ULTRA LOW-POWER ELECTRONICS AND DESIGN

Ultra Low-Power Electronics and Design

Edited by

Enrico Macii
Politecnico di Torino,
Italy

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 1-4020-8076-X
Print ISBN: 1-4020-8075-1

Created in the United States of America

Visit Springer's eBookstore at: http://www.ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

Contents

Contributors

A. Acquaviva	Università di Urbino
L. Benini	Università di Bologna
D. Bertozzi	Università di Bologna
D. Blaauw	University of Michigan, Ann Arbor
A. Bogliolo	Università di Urbino
A. Bona	STMicroelectronics
C. Brandolese	Politecnico di Milano
W.C. Cheng	University of Southern California
G. De Micheli	Stanford University
N. Dutt	University of California, Irvine
W. Fornaciari	Politecnico di Milano
F. Gaffiot	Ecole Centrale de Lyon
J. Gautier	CEA-DRT–LETI/D2NT–CEA/GRE
A. Gordon-Ross	University of California, Riverside
R. Gupta	University of California, San Diego
C. Heer	Infineon Technologies AG
M. J. Irwin	Pennsylvania State University
I. Kadayif	Canakkale Onsekiz Mart University
M. Kandemir	Pennsylvania State University
B. Kienhuis	Leiden
I. Kolcu	UMIST
E. Lattanzi	Università di Urbino
D. Lee	University of Michigan, Ann Arbor
A. Macii	Politecnico di Torino
S. Mohapatra	University of California, Irvine
I. O'Connor	Ecole Centrale de Lyon
K. Patel	Politecnico di Torino
M. Pedram	University of Southern California
C. Pereira	University of California, San Diego
C. Piguet	CSEM
M. Poncino	Università di Verona
F. Salice	Politecnico di Milano
P. Schaumont	University of California, Los Angeles
U. Schlichtmann	Technische Universität München
D. Sylvester	University of Michigan, Ann Arbor

F. Vahid	University of California, Riverside and University of California, Irvine
N. Venkatasubramanian	University of California, Irvine
I. Verbauwhede	University of California, Los Angeles and K.U.Leuven
N. Vijaykrishnan	Pennsylvania State University
V. Zaccaria	STMicroelectronics
R. Zafalon	STMicroelectronics
B. Zhai	University of Michigan, Ann Arbor
C. Zhang	University of California, Riverside

Preface

Today we are beginning to have to face up to the consequences of the stunning success of Moore's Law, that astute observation by Intel's Gordon Moore which predicts that integrated circuit transistor densities will double every 12 to 18 months. This observation has now held true for the last 25 years or more, and there are many indications that it will continue to hold true for many years to come. This book appears at a time when the first examples of complex circuits in 65nm CMOS technology are beginning to appear, and these products already must take advantage of many of the techniques to be discussed and developed in this book. So why then should our increasing success at miniaturization, as evidenced by the success of Moore's Law, be creating so many new difficulties in power management in circuit designs?

The principal source and the physical origin of the problem lies in the differential scaling rates of the many factors that contribute to power dissipation in an IC – transistor speed/density product goes up faster than the energy per transition comes down, so the power dissipation per unit area increases in a general sense as the technology evolves.

Secondly, the "natural" transistor switching speed increase from one generation to the next is becoming downgraded due to the greater parasitic losses in the wiring of the devices. The technologists are offsetting this problem to some extent by introducing lower permittivity dielectrics ("low-k") and lower resistivity conductors (copper) – but nonetheless to get the needed circuit performance, higher speed devices using techniques such as silicon-on-insulator (SOI) substrates, enhanced carrier mobility ("strained silicon") and higher field ("overdrive") operation are driving power densities ever upwards. In many cases, these new device architectures are increasingly leaky, so static power dissipation becomes a major headache in power management, especially for portable applications.

A third factor is system or application driven – having all this integration capability available encourages us to combine many different functional blocks into one system IC. This means that in many cases, a large part of the chip's required functionality will come from software executing on and between multiple on-chip execution units; how the optimum partitioning between hardware architecture and software implementation is obtained is a vast subject, but clearly some implementations will be more energy efficient than others. Given that, in many of today's designs, more than 50% of the total development effort is on the software that runs on the chip, getting this partitioning right in terms of power dissipation can be critical to the success of (or instrumental in the failure of!) the product.

A final motivation comes from the practical and environmental consequences of how we design our chips – state-of-the-art high performance circuits are dissipating up to 100W per square centimeter – we only need 500 square meters of such silicon to soak up the output of a small nuclear power station. A related argument, based on battery lifetime, shows that the "converged" mobile phone application combining telephony, data transmission, multimedia and PDA functions that will appear shortly is demanding power at the limit of lithium-ion or even methanol-water fuel cell battery technology. We have to solve the power issue by a combination of design and process technology innovations; examples of current approaches to power management include multiple transistor thresholds, triple gate oxide, dynamic supply voltage adjustment and memory architectures.

Multiple transistor thresholds is a technique, practiced for several years now, that allows the designer to use high performance (low Vt) devices where he needs the speed, and low leakage (high Vt) devices elsewhere. This benefits both static power consumption (through less sub-threshold leakage) and dynamic power consumption (through lower overall switching currents). High threshold devices can also be used to gate the supplies to different parts of the circuit, allowing blocks to be put to sleep until needed.

Similar to the previous technique, triple gate oxide (TGO) allows circuit partitioning between those parts that need performance and other areas of the circuit that don't. It has the additional benefit of acting on both sub-threshold leakage and gate leakage. The third oxide is used for I/O and possibly mixed-signal. It is expected over the next few years that the process technologists will eventually replace the traditional silicon dioxide gate dielectric of the CMOS devices by new materials such as rare earth oxides with much higher dielectric constants that will allow the gate leakage problem to be completely suppressed.

Dynamic supply voltage adjustment allows the supply voltage to different blocks of the circuit to be adjusted dynamically in response to the immediate performance needs for the block – this very sophisticated technique will take some time to mature.

Finally, many, if not most, advanced devices use very large amounts of memory for which the contents may have to be maintained during standby; this consumes a substantial amount of power, either through refreshing dynamic RAM or through the array leakage for static RAM. Traditional non-volatile memories have writing times that are orders of magnitude too slow to allow them to substitute these on-chip memories. New developments, such as MRAM, offer the possibility of SRAM-like performance coupled with unlimited endurance and data retention, making them potential candidates to replace the traditional on-chip memories and remove this component of standby power consumption.

Most of the approaches to power management described briefly above will be employed in 65nm circuits, but there are a lot more good ideas waiting to be applied to the problem, many of which you will find clearly and concisely explained in this book.

Mike Thompson, Philippe Magarshack

STMicroelectronics, Central R&D
Crolles, France

Introduction

ULTRA LOW-POWER ELECTRONICS AND DESIGN

Enrico Macii
Politecnico di Torino

Power consumption is a key limitation in many electronic systems today, ranging from mobile telecom to portable and desktop computing systems, especially when moving to nanometer technologies. Power is also a showstopper for many emerging applications like ambient intelligence and sensor networks. Consequently, new design techniques and methodologies are needed to control and limit power consumption.

The 2004 edition of the DATE (Design Automation and Test in Europe) conference has devoted an entire Special Focus Day to the power problem and its implications on the design of future electronic systems. In particular, keynote presentations and invited talks by outstanding researchers in the field of low-power design, as well as several technical papers from the regular conference sessions have addressed the difficulties ahead and advanced strategies and principles for achieving ultra low-power design solutions. Purpose of this book is to integrate into a single volume a selection of these contributions, duly extended and transformed by the authors into chapters proposing a mix of tutorial material and advanced research results.

The manuscript consists of a total of 14 chapters, addressing different aspects of ultra low-power electronics and design. Chapter 1 opens the volume by providing an insight to innovative transistor devices that are capable of operating with a very low threshold voltage, thus contributing to a significant reduction of the dynamic component of power consumption. Solutions for limiting leakage power during stand-by mode are also discussed. The chapter closes with a quick overview of low-power design techniques applicable at the logic level, including multi-V_{dd}, multi-V_{th} and hybrid approaches.

Chapter 2 focuses on the problem of reducing power in the interconnect network by investigating alternatives to traditional metal wires. In fact, according to the 2003 ITRS roadmap, metallic interconnections may not be able to provide enough transmission speed and to keep power under control for the upcoming technology nodes (65nm and below). A possible solution, explored in the chapter, consists of the adoption of optical interconnect networks. Two applications are presented: Clock distribution and data communication using wavelength division multiplexing.

In Chapter 3, the power consumption problem is faced from the technology point of view by looking at innovative nano-devices, such as single-electron or few-electron transistors. The low-power characteristics and potential of these devices are reviewed in details. Other devices, including carbon nano-tube transistors, resonant tunnelling diodes and quantum cellular automata are also treated.

Chapter 4 is entirely dedicated to advanced design methodologies for reducing sub-threshold and gate leakage currents in deep-submicron CMOS circuits by properly choosing the states to which gates have to be driven when in stand-by mode, as well as the values of the threshold voltage and of the gate oxide thickness. The authors formulate the optimization problem for simultaneous state/V_{th} and state/V_{th}/T_{ox} assignments under delay constraints and propose both an exact method for its optimal solution and two practical heuristics with reasonable run-time. Experimental results obtained on a number of benchmark circuits demonstrate the viability of the proposed methodology.

Chapter 5 is concerned with the issue of minimizing power consumption of the memory subsystem in complex, multi-processor systems-on-chip (MPSoCs), such as those employed in multi-media applications. The focus is on design solutions and methods for synthesizing memory architectures containing both single-ported and multi-ported memory banks. Power efficiency is achieved by casting the memory partitioning design paradigm to the case of heterogeneous memory structures, in which data need to be accessed in a shared manner by different processing units.

Chapter 6 addresses the relevant problem of minimizing the power consumed by the cache hierarchy of a microprocessor. Several design techniques are discussed, including application-driven automatic and dynamic cache parameter tuning, adoption of configurable victim buffers and frequent-value data encoding and compression.

Power optimization for parallel, variable-voltage/frequency processors is the subject of Chapter 7. Given a processor with such an architecture, this chapter investigates the energy/performance tradeoffs that can be spanned in parallelizing array-intensive applications, taking into account the possibility that individual processing units can operate at different voltage/frequency levels. In assigning voltage levels to processing units, compiler analysis is used to reveal hetherogeneity between the loads of the different units in parallel execution.

Chapter 8 provides guidelines for the design and implementation of DSP and multi-media applications onto programmable embedded platforms. The RINGS architecture is first introduced, followed by a detailed discussion on power-efficient design of some of the platform components, namely, the DSPs. Next, design exploration, co-design and co-simulation challenges are addressed, with the goal of offering to the designers the capability of including into the final architecture the right level of programmability (or reconfigurability) to guarantee the required balance between system performance and power consumption.
Chapter 9 targets software power minimization through source code optimization. Different classes of code transformations are first reviewed; next, the chapter outlines a flow for the estimation of the effects that the application of such transformations may have on the power consumed by a software application. At the core of the estimation methodology there is the development of power models that allow the decoupling of processor-independent analysis from all the aspects that are tightly related to processor architecture and implementation. The proposed approach to software power minimization is validated through several experiments conducted on a number of embedded processors for different types of benchmark applications.

Reduction of the power consumed by TFT liquid crystal displays, such as those commonly used in consumer electronic products is the subject of Chapter 10. More specifically, techniques for reducing power consumption of transmissive TFT-LCDs using a cold cathode fluorescent lamp backlight are proposed. The rationale behind such techniques is that the transmittance function of the TFT-LCD panel can be adjusted (i.e., scaled) while meeting an upper bound on a contrast distortion metric. Experimental results show that significant power savings can be achieved for still images with very little penalty in image contrast.

Chapter 11 addresses the issue of efficiently accessing remote memories from wireless systems. This problem is particularly important for devices such as palmtops and PDAs, for which local memory space is at a premium and networked memory access is required to support virtual memory swapping. The chapter explores performance and energy of network swapping in comparison with swapping on local microdrives and FLASH memories. Results show that remote swapping over power-manageable wireless network interface cards can be more efficient than local swapping and that both energy and performance can be optimized by means of power-aware reshaping of data requests. In other words, dummy data accesses can be preemptively inserted in the source code to reshape page requests in order to significantly improve the effectiveness of dynamic power management.

Chapter 12 focuses on communication architectures for multi-processor SoCs. The network-on-chip (NoC) paradigm is reviewed, touching upon several issues related to power optimization of such kinds of communication architectures. The analysis goes on a layer-by-layer basis, and particular emphasis is given to customized, domain-specific networks, which represent the most promising scenario for communication-energy minimization in multi-processor platforms.

Chapter 13 provides a natural follow up to the theory of NoCs covered in the previous chapter by describing an industrial application of this type of communication architecture. In particular, the authors introduce an innovative methodology for automatically generating the power models of a versatile and parametric on-chip communication IP, namely the STBus by STMicroelectronics. The methodology is validated on a multi-processor hardware platform including four ARM cores accessing a number of peripheral targets, such as SRAM banks, interrupt slaves and ROM memories.

The last contribution, offered in Chapter 14, proposes an integrated end-to-end power management approach for mobile video streaming applications that unifies low-level architectural optimizations (e.g., CPU, memory, registers), OS power-saving mechanisms (e.g., dynamic voltage scaling) and adaptive middleware techniques (e.g., admission control, trans-coding, network traffic regulation). Specifically, interaction parameters between the different levels are identified and optimized to achieve a reduction in the power consumption.

Closing this introductory chapter, the editor would like to thank all the authors for their effort in producing their outstanding contributions in a very short time. A special thank goes to Mike Thompson and Philippe Magarshack of STMicroelectronics for their keynote presentation at DATE 2004 and for writing the foreword to this book. The editor would also like to acknowledge the support offered by Mark De Jongh and the Kluwer staff during the preparation of the final version of the manuscript. Last, but not least, the editor is grateful to Agnieszka Furman for taking care of most of the “dirty work” related to book editing, paging and preparation of the camera-ready material.