadth of the technical areas involved, it is evident that it will be essential to apply multidisciplinary approaches to achieve meaningful solutions that will enhance the AmIware technologies.

Subject Index

Author Index