



UNIVERSITÉ CATHOLIQUE DE LOUVAIN  
ÉCOLE POLYTECHNIQUE DE LOUVAIN  
DÉPARTEMENT D'ÉLECTRICITÉ

MICROELECTRONICS LABORATORY

## **Pushing Ultra-Low-Power Digital Circuits into the Nanometer Era**

**David Bol**

Thesis submitted in partial fulfillment  
of the requirements for the degree of  
*Docteur en sciences appliquées*

Dissertation committee:

Prof. Jean-Didier Legat (Microelectronics Laboratory, UCL), advisor  
Prof. Denis Flandre (Microelectronics Laboratory, UCL),  
Prof. Jean-Pierre Raskin (Microwave Laboratory, UCL),  
Prof. Christian Piguet (CSEM, Neuchâtel - EPFL, Lausanne),  
Prof. Kaushik Roy (Purdue University, West Lafayette),  
Prof. Andrei Vladimirescu (University of California, Berkeley - ISEP, Paris),  
Prof. Luc Vandendorpe (Electricity Department, UCL), President

December 2008



---

# PUSHING ULTRA-LOW-POWER DIGITAL CIRCUITS INTO THE NANOMETER ERA

---

**David Bol**

Ecole Polytechnique de Louvain  
Microelectronics laboratory



Université catholique de Louvain  
Louvain-la-Neuve (Belgium)



*À mon grand-père*



# CONTENTS

---

Acknowledgments	xi
Abstract	xiii
Acronyms	xv
List of notations	xvii
<b>Introduction</b>	<b>xxi</b>
I.1 Moore's law and technology scaling	xxii
I.2 Ultra-low-power applications	xxiii
I.3 Frequency/voltage-scaled subthreshold digital circuits	xxiii
I.4 Thesis outline	xxv
Author's publication list	xxxi
<b>1 Power and Energy Consumption of FVS Digital Circuits</b>	<b>1</b>
1.1 Introduction	3
1.2 Constraints on digital circuits	3
1.2.1 Robustness constraint	3
1.2.2 Throughput constraint	5
1.3 Sources of power and energy consumption	7
1.3.1 Dynamic power consumption	7
1.3.2 Static power dissipation	8
1.3.3 Energy per operation	13
1.4 Practical power and energy under robustness and throughput constraints	14
1.4.1 Frequency/voltage scaling scheme	14
1.4.2 Comparison with classic operating schemes	17
1.4.3 Comparison with sleep-mode operating scheme	18
1.4.4 Impact of the temperature	20
1.4.5 Impact of circuit/application parameters	20
1.5 Conclusion	23
	<b>vii</b>

<b>2</b>	<b>Impact of Technology Scaling on Subthreshold Logic</b>	<b>27</b>
2.1	Introduction	29
2.2	Technology scaling	30
2.2.1	Scaling theory	30
2.2.2	Technology scaling in the nanometer era	30
2.2.3	Considered device models	31
2.3	Impact on MOSFET subthreshold operation	32
2.3.1	Subthreshold region of operation	32
2.3.2	Subthreshold drain current	33
2.3.3	Capacitances in subthreshold regime	35
2.3.4	Variability in subthreshold regime	38
2.4	Impact on subthreshold logic	40
2.4.1	Static noise margins and functional yield	41
2.4.2	Subthreshold circuit delay	44
2.4.3	Dynamic power consumption	47
2.4.4	Static power consumption	47
2.4.5	Minimum-energy point	48
2.4.6	Practical power and energy under robustness and throughput constraints	50
2.5	Results validation	54
2.6	Conclusion	54
<b>3</b>	<b>Optimum Nanometer Devices and Technologies for Minimum-Energy Subthreshold Circuits</b>	<b>61</b>
3.1	Introduction	63
3.2	Background and related work	64
3.2.1	Minimum-energy point modeling	64
3.2.2	Device optimization for minimum energy	65
3.3	Pre-Silicon bulk MOSFET compact models for subthreshold circuit simulation	66
3.4	Limitations from nanometer MOSFET effects	69
3.5	Impact of nanometer MOSFET parameters	71
3.5.1	Gate length impact	71
3.5.2	Threshold voltage impact	73
3.5.3	Oxide thickness impact	73
3.6	Optimum technology and device selection	74
3.6.1	Technology flavor comparison	74
3.6.2	Optimum device selection	78



3.7	Fully-depleted SOI technology	80
3.7.1	Pre-Silicon FD SOI MOSFET compact models	80
3.7.2	Subthreshold characteristics of FD SOI MOSFETs	85
3.7.3	Minimum-energy subthreshold circuits in FD SOI technology	86
3.8	Conclusion	88
<b>4</b>	<b>Design Choices for Practical Energy Minimization in Nanometer Subthreshold Circuits</b>	<b>93</b>
4.1	Introduction	95
4.2	Technology selection	96
4.2.1	Bulk vs. FD SOI	96
4.2.2	Technology flavor selection	96
4.2.3	MOSFET selection	99
4.2.4	Independent dual- $V_t$ assignment	100
4.2.5	Discussion	104
4.3	Body biasing for circuit adaptation	105
4.3.1	Impact of global process/temperature corners	105
4.3.2	Effects of body bias on subthreshold MOSFET operation	107
4.3.3	Impact of body bias on practical energy	108
4.3.4	Circuit adaptation	111
4.3.5	Discussion	112
4.4	Sleep-mode techniques	114
4.4.1	Impact of dynamic reverse body biasing on practical energy	115
4.4.2	Impact of power gating on practical energy	117
4.4.3	Discussion	123
4.5	Conclusion	125
<b>5</b>	<b>Building Ultra-Low-Power High-Temperature Digital Circuits in Standard SOI Technology</b>	<b>129</b>
5.1	Introduction	131
5.2	High-temperature MOSFET behavior	131
5.3	ULP transistor	132
5.3.1	Principle	133
5.3.2	Leakage reduction mechanism	133
5.3.3	$I_D/V_{GS}$ characteristics	135
5.4	ULP logic style	137

5.4.1	Architecture and layout	137
5.4.2	DC behaviour	138
5.4.3	Impact of intrinsic variability on robustness	139
5.4.4	Performance evaluation	142
5.5	Impact of PVT variations on performances	144
5.5.1	Process variations	144
5.5.2	Voltage variations	145
5.5.3	Temperature variations	146
5.6	Validation of ULP logic style	146
5.6.1	Measurement of ring-oscillator test vehicle	147
5.6.2	Simulation of a benchmark multiplier	148
5.6.3	Comparison with other leakage-reduction techniques	148
5.7	Conclusion	150
<b>Conclusions and perspectives</b>		<b>153</b>
Postface		161
<b>Appendix A: Roadmap for Nanometer Ultra-Low-Power Circuits</b>		<b>163</b>
A.1	Technology/circuit specifications for optimum subthreshold circuits	163
A.2	A possible technology/circuit roadmap for nanometer ultra-low-power circuits	166
<b>Appendix B: Description of the Circuit Simulation Benchmark</b>		<b>171</b>
B.1	8-bit RCA benchmark multiplier	171
B.2	Simulation setup	171
<b>Appendix C: BSIM4 Pre-Silicon Nanometer MOSFET Model Cards</b>		<b>175</b>

# ACKNOWLEDGMENTS

---

As you might guess, carrying out a Ph.D research project alone in front of his desk is a difficult task. When it comes to high-end technological topics such as nanoelectronics, it becomes pretty impossible without the help of several key people. Therefore, I would like to thank them here for their support. I apologize in advance for the ones I might forget.

First of all, I would like to thank my advisor Prof. Jean-Didier Legat not only for giving me the opportunity to live this adventure but also for actually convincing me that I had to take my chance on the Ph.D cursus. I would like to thank him too for giving me this (controlled) freedom and encouraging me to do my own choices, even if sometimes they turned out to be dead ends. I definitively learned a lot that way.

I am so grateful to Prof. Denis Flandre for his continuous guidance during the last two years and his tough review of my papers. I think some of them would not have been accepted without his suggestions. I am further very grateful to him for finding an opportunity for me to spend several months under the Spanish sun at Seville, for offering me to present tutorials in interesting conferences and to patent the ULP logic style.

I also would like to express my thanks to Prof. Jean-Pierre Raskin, who accepted to be part of my guidance committee despite the gap between our research topics, for the interesting and entertaining discussions.

My gratitude also goes to the professors that accepted to be part of my examination committee: Prof. Christian Piguët (I am glad to see that we have reached some kind of agreement on the interests of nanometer technologies for ULP applications), Prof. Kaushik Roy and Prof. Andrei Vladimirescu (thank you for the wine selection!). I would like to thank them for their valuable comments and the tough but nice discussion at the private defense, I really enjoyed it. Finally, I am grateful to Prof. Luc Vandendorpe who did a great job chairing the private defense.

Next, my thanks go to my co-authors mainly for their patience regarding my fussiness about expressing things right (= my way) and plotting nice graphics (= my taste), especially when I was not the first author: María José Avedillo, Marc Baltus, Laurent Demeûs, Julien De Vos, Ilham Hassoune, Dina Kamel, David Levacq, Philippe Manet, Gueric Meurice, José María Quintana, Jean-Jacques Quisquater, Cesar Roda Neve (this is especially true for you!). Amongst them, I owe special thanks to Dina Kamel who took a whole night correcting parts of this dissertation under pressure.

I also would like to thank my colleagues from the SPARC and SOI groups at UCL Microelectronics lab (past and present) for the nice atmosphere (and the

long layout nights before tape-out deadlines for some of them): Aryan Afzalian, Nicolas André, Olivier Bulteel, Thibault Delavallée, Sylvain Druart, Majid El Kaamouchi and Mostafa Emam (Microwave lab), Geoffroy Gosset, Igor Loïselle, Angelo Kuti Lusala, Valeriya Kilchytska, Rémi Pampin, Guillaume Pollissard, François Mace, Luis Moreno, Bertrand Rousseau, Bertrand Rue, Laurent Vancaille. Many thanks to the members of all other groups from the Microelectronics and Microwave labs, especially (but not limited to) Stéphane Burignat, Laurent Francis, Pierre-Olivier Mouthuy, Olivier Pereira, François-Xavier Standaert, Dana Serban and all the professors. I also address special thanks to Frédéric Vrins for his nice Latex template. I am grateful to the administrative and technical staffs for their help: Anne Adant, Sylvie Baudine, André Crahay and the whole clean-room team, Isabelle Dargent, Emmanuel De Pauw, Christel Derzelle, Christian Renaux, Viviane Sauvage and Pascal Simon. I particularly thank Brigitte Dupont for the hardware and software computer support regarding EDA tools, which was always quick and friendly even when I needed a tool to be installed within the next 15 minutes.

I am grateful to the Fonds National de la Recherche Scientifique of Belgium and Walloon Region for funding this research.

Je voudrais également profiter de cette occasion pour remercier (en français) mes parents pour leur soutien - moral pendant mes 4 années de thèse, et logistique pendant les derniers jours de rédaction. Merci aussi aux autres membres de ma famille et belle-famille ainsi que mes amis pour leur enthousiasme (avant), leurs encouragements (pendant) et leurs félicitations (après).

Je remercie tout spécialement la femme de ma vie, Céline, qui, sans totalement comprendre l'objet de cette thèse, a su percevoir, mieux que quiconque, que mon épanouissement professionnel passait par l'accomplissement d'une recherche dont je serais fier. Merci à elle de m'avoir encouragé, dans les moments de remise en question, à faire les choix qui me satisfaisaient même s'ils impliquaient quelques centaines d'heures supplémentaires.

Last but certainly not least, je voudrais finalement remercier mon collègue (ex-et futur), co-auteur, co-organisateur de conférence, co-designer, co-mémorant, camarade de cours, co-koteur, témoin de mariage mais avant tout ami, Renaud Ambroise, pour avoir partagé tant de discussions sur le pourquoi du comment de l'électronique digitale (aaah les facteurs de mérite!), des effets canal court et autres courants de fuite, du bruit de substrat, des outils de conception et *design kits*, des sauts en BMX et *Questions pour un Champion*. Merci à lui d'avoir pris spontanément tant de tâches (parfois ingrates) à sa charge pendant cette dernière année pour me permettre de terminer cette thèse dans des délais acceptables. Sans lui, j'en serais encore au chapitre 3. Mille fois, merci!

David

# ABSTRACT

---

Over the last decade, ultra-low-power (ULP) design of integrated circuits has become a vibrant research field for emerging applications such as sensor networks, biomedical devices or RFID tags. In these applications, the circuits typically feature a low computational load but need to operate for a long time on small batteries or to harvest power from the environment. The energy and power consumptions are thus the main figures of merit.

Owing to their low computational load, ULP applications require low-to-medium data/operation throughputs (10k - 10MOp/s). Power consumption of CMOS digital circuits for these applications is thus minimized through joint scaling of the clock frequency  $f_{clk}$  and the supply voltage  $V_{dd}$  to the functional and speed limits, whereas energy per operation is minimized when lowering  $V_{dd}$  to the so-called minimum-energy point. Operating at the minimum-energy point provides minimum energy level by balancing dynamic energy due to capacitance switching and static energy due to the integration of leakage currents over the execution time of the operation, equal to the circuit delay. This often occurs for  $V_{dd}$  values ranging from 0.2 to 0.4V. At these voltages, MOSFET devices operate in subthreshold regime, i.e. the on- and off-state drain currents are subthreshold currents, which exponentially depend on the gate bias and on the threshold voltage. Under these conditions, digital circuits are called subthreshold logic circuits.

At the same time, Moore's-law-driven technology scaling leads to the development of nanoscale CMOS processes featuring severe drawbacks such as high leakage currents, short-channel effects and device variability. Given this evolution of IC technology driven by Moore's law and the specifications of ULP applications, it is not clear whether ULP applications benefit from CMOS technology scaling, when reaching the nanometer era. This is the focus of this dissertation: investigation of the porting of digital circuits for ULP applications into nanometer CMOS technologies, by raising two questions:

- What is the impact of nanometer CMOS technology scaling on ultra-low-power digital circuits ?
- How to benefit from the circuit size reduction while keeping robustness and power/energy consumption under control ?

To answer first question, we propose a strong framework to support the analysis of energy efficiency in frequency/voltage-scaled digital circuits and we use it to carry out a detailed investigation of the impact of technology scaling on subthreshold circuits. We report three major issues that we then try to fix to answer the second question.

First, minimum-energy level increases when reaching 45 nm technology node. We propose an optimum MOSFET selection in standard nanometer bulk technology, which favors thin-oxide low- $V_t$  with an upsized gate length. The use of such optimum devices in subthreshold circuits leads to 40% energy saving. We also show that fully-depleted SOI technology with undoped-channel devices is very interesting for subthreshold circuits as it brings up to 60% energy saving with delay improvement as an extra benefit.

Second, energy in low-throughput applications becomes much higher than minimum-energy level in nanometer technologies. To solve this issue, we propose an appropriate technology flavor selection in versatile yet standard 45 nm technology and demonstrate the inefficiency of dual- $V_t$  assignment in nanometer subthreshold circuits. We show that adaptive reverse body biasing can be used to compensate for global process/temperature variations or dynamic workload variations. When using a power-gating technique for managing stand-by periods, we further propose to engineer the power switch in order to improve energy-efficiency in nanometer subthreshold circuits. Combining all these techniques can be used for keeping minimum energy per operation over a wide range of operating conditions.

Finally, high-temperature ( $> 150^\circ\text{C}$ ) operation increases energy consumption of ULP applications by two orders of magnitude due to leakage currents. We propose a novel ULP logic style to reduce leakage currents by three orders of magnitude at the expense of circuit delay. Additionally, ULP logic gates feature unique hysteresis property, which leads to high static noise margins and robustness as an extra benefit. We demonstrate that ULP logic style can be used as a low-cost and straightforward technique to build ULP circuits for high-temperature applications in standard SOI technology.

## ACRONYMS

---

ABB	Adaptive body bias
ASV	Adaptive supply voltage
BOX	Buried oxide
BTBT	Band-to-band tunneling
CD	Critical dimensions
CMOS	Complementary metal-oxide semiconductor
DFVS	Dynamic frequency/voltage scaling
DIBL	Drain-induced barrier lowering
FBB	Forward body bias
FD	Fully depleted
FS	Frequency scaling
FVS	Frequency/voltage scaling
FO	Fan out
GIDL	Gate-induced drain leakage
GP	General purpose
HP	High performance
IC	Integrated circuit
LOP	Low operating power
LP	Low power
LSTP	Low standby power
MOSFET	Metal-oxide semiconductor field effect transistor
MTCMOS	Multi-threshold CMOS
NMOS	N-type MOSFET
PG	Power gating
PMOS	P-type MOSFET
PS	Power switch
PTM	Predictive technology model
RBB	Reverse body bias

RCA	Ripple-carry array
RDF	Random dopant fluctuation
RFID	Radio-frequency identification
SNM	Static noise margin
SOI	Silicon on insulator
SRAM	Static random-access memory
ULP	Ultra low power
VTCMOS	Virtual-threshold CMOS



# LIST OF NOTATIONS

---

## Device-level symbols

$\epsilon_{ox}$	Oxide dielectric permittivity	$[F/m]$
$\epsilon_{Si}$	Silicon dielectric permittivity	$[F/m]$
$\gamma$	Linearized body-effect coefficient	$[mV/V]$
$\eta$	Drain-induced-barrier-lowering coefficient	$[mV/V]$
$\mu_0$	Zero-bias mobility	$[m^2/(V.s)]$
$\sigma_{V_t}$	Threshold-voltage standard deviation	$[mV]$
$\sigma_{L_g}$	Gate-length standard deviation	$[nm]$
$C_{dep}$	Channel depletion-layer capacitance	$[fF/\mu m^2]$
$C_g$	Intrinsic gate capacitance	$[fF/\mu m]$
$C_{g,par}$	Parasitic gate capacitance	$[fF/\mu m]$
$C_{if}$	Gate inner fringing capacitance	$[fF/\mu m]$
$C_j$	Source/drain junction capacitance	$[fF/\mu m]$
$C_{of}$	Gate outer fringing capacitance	$[fF/\mu m]$
$C_{ov}$	Gate overlap capacitance	$[fF/\mu m]$
$C_{ox}$	Gate-oxide capacitance	$[fF/\mu m^2]$
$I_0$	Subthreshold reference current	$[A/\mu m]$
$I_{gate}$	Gate leakage current	$[A/\mu m]$
$I_{junc}$	Junction leakage current	$[A/\mu m]$
$I_{off}$	Off-state drain current	$[A/\mu m]$
$I_{on}$	On-state drain current	$[A/\mu m]$
$I_{sub}$	Subthreshold drain current	$[A/\mu m]$
$L_{eff}$	Effective channel length	$[nm]$
$L_g$	Gate length (printed unless otherwise specified)	$[nm]$

$L_{gsd}$	Distance between gate and source/drain contacts	[nm]
$l_t$	Characteristic length of short-channel effects	[nm]
$n$	Body-effect factor	[—]
$N_{ch}$	Channel doping	[#/cm <sup>3</sup> ]
$S$	Subthreshold swing	[mV/dec]
$T_{ox}$	Gate-oxide thickness	[nm]
$U_{th}$	Thermal voltage	[mV]
$V_{bs}$	Body-to-source voltage	[V]
$V_{ds}$	Drain-to-source voltage	[V]
$V_{gs}$	Gate-to-source voltage	[V]
$V_t$	Threshold voltage	[V]
$V_{t0}$	Zero-bias threshold voltage	[V]
$W$	Channel width	[nm]
$X_{dep}$	Channel depletion depth	[nm]
$X_j$	Source/drain diffusion depth	[nm]

### Circuit-level symbols

$\alpha_F$	Activity factor	[—]
$C_L$	Load capacitance	[fF]
$C_{sw}$	Switched capacitance to perform an operation	[fF]
$E_{dyn}$	Dynamic energy per operation	[J]
$E_{min}$	Minimum energy per operation	[J]
$E_{op}$	Energy per operation	[J]
$E_{stat}$	Static energy per operation	[J]
$f_{clk}$	Clock frequency	[Hz]
$f_{clk,opt}$	Clock frequency of minimum-energy point	[Hz]
$f_{op}$	Operation throughput	[Op/s]
$F_{SNM}$	Static-noise-margin factor	[—]
$I_{leak}$	Circuit leakage current	[A]

$k_{DIBL}$	DIBL-induced delay factor	$[-]$
$L_D$	Logic depth	$[\#]$
$N_{nodes}$	Number of nodes in a circuit	$[\#]$
$N_{sw}$	Number of node switchings to perform an operation	$[\#]$
$P_{dyn}$	Dynamic power consumption	$[W]$
$P_{inst}$	Instantaneous power consumption	$[W]$
$P_{stat}$	Static power consumption	$[W]$
$P_{sc}$	Short-circuit power dissipation	$[W]$
$P_{sw}$	Switching power dissipation	$[W]$
$T_{del}$	Critical-path delay	$[s]$
$T_{op}$	Operation execution time	$[s]$
$V_{BB}$	Body bias voltage	$[V]$
$V_{dd}$	Supply voltage	$[V]$
$V_{dd,opt}$	Supply voltage of minimum-energy point	$[V]$



# INTRODUCTION

---

The story begins on a bright summer day of 1958 at Texas Instrument in Dallas. As a new employee, Jack Kilby had no vacation time that summer. When in the deserted laboratory he successfully built a small electronic circuit integrated onto a single slice of Germanium, he did not know he was not only about to revolutionize the electronic market but our everyday's life [1]. Fifty years later, his invention known as the integrated circuit (IC) fills up houses and offices, cars and planes, schools and hospitals, purses and pockets.

Behind the incredible evolution of IC market lay a cost reduction and a functionality increase, owing to the exponentially-growing number of integrated semiconductor devices on the same chip, known as Moore's law [2]. This growth in integrated device number relies on the shrinking of device feature size or technology scaling trend. Whereas Moore's law mainly targets high-performance applications, the success of IC market has lead to a diversification of the applications: from high-end ultra-fast super computers and servers to low-power portable devices, such as laptop computers and cell phones. Moreover, over the last decade, a new class of applications has emerged: ultra-low-power (ULP) applications such as radio-frequency identification (RFID) tags [3], wireless sensor networks [4] and biomedical devices [5]. ULP applications require minute power consumption with very loose speed requirements.

Modern semiconductor devices are CMOS field-effect transistors and today's minimum feature size is getting close to the nanometer level, with gate length as small as 30 nm in the so-called "45 nm" commercial technologies [6]. When reaching the nanometer era, two major detrimental side effects arise: leakage currents and device variability. Given this evolution of IC technology driven by Moore's law and the specifications of ULP applications, it is not clear whether ULP applications benefit from CMOS technology scaling, when reaching the nanometer era. This is the focus of this dissertation: investigation of the porting of digital circuits for ULP applications into nanometer CMOS technologies, by raising these questions:

- What is the impact of nanometer CMOS technology scaling on ultra-low-power digital circuits ?
- How to benefit from the circuit size reduction while keeping robustness and power/energy consumption under control ?

In this general introduction, we briefly introduce the concepts that motivated this work: technology scaling, ULP applications and circuits, before sketching the outline of the text.

## 1.1 MOORE'S LAW AND TECHNOLOGY SCALING

Moore's law denomination comes from Gordon E. Moore, co-founder of Intel. In 1965, he observed that the number of transistors per chip was roughly doubling every two years [2], thereby increasing the functionality per chip. It comes with an increase of the chip clock frequency to perform more operations per second. This trend is enabled by the famous technology scaling, which results in speed improvement for logic gates to support the clock frequency increase as well as a reduction of the energy required to perform a given operation.

Despite this reduction of the energy per operation, the total power consumption per chip dramatically suffers from the exponential growth in integrated device number and from the clock frequency increase. Nevertheless, technology scaling is driven by the lucrative market of high-performance applications and, historically, power consumption thus remained a second concern until the introduction of battery-operated portable devices in the early 90's [7]. For these low-power applications, limiting power consumption is equally important as increasing the speed. Therefore, we have seen a significant reduction of the supply voltage  $V_{dd}$  to limit dynamic power consumption [8]. These were golden years for IC designers, that are often referred to as the "happy-scaling era". Indeed, at that time, the concerted technology and voltage scalings were able to keep power consumption under control while meeting Moore's-law increasing integration density.

In the late 90's, clouds came in this beautiful picture in the shape of leakage currents [9]. Indeed, as the MOSFET threshold voltage  $V_t$  has to be scaled according to the supply voltage for maintaining speed improvement, the sub-threshold leakage current exponentially increases, thereby making static energy a primary concern. Moreover, a few years later, the scaling of gate oxide thickness  $T_{ox}$  also resulted in prohibitive gate-oxide tunneling leakage. Today, the happy-scaling era is over and technology designers have to limit  $V_{dd}$ ,  $V_t$  and  $T_{ox}$  scaling [10]. Nevertheless, Moore's law is still ruling the IC market to increase the functionality of electronic devices in the communication era. This is needed in both ultra-fast servers to support the development of the world-wide web, and in portable devices such as cell phones, personal digital assistants, music/video players or global positioning system (GPS) receivers to meet the demand for increasing portable service and entertainment. Technology scaling has thus to be pushed further.

The consequence is a new technology scaling trend, which keeps  $V_{dd}$ ,  $V_t$  and  $T_{ox}$  roughly constant, whereas the device area is constantly reduced. As a side effect, this trend puts an exacerbated pressure on the devices, which leads to short-channel effects and loss of channel control by the gate [11]. Moreover, reaching the nanometer era also means reaching the atom dimensions. Devices are so small that the number of dopants in their channel becomes discrete. As this discrete number is random by nature, it implies a high variability of total channel doping level and dopant placement i.e. random dopant fluctuations [12], which directly results in  $V_t$  and thus performance variability [13]. Similarly, as

the resolution of gate patterning process is hardly improved, manufactured gate edges no longer appear straight, when scaling the gate size [14]. This line edge roughness also results in device performance variability, which enforces IC designers to take sufficient safety design margins to ensure the circuits to work, thereby reducing the benefit of scaling.

## **I.2 ULTRA-LOW-POWER APPLICATIONS**

The historical example of ULP application is the electronic wristwatch, whose first prototype was developed in the 60's at the *Centre Électronique Horloger*, Neuchâtel, consuming less than  $30\ \mu\text{A}$  from a 1.3V supply voltage [7]. It remained the only ULP application for thirty years. Indeed, it is only a decade ago that this new class of IC applications actually arose. Amongst them, wireless sensor networks came along with the concept of ubiquitous computing or ambient intelligence. Such networks feature up to thousands of intelligent nodes that sense their environment, process data and transmit the resulting information to an end-user that can then act on it [4]. Applications include monitoring of habitat, structures and industrial processes. Radio-frequency identifier (RFID) tags is another new ULP application [3]. The tags are used to wirelessly identify an object, an animal or a person. Another category of ULP applications are biomedical devices, whether implanted or not such as hearing aids, cochlear implants, health care monitoring devices or body-area sensor networks, that can improve the quality of life for many people [5].

Yet varied, all these applications share a common characteristic: their low computational load, and a common constraint: a minute energy/power consumption. Indeed, these applications either have to operate for a long time on small batteries i.e. with low energy capacity or harvest power from the environment or from a wireless link.

Notice that, yet related, the important figures of merit to consider are different, depending on the energy/power source - battery or environment harvest. In battery-operated systems, the energy to perform an operation has to be low enough to sustain a reasonable battery life. As a reference level, a  $1\text{cm}^3$  Lithium battery has 1.5 kJ capacity, which means that it can deliver  $10\ \mu\text{W}$  continuously for 5 years [15]. In environment harvesting systems, the available power that can be harvested is small and it is thus the maximum instantaneous power consumption that has to be limited. Indeed, a  $1\text{cm}^2$  solar cell for example can only deliver  $3.2\ \mu\text{W}$  indoor [16].

## **I.3 FREQUENCY/VOLTAGE-SCALED SUBTHRESHOLD DIGITAL CIRCUITS**

The low computational load of ULP applications means that ULP digital circuits have to support low-to-medium data or operation throughputs. The clock frequency of ULP circuits can thus be drastically reduced, thereby relaxing speed

constraints. Speed can thus be traded off for reduction of instantaneous power, or energy per operation. On one hand, dynamic power/energy consumption in digital circuits due to capacitance switching is quadratically reduced by lowering the supply voltage  $V_{dd}$ , thereby making frequency/voltage-scaled (FVS) circuits very efficient for ULP applications. On another hand, static power due to subthreshold leakage current is exponentially reduced by increasing the threshold voltage  $V_t$ . These optimizations can be pushed to the limit where  $V_{dd}$  is lower than  $V_t$  and MOSFETs operate in subthreshold or weak-inversion regime. Under this condition, the circuits thus use subthreshold leakage as the active drain current, which depends exponentially on the threshold voltage and the MOSFET bias voltages.

Subthreshold operation has been first suggested by Swanson and Meindl [17]. Back in 1972, they showed that an inverter could operate under a supply voltage down to 100 mV. Subthreshold operation of analog circuits was demonstrated by Vittoz and Fellrath four years later at the *1976 European Solid-State Circuits Conference* (ESSCIRC) [18]. It is worth mentioning that the audience suggested that such circuits could not be reliable, as they operate with leakage currents [19]. However, the amplitude-regulated crystal oscillator that was presented has since been integrated in billions of electronic wristwatches.

Although analog subthreshold circuits receive attention thanks to the wristwatch application, digital subthreshold circuits remain in the shadow until the *1999 IEEE/ACM International Symposium on Low-Power Electronics and Design* (ISLPED), where Soeleman and Roy showed that operation of CMOS and pseudo-NMOS logic gates down to 0.3 V leads to nearly two-orders-of-magnitude power-delay product saving, in  $0.35\text{ }\mu\text{m}$  technology [20]. In 2002, Kao *et al.* demonstrated the operation of a multiply-accumulator unit down to 175 mV in  $0.14\text{ }\mu\text{m}$  technology [21]. They showed that energy per operation can be minimized by operating at an optimum  $V_{dd}$ , which balances dynamic and static energy and they reported measurement of this optimum  $V_{dd}$  below 0.5 V, deep in the subthreshold region. This concept of minimum energy point has since then become a vibrant research direction. In 2008, a decade after Soeleman's first subthreshold-logic paper, there have been numerous successful subthreshold circuit implementations, the most advanced one being a complete subthreshold microcontroller with embedded SRAM and DC-DC converter in 65 nm technology for biomedical applications, which was designed in collaboration between the Massachusetts Institute of Technology and Texas Instruments [22]. In parallel, numerous studies on technology optimizations and design techniques for subthreshold digital circuits have been carried out, making ULP subthreshold design a vibrant research area in digital electronics.

It is worth mentioning that the interest of subthreshold circuits is not limited to pure ULP applications. Indeed, it has recently been proposed to use the minimum-energy property of subthreshold circuits for two other class of applications, in mass-producton markets. First, Zhai *et al.* and Calhoun *et al.* suggested to extend the traditional range of dynamic frequency-voltage scaling (DFVS) scheme to the minimum energy point, down in the subthreshold region [23, 24].



General-purpose microprocessors in portable devices such as laptop computers or smart phones can benefit from this ULP mode to save energy when doing background computation or maintenance tasks that do not require high throughputs. Second, Zhai *et al.* and Sze *et al.* proposed to combine minimum-energy subthreshold operation with highly-parallelized architecture for acceptable speed performances [25, 26], thereby improving the energy efficiency of digital-signal processors for wireless applications. In this dissertation, we focus on ULP digital circuits for both pure ULP applications (niche market) and ULP-mode consumer portable applications (mass-production market).

## I.4 THESIS OUTLINE

The exponential dependence of subthreshold drain current on  $V_t$  and bias voltages increases the sensitivity of active drain current in subthreshold circuits against operating conditions and device parameters. The new effects induced by technology scaling in the nanometer era are thus magnified by subthreshold operation for ULP applications. This dissertation contains two aspects: analysis and solution proposal. The first two chapters are analysis-driven and try to answer the question “What is the impact of nanometer CMOS technology scaling on ultra-low-power digital circuits?”, revealing several new issues. In each of Chapters 3 to 5, we then try to fix one of the issues we pointed out, by balancing in-depth analysis and solution proposals to answer the question “How to benefit from the circuit size reduction while keeping robustness and power/energy consumption under control?”. This is the outline of the text.

**Chapter 1.** As a preliminary discussion, we have a brief look at the power and energy consumption of frequency/voltage-scaled (FVS) digital CMOS circuits, under robustness and throughput constraints. We present the sources of power/energy consumption and then show the evolution of practical power and energy under static FVS scheme from high-performance to ULP applications by using a unified representation for a wide throughput range. It allows us to clearly distinguish the context of ULP applications and highlight the benefit of frequency/voltage scaling down to the subthreshold regime. We show that the application throughput space can be divided in three regions [CP6], depending on the constraint that sets the limit on minimum supply voltage (robustness or throughput), and the dominating power/energy component (dynamic or static). This can be used as a strong framework to support the analysis of energy efficiency in FVS circuits. Moreover, we point out 2 important figures of merit of frequency/voltage-scaled ULP circuits: minimum-power range and minimum-energy point. It may be used for fast evaluation of the power/energy efficiency of ULP circuits, although practical power and energy consumption cannot be restricted to minimum power and energy levels, which can only be reached at particular application throughputs.

**Chapter 2.** We focus on subthreshold logic and analyze the impact of CMOS technology scaling from  $0.25\ \mu\text{m}$  to 32 nm node [JP2][CP3]. The analysis is first carried out at device level. It shows that worst-case subthreshold  $I_{on}$  increases with constant-field scaling trend until 90 nm node and then saturates because of subthreshold swing, drain-induced barrier lowering (DIBL) and variability increase. Fringing capacitances due to slow scaling of gate-stack height also exhibit a worrying increase.

At circuit level, the analysis shows that minimum supply voltage  $V_{dd}$  of subthreshold circuits jumps from speed to robustness limitation when migrating to smaller technology nodes. Instantaneous power consumption in low-throughput applications suffer from an extension of the minimum-power range and the increase of minimum-power level. Regarding energy per operation, we first report that minimum-energy level is reduced when migrating to 90 nm node thanks to dynamic energy reduction. It then increases as static energy does. Second, we show that technology scaling shifts the minimum-energy point towards higher throughput values. This shift combined with the reduction of minimum-energy level enables considerable practical energy savings at medium throughputs when migrating to 90/65 nm nodes. However, at 45/32 nm nodes, this benefit is outweighed by static energy. Moreover, for low-throughput applications, practical energy increases by 2 orders of magnitude when migrating from 180/90 nm to 45/32 nm node.

**Chapter 3.** As shown in Chapter 2, minimum energy in subthreshold circuits increases from 90 nm node, whereas its previously-reported  $C_L S^2$  figure of merit decreases. In this chapter, we first explain the new effects that make minimum energy rise in nanometer technology: DIBL, gate leakage and device variability. We then study the impact of nanometer MOSFET parameters on minimum energy. We show that traditional technology flavors are not adapted to minimum-energy subthreshold circuits and we propose an optimum device selection to improve energy efficiency, at circuit level, i.e. without any process modification. At 45 nm node, we show that the use of thin-oxide low- $V_t$  devices in a high-performance technology flavor with gate length upsized by 15 to 25 nm reduces minimum-energy level by 35-40%, with mitigation of delay variability as an extra benefit. This study draws a new route for device optimization towards ultimate subthreshold circuits, indicating that efforts should be devoted to minimizing subthreshold swing, DIBL and variability, while gate leakage increase can be tolerated provided that it remains below the subthreshold leakage level.

Finally, we investigate the potential of ultra-thin-body fully-depleted (FD) Silicon-on-insulator (SOI) technology to reduce minimum energy [CP5]. In standard 45 nm high-performance technology, FD SOI brings 45% minimum-energy reduction at minimum gate length, thanks to subthreshold swing improvement, capacitance reduction and variability mitigation. The combination of an undoped channel with a metal gate further increases this improvement, yielding a 60% minimum-energy reduction as compared to bulk, thanks to outstanding variability mitigation.

**Chapter 4.** As shown in Chapter 2, practical energy per operation under robustness and throughput constraints can be far higher than minimum energy. In this chapter, we revisit classical circuit design choices in the light of nanometer subthreshold digital circuits for ULP applications, the design target being to make practical energy reach the minimum energy level [CP6]. We show that fully-depleted SOI brings important practical-energy savings for the whole throughput range of ULP applications. We also demonstrate that the versatility of nanometer technologies is a powerful option to minimize practical energy, as it allows to shift minimum-energy point to different application throughputs. Nevertheless, we demonstrate that independent dual- $V_t$  assignment is inefficient in nanometer subthreshold circuits because of the large delay difference between std- and high- $V_t$  logic gates and the high variability of short paths.

We then show that adaptive reverse body biasing with negative voltage is an efficient technique to compensate for modeling errors or global process/temperature variations. It allows to limit design margins while keeping minimum-energy point at the target application throughput, under various operating conditions. On the contrary, forward body biasing suffer from increased minimum-energy level and bad behavior with discrete bias voltage values. Moreover at 45 nm node, we point out that reverse body biasing is only efficient in low-power technology flavor and we suggest that at next nodes it may no longer be practical because of decreasing body-bias coefficient and increasing band-to-band tunneling leakage.

Finally, we investigate the efficiency of sleep-mode techniques - dynamic reverse body biasing and power gating - for reducing active and stand-by leakage. For active-leakage reduction, sleep-mode techniques are less efficient than technology selection and static reverse body biasing, as they suffer from the energy overhead associated to mode transition. However, for reducing stand-by leakage, power gating is a very efficient technique in nanometer subthreshold circuits. Nevertheless, we showed that circuit robustness can be under risk when using badly-sized power switches and that engineering the power switch can bring significant energy reduction with lower robustness degradation.

**Chapter 5.** In high-temperature environments ( $> 150^\circ\text{C}$ ), static power/energy consumption completely dominates, even at  $0.13\mu\text{m}$  node. As no technology option in scaled technology nodes solves this issue, we propose a new logic style, named Ultra-Low-Power (ULP), which achieves negative  $V_{gs}$  self-biasing, to benefit from the small area and low dynamic power of scaled technologies while keeping ultra-low leakage, even at high temperature [CP1][PA1]. In  $0.13\mu\text{m}$  partially-depleted SOI CMOS technology, ULP logic style reduces static power consumption at  $200^\circ\text{C}$  by 3 orders of magnitude at the expense of increased delay and area, with good robustness against process variations [CP2][JP1]. Moreover, ULP logic gates feature excellent noise robustness thanks to SNM higher than  $V_{dd}/2$ , which is never achieved in standard CMOS logic style. Functionality of ULP logic style is demonstrated by measurement results of

ULP-inverter ring oscillators in  $0.13\mu\text{m}$  technology.

**Conclusions and appendixes.** We finally summarize the results and give concluding remarks with research perspectives in the general conclusion. Additionally, this dissertation comes with three appendixes. In Appendix A, we use the results from this thesis to derive the technology and circuits specifications for nanometer subthreshold circuit that we then combine into a possible roadmap for nanometer ULP circuits. Appendix B is a description of the 8-bit multiplier, which is used as a benchmark of ULP circuits throughout the dissertation. In Appendix C, we provide details about the pre-Silicon BSIM4 model cards that we generated for Spice simulation of nanometer subthreshold circuits in Chapters 2 and 3.

**Note:** regarding the applications mentioned in this general introduction, notice that Chapter 3 mainly targets ULP-mode operation in consumer low-power/wireless applications, while Chapters 4 and 5 target pure ULP applications for standard (consumer/biomedical/industrial) and high-temperature (industrial) environments, respectively.

## REFERENCES

1. “An Interview with Jack Kilby”, Texas Instruments, available at [www.ti.com/corp/docs/kilbyctr/interview.shtml](http://www.ti.com/corp/docs/kilbyctr/interview.shtml).
2. G. E. Moore, “Cramming more components onto integrated circuits”, in *Electronics*, vol. 38, no. 8, 4 p., Apr. 1965.
3. R. Weinstein, “RFID: a technical overview and its application to the enterprise”, in *IT Professional*, vol. 7, no. 3, pp. 27-33, May-Jun. 2005.
4. B. Warneke, M. Last, B. Liebowitz and K. S. J. Pister, “Smart Dust: communicating with a cubic millimeter computer”, in *Computers*, vol. 34, no. 1, pp. 44-51, Jan. 2001.
5. I. Korhonen, J. Pärkkä and M. Van Gils, “Health monitoring in the home of the future”, in *IEEE Eng. Medicine Biology Mag.*, vol. 22, no. 3, pp. 66-73, May/Jun. 2003.
6. K. Mistry *et al.*, “A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging”, in *Dig. IEEE Int. Electron Dev. Meeting*, pp. 247-250, 2007.
7. Ch. Piguet, “History of low-power electronics”, in *Low-Power Electronics Design*, Ch. Piguet Ed., CRC Press, pp. 1.1-15, 2005.
8. A. Wang, B. H. Calhoun and A. P. Chandrakasan, “Survey of low-voltage implementations”, in *Sub-Threshold Design for Ultra-Low-Power Systems*, A. P. Chandrakasan Ed., Springer, pp. 11-23, 2005.
9. K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, “Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits”, in *Proc. IEEE*, vol. 91, no. 2, pp. 305-327, Feb. 2003.
10. E. J. Nowak, “Maintaining the benefits of CMOS scaling when scaling bogs down”, in *IBM J. Research and Development*, vol. 46, no. 2/3, pp. 169-180, Mar./May 2002.
11. Y. Taur, “CMOS design near the limit of scaling”, in *IBM J. Research and Development*, vol. 46, no. 2/3, pp. 213-222, Mar./May 2002.
12. A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 $\mu$ m MOSFET: 3-D atomistic simulation study”, in *IEEE Trans. Electron Dev.*, vol. 45, no. 12, pp. 2505-2513, Dec. 1998.
13. K. A. Bowman, X. Tang, J. C. Eble and J. D. Meindl, “Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance”, in *IEEE J. Solid-State Circuits*, vol. 35, no. 8, pp. 1186-1193, May 2000.
14. J. A. Croon, G. Storms, S. Winkelmeier, I. Pollentier, M. Ercken, S. Decoutere, W. Sansen and H. E. Maes, “Line edge roughness: characterization, modeling and impact on device behavior”, in *Dig. IEEE Int. Electron Dev. Meeting*, pp. 307-310, 2002.
15. R. Hahn and H. Reichl, “Batteries and power supplies for wearable and ubiquitous computing”, in *Proc. 3es Int. Symp. Wearable Computers*, pp. 168-169, 1999.
16. *Panasonic Solar Cells Handbook '98/'99*, Matsushita Battery Industrial Co., Ltd., Aug 1998.

17. R. M. Swanson and J. D. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits", in *IEEE J. Solid-State Circuits*, vol. 7, no. 2, pp. 146-153, Apr. 1972.
18. E. Vittoz and J. Fellrath, "New analog CMOS IC's based on weak inversion operation", in *Proc. European Solid-State Circuits Conf.*, pp. 12-13, 1976.
19. E. A. Vittoz, "Origins of weak inversion (or sub-threshold) circuit design", in *Sub-Threshold Design for Ultra-Low-Power Systems*, A. P. Chandrakasan Ed., Springer, pp. 11-23, 2005.
20. H. Soeleman and K. Roy, "Ultra-low power digital subthreshold logic circuits", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 94-96, 1999.
21. J. T. Kao, M. Masayuki and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture", in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545-1554, Nov. 2002.
22. J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann and A. P. Chandrakasan, "A 65nm sub- $V_t$  microcontroller with integrated SRAM and switched-capacitor DC-DC converter", in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 318-319, 2008.
23. B. Zhai, D. Blaauw, D. Sylvester and K. Flautner, "The limit of dynamic voltage scaling and insomnia dynamic voltage scaling", in *IEEE Trans. VLSI Syst.*, vol. 13, no. 11, pp. 1239-1252, Nov. 2005.
24. B. H. Calhoun and A. P. Chandrakasan: "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering", in *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 238-245, Jan. 2006.
25. B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge and D. Sylvester, "Energy efficient near-threshold chip multi-processing", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 32-37, 2007.
26. V. Sze and A. P. Chandrakasan, "A 0.4-V UWB baseband processor", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 262-267, 2007.

# AUTHOR'S PUBLICATION LIST

---

## Related journal papers

- JP1. D. Bol, J. De Vos, R. Ambroise, D. Flandre and J.-D. Legat, "Building ultra-low-power high-temperature digital circuits in standard high-performance SOI technology", in *Solid-State Electronics*, vol. 52, no. 12, pp. 1939-1945, Dec. 2008.
- JP2. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Interests and limitations of technology scaling for subthreshold logic", in *IEEE Trans. on VLSI Systems*, in press, 12 p., 2009.

## Related conference papers

- CP1. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Building ultra-low-power low-frequency digital circuits with high-speed devices", in *Proc. IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1404-1407, 2007.
- CP2. D. Bol, D. Flandre and J.-D. Legat, "Ultra-low-power logic style for low-frequency high-temperature applications", in *Proc. EuroSOI workshop*, pp. 33-34, 2008.
- CP3. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Impact of technology scaling on digital subthreshold circuits", in *Proc. IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 179-184, 2008.
- CP4. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Channel length upsize for robust and compact subthreshold SRAM", in *Proc. 7è journées d'études Faible Tension Faible Consommation (FTFC)*, pp. 117-120, 2008.
- CP5. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Sub-45nm fully-depleted SOI CMOS subthreshold logic for ultra-low-power applications", in *Proc. IEEE International SOI Conference*, pp. 57-58, 2008. This poster presentation was awarded as Best Poster of the conference.
- CP6. D. Bol, R. Ambroise, D. Flandre and J.-D. Legat, "Analysis and minimization of practical energy in 45nm subthreshold logic circuits", in *Proc. IEEE International Conference on Computer Design (ICCD)*, pp 294-300, 2008. This paper was awarded as Best Paper of Logic and Circuits Track.

## Patent

- PA1. D. Bol, D. Flandre and J.-D. Legat, "Ultra-low-power circuit", EP2008/055239, 2008.

## Short-course invited presentations

- SC1. D. Bol and D. Flandre, "Technology scaling for ultra-low-power circuits - Is mainstream technology adapted to special design ?", in tutorial on "Process and device issues from a circuit point of view" at the *9<sup>th</sup> Conference on Ultimate Integration on Silicon (ULIS)* and in tutorial on "Ultra-low-power design" at the *7<sup>e</sup> journées d'études Faible Tension Faible Consommation (FTFC)*, 2008.
- SC2. D. Bol and D. Flandre, "Fully-depleted SOI for nanometer subthreshold circuits", in tutorial on "FD SOI" of the *Thematic Network on Silicon on Insulator Technology, Devices and Circuits (EuroSOI)*, 2008.
- SC3. D. Bol, "Digital design on SOI in the nanometer era - from high-performance to ultra-low-power circuits", in tutorial on "SOI design" of the *5<sup>th</sup> Workshop of the Thematic Network on Silicon on Insulator Technology, Devices and Circuits (EuroSOI)*, 2009.

## Unrelated papers

- UP1. D. Bol, I. Hassoune, D. Levacq, D. Flandre and J.-D. Legat, "Efficient multiple-valued signed-digit full adder based on NDR MOS structures and its application to an n-bit current-mode constant-time adder", in *J. Multiple-Valued Logic and Soft-Computing*, vol. 13, no. 1, pp. 61-78, 2007.
- UP2. D. Bol, M. J. Avedillo, J. M. Quintana, D. Flandre and J.-D. Legat, "Investigation of monostable-bistable transition logic element circuit based on ultra-low power diodes", in *Proc. EuroSOI workshop*, pp. 51-52, 2006.
- UP3. D. Bol, M. J. Avedillo, J. M. Quintana and J.-D. Legat, "MOBILE digital circuits based on negative-differential-resistance MOS structures", in *Proc. Conference on Design Circuits and Integrated Systems (DCIS)*, 5 p., 2006.
- UP4. D. Bol, J. M. Quintana, M. J. Avedillo and J.-D. Legat, "Monostable-bistable transition logic elements: threshold logic vs. boolean logic comparison", in *Proc. IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1049-1052, 2006.
- UP5. D. Bol, R. Ambroise, C. Roda Neve, J.-P. Raskin and D. Flandre, "Wide-band characterization and modeling of digital substrate noise in SOI technology", in *Proc. IEEE International SOI Conference*, pp. 133-134, 2007.



**Co-authored unrelated papers**

- CO1. I. Hassoune, A. Drummond, A. Gaudissart, D. Bol, D. Levacq, D. Flandre and J.-D. Legat, "A new multi-valued current-mode adder based on negative-differential resistance using ULP diodes", in *Solid-State Electronics*, vol. 49, no. 7, Jul. 2005, pp. 1185-1191.
- CO2. Ph. Manet, R. Ambroise, D. Bol, M. Baltus and J.-D. Legat, "Low power techniques applied to a 80C51 microcontroller for high temperature applications", in *Journal of Low-Power Electronics*, vol. 2, no. 1, pp. 95-104, Apr. 2006.
- CO3. Ph. Manet, D. Bol, R. Ambroise and J.-D. Legat, "Low-power techniques applied to a 80C51 microcontroller for high temperature applications", in *Proc. International Workshop Power and Timing Modelling, Optimization and Simulation (PATMOS)*, LNCS 3728, pp. 19-29, 2005.
- CO4. Ph. Manet, R. Ambroise, D. Bol, M. Baltus, L. Demeûs and J.-D. Legat, "Low-power 80C51 microcontroller in SOI high temperature technology", in *Proc. International Conference on High-Temperature Electronics (HITEN)*, 4 p., 2005.
- CO5. G. Meurice de Dormale, R. Ambroise, D. Bol, J.-J. Quisquater and J.-D. Legat, "Low-cost elliptic curve digital signature coprocessor for smart cards", in *Proc. IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP)*, pp. 347-353, 2006.
- CO6. Ph. Manet, D. Bol, R. Ambroise, M. Baltus, J. Creteur, L. Demeûs and J.-D. Legat, "High-temperature characterization of a low power HT SOI 80C51", in *Proc. International Conference High-Temperature Electronics (HiTEC)*, 4 p., 2006.
- CO7. J. De Vos, D. Bol and D. Flandre, "Cellule SRAM 12 transistors à ultra faible courant de fuite", in *Proc. 7<sup>è</sup> journées d'études Faible Tension Faible Consommation (FTFC)*, pp. 111-115, 2008.
- CO8. C. Roda Neve, D. Bol, R. Ambroise, J.-P. Raskin and D. Flandre, "Digital substrate noise reduction by low power circuit operation and SOI technology", in *Proc. 7<sup>è</sup> journées d'études Faible Tension Faible Consommation (FTFC)*, pp. 23-28, 2008.
- CO9. D. Kamel, D. Bol and D. Flandre, "Impact of layout style and parasitic capacitances in full adder", in *Proc. IEEE International SOI Conference*, pp. 97-98, 2008.



*It's a 5 volt world,  
and to change to 1.5 volt  
would mean that the whole world  
would have to change !*

***Gordon Moore***



## CHAPTER 1

---

# POWER AND ENERGY CONSUMPTION OF FREQUENCY/VOLTAGE-SCALED DIGITAL CIRCUITS

---

## Abstract

---

As a preliminary discussion, we have a brief look at the power and energy consumption of frequency/voltage-scaled (FVS) digital CMOS circuits, under robustness and throughput constraints. We present the sources of power/energy consumption and then show the evolution of practical power and energy under static FVS scheme from high-performance to ULP applications by using a unified representation for a wide throughput range. It allows us to clearly distinguish the context of ULP applications and highlight the benefit of frequency/voltage scaling down to the subthreshold regime. We show that the application throughput space can be divided in three regions [CP6], depending on the constraint that sets the limit on minimum supply voltage (robustness or throughput), and the dominating power/energy component (dynamic or static). This can be used as a strong framework to support the analysis of energy efficiency in FVS circuits. Moreover, we point out 2 important figures of merit of frequency/voltage-scaled ULP circuits: minimum-power range and minimum-energy point. It may be used for fast evaluation of the power/energy efficiency of ULP circuits, although practical power and energy consumption cannot be restricted to minimum power and energy levels, which can only be reached at particular application throughputs.

## Contents

---

1.1	Introduction	3
1.2	Constraints on digital circuits	3
1.3	Sources of power and energy consumption	7
1.4	Practical power and energy under robustness and throughput constraints	14
1.5	Conclusion	23

---

## 1.1 INTRODUCTION

Static frequency/voltage scaling (FVS) is a very efficient technique to reduce power/energy consumption of digital circuits for ULP applications whose speed requirements are not stringent [1, 2, 3]. In this prerequisite chapter, we cover the basics of power/energy consumption of CMOS digital circuits in the light of an FVS circuit for ULP applications under robustness and throughput constraints. In Section 1.2, we first introduce the general constraints on digital circuits and how we consider these constraints throughout this dissertation. We then briefly review the sources of power and energy consumption of digital circuits in Section 1.3. Finally, in Section 1.4, we show the evolution of practical power/energy under robustness and throughput constraints, when moving from high-performance to ULP applications. Based on simulations of a benchmark multiplier in  $0.13\mu\text{m}$  technology, we compare FVS with classic operating schemes and investigate the impact of operating temperature and circuit/application parameters.

## 1.2 CONSTRAINTS ON DIGITAL CIRCUITS

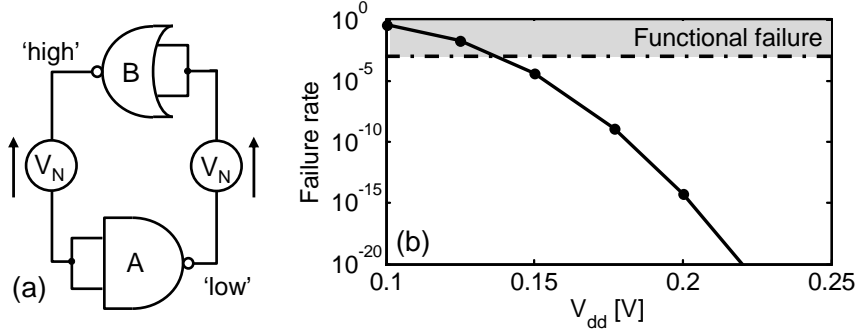
As economics rules IC market, the first important constraint is probably the cost. The cost has three main components: the design, the manufacturing (including test and packaging) and the raw material. An easy-designed circuit, with simple process steps and a small Silicon die area will be cheap. The importance of these factors is balanced by the scale factor related to the number of chips under production.

ULP applications is a niche market as the products have to be low-cost. Detailed cost modeling is beyond the scope of this dissertation and we will content ourself to rough qualitative cost considerations. As process modifications are hardly available to small customers of large foundries (niche-market fabless design companies), we assume that a process step modification is prohibitively expensive for ULP applications. We thus consider the cost of raw material, i.e. Silicon wafers, as the most important one.

Beyond cost, there are two main constraints that digital IC have to meet: robustness and throughput.

### 1.2.1 Robustness constraint

IC's have to be robust. First, a high percentage of the manufactured chips have to be qualified as functional and ready for sale, it is the chip test yield [4]. Second, amongst these qualified chips, a high percentage have to remain functional for a sufficient life time. This is yield over time or reliability [5]. In this dissertation, we do not consider manufacturing defects that can cause bad wafer test yield because it is process related. Similarly, we do not consider failure mechanisms such as gate oxide breakdown, hot carrier effects and electromigration nor we consider aging effects such as threshold voltage shift from negative-bias temperature instability



**Fig. 1.1.** Benchmark circuit to compute SNM distribution of logic gates (a) [8] and Monte-Carlo Spice-simulated functional failure rate vs.  $V_{dd}$  (b) in  $0.13\ \mu\text{m}$  standard bulk CMOS technology. Functional yield sets a limit on the minimum functional  $V_{dd}$ .

or radiation doses. Indeed, as FVS circuits for ULP applications feature both low voltage and current levels, we expect reliability effects to be less pronounced than in high-performance circuits.

Nevertheless, the low voltage and current levels imply a magnified sensitivity against device variability inherent to nanometer technologies [6]. We thus limit the robustness considerations to the most important threat of ULP circuits: logic functionality failure because of variability-induced bad logic levels [7]. Indeed, the operation at low supply voltage  $V_{dd}$  implies a low  $I_{on}/I_{off}$  current ratio and high current variability. It can in turn lead to bad output logic level of a gate, which would not be recognized as the correct logic level by the next gate. Some gates can thus exhibit functional failure, leading to bad functional yield of the circuit. In [8], Kwong *et al.* propose an efficient method to extract functional yield of digital gates. This method, inherited from the extraction of SRAM-cell SNM, is based on a statistical computation of static noise margins (SNM) of coupled gates depicted in Fig. 1.1(a), through Monte-Carlo Spice simulations in the context of process variability. A negative SNM means that the output low logic level  $V_{OL}$  of logic gate A is higher than  $V_{IL}$  (maximum input voltage recognized as low logic level) of logic gate B, and vice versa. The gates can thus not be operated at this supply voltage. The worst case consists of simulating as A a NAND gate, that has the highest  $V_{OL}$ , with as B a NOR gate, that has the lowest  $V_{IL}$ , as B. This structure is statistically simulated to extract the SNM distribution at a given  $V_{dd}$  supply voltage. A “functional yield” is then defined as the proportion of occurrences from the distribution having positive SNM. In this dissertation unless otherwise mentioned, we arbitrarily specify a functional-yield constraint of 99.9% meaning that the worst-case SNM computed with 99.9% confidence interval ( $3\sigma$  tail) is positive.

Exact SNM value depends on the circuit architecture through the kind of logic gates it uses and their arrangement. Indeed, two-cascaded inverters fea-



ture a better SNM than a 3-input NAND gate following a 3-input NOR gate. Nevertheless, computing the exact worst-case SNM of the whole circuit for all possible input patterns would require prohibitively-long Monte-Carlo Spice simulations. Pu *et al.* propose in [9] a faster simulation method based on MOSFET equivalent-resistance modeling. This method is shown to predict worst-case SNM with less than 1.5% deviation as compared to full-circuit Monte-Carlo Spice simulation, with a run-time reduction by 5 orders of magnitude. Nevertheless, the target of this dissertation is to investigate the main phenomena in nanometer ULP digital circuits rather than assessing the robustness of one given design. We therefore stick to Kwong's method [8] for the sake of generality.

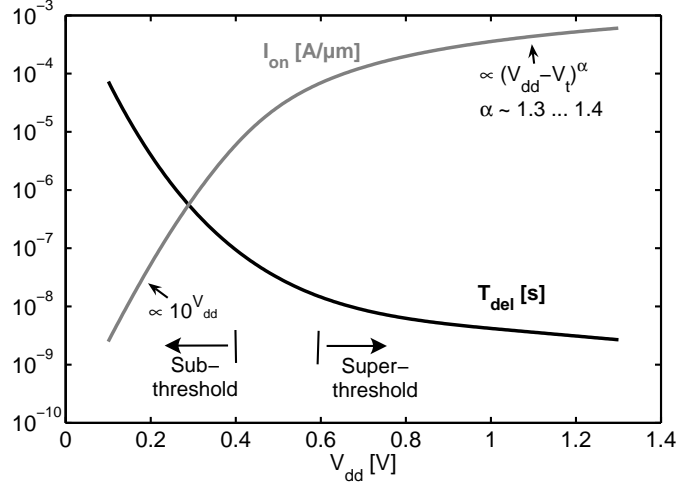
We use this method to extract the functional yield in an industrial 0.13  $\mu\text{m}$  standard bulk CMOS technology by 10k-point Monte-Carlo Spice simulations with production MOSFET compact models. Simulated functional failure rate ( $= 1 - \text{functional yield}$ ) is shown vs.  $V_{dd}$  in Fig. 1.1(b). Failure rate dramatically increases when lowering  $V_{dd}$  below 0.2V. In order to ensure a functional yield of 99.9%, i.e. keeping functional failure rate lower than 0.1%,  $V_{dd}$  has to be kept higher than 0.14V. This shows that functional yield sets a limit on the minimum functional  $V_{dd}$ .

The low voltage and current levels also increase the sensitivity against radiation- and ambient-noise-induced soft errors [10]. Nevertheless, we do not explicitly address this issue as robustness against soft errors is intrinsically included in the SNM metric, i.e. a logic gate with highest SNM is less sensitive to soft-errors.

### 1.2.2 Throughput constraint

Correctly performing an operation is compulsory yet not sufficient. In addition, a sufficient number of operations has to be performed within a given time, leading to timing or speed constraints, which can be expressed as a target data/operation throughput  $f_{op}$  to support. Although they are very loose for ULP applications, throughput constraints do exist. For example, a temperature sensor monitoring an industrial process may have to sense the temperature, process the data and transmit it, once in a second. Similarly, an RFID tag granting access to the London subway has to answer to the reader's challenge within a second, and a hearing aid has to support the ear bandwidth.

These system-level constraints can be translated into circuit-level constraints depending on the system architecture. ULP system architecture is beyond the scope of this dissertation and we assume that ULP circuits have to support operation throughputs  $f_{op}$  between 10k and 10M operations per second (Op/s), depending on the application. Throughout the text, we will focus on this wide throughput range in order to get results that are representative of a wide application spectrum. In this dissertation, we consider simple circuits i.e. without parallelism nor sequentialism for technology/device/design benchmarking, unless otherwise specified. They thus perform one operation per clock cycle and the minimum clock frequency  $f_{clk}$  to meet the constraint is directly equal to the



**Fig. 1.2.** Drain current  $I_{on}$  of on-state NMOS vs.  $V_{dd}$  and corresponding critical-path delay  $T_{del}$  (Spice simulation of an 8-bit RCA benchmark multiplier in  $0.13\mu\text{m}$  standard bulk technology)

target operation throughput  $f_{op}$ . Consequently, the throughput constraint simply implies that critical-path delay  $T_{del}$  has to be lower than the corresponding operation period  $T_{op} = 1/f_{op}$ <sup>1</sup>.

The delay of a CMOS logic gate is proportional to  $C_L V_{dd}/I_{on}$  where  $C_L$  is the typical load capacitance and  $I_{on}$  the MOSFET drain current in on-state i.e. at gate-to-source  $V_{gs}$  and drain-to-source  $V_{ds}$  voltages both equal to  $V_{dd}$  [1]. The throughput constraint on digital circuits can thus be expressed as:

$$T_{del} \propto L_D \times \frac{C_L V_{dd}}{I_{on}} < \frac{1}{f_{op}} \quad (1.1)$$

where  $L_D$  is the logic depth i.e. the number of gates in the critical path. Let us consider an 8-bit ripple-carry array (RCA) multiplier as a benchmark circuit, which will be considered as a benchmark of ULP circuits throughout this dissertation. Details on this circuit can be found in Appendix B. We simulate its worst-case delay with Spice simulator in the industrial  $0.13\mu\text{m}$  CMOS technology with 1.2V nominal  $V_{dd}$  and 0.4V  $V_t$  threshold voltage. The extracted worst-case delay is plotted vs.  $V_{dd}$  in Fig. 1.2 with the corresponding NMOS  $I_{on}$ . It shows that lowering  $V_{dd}$  results in delay increase because of significant  $I_{on}$  reduction. In this figure, we notice that the delay dependence on  $V_{dd}$  is much more important at low voltages. This is due to the regime of operation of MOSFETs, which changes at low voltages. Indeed, when  $V_{gs}$  is higher than the threshold voltage

<sup>1</sup>Notice that  $T_{op}$  is expressed in seconds while  $f_{op}$  is given in Op/s. Strictly speaking, we thus assume in these relationships an implicit translation of  $f_{op}$  into  $\text{s}^{-1}$ .

$V_t$ , MOSFET are in strong-inversion regime also called super-threshold regime and the drain current almost linearly depends on  $(V_{gs}-V_t)$  in modern technologies, where velocity saturation is important [11]. Consequently  $I_{on}$  also linearly depends on  $(V_{dd}-V_t)$ . When  $V_{dd}$  and thus  $V_{gs}$  are lower than  $V_t$ , MOSFETs are in weak-inversion or subthreshold regime and  $I_{on}$  exponentially depends on  $V_{dd}$  [12].

Fig. 1.2 shows that throughput constraint sets another limit on the minimum operating  $V_{dd}$ , through the maximum circuit delay. This minimum  $V_{dd}$  depends on the application as the throughput constraint does.

### 1.3 SOURCES OF POWER AND ENERGY CONSUMPTION

As explained in the general introduction, the important figure of merit for ULP applications is the power/energy consumption, depending on the power/energy source. Maximum instantaneous power has to be considered for applications that harvest power from their environment unless sufficient energy can be stored in the system, whereas energy per operation has to be considered for battery-operated applications. Let us first focus on instantaneous power consumption  $P_{inst}$ , which is composed of dynamic  $P_{dyn}$  and static  $P_{stat}$  components:

$$P_{inst} = P_{dyn} + P_{stat} . \quad (1.2)$$

In CMOS digital circuits, dynamic power is only consumed when the circuit performs computation, whereas static power comes from leakage currents constantly flowing through the circuit even when no computation is performed.

#### 1.3.1 Dynamic power consumption

##### *Capacitance switching*

In order to perform computation, some current is retrieved from the power supply source to charge the internal circuit capacitances, this is the switching power  $P_{sw}$  [1]:

$$P_{sw} = \frac{1}{2} N_{nodes} \alpha_F C_L V_{dd}^2 f_{clk} , \quad (1.3)$$

where  $\alpha_F$  is the activity factor,  $N_{nodes}$  the number of nodes in the circuit,  $C_L$  the mean load capacitance per node,  $V_{dd}$  the supply voltage and  $f_{clk}$  the operation throughput. Recall that we consider simple circuits that execute an operation in one clock cycle and thus  $f_{clk} = f_{op}$  in general. In this equation, the load capacitance is the only parameter related to the technology. It is composed of intrinsic gate capacitance, parasitic gate capacitance, junction and routing capacitances. The architectural parameters are  $N_{nodes}$  and  $\alpha_F$ . Timing characteristics of the circuit also impacts  $\alpha_F$  as unbalanced data paths generate parasitic switching (glitches), thereby making  $\alpha_F$  rise.

### Short circuit current

The switching of logic gates also generates another kind of dynamic power consumption. During an input transition of a CMOS logic gate, there is a direct current path from the power supply to the ground because during a short time  $\Delta t_{sc}$ , none of the NMOS and PMOS devices are totally shut down. This makes some short-circuit current flow through the gates at each transition. The associated short-circuit power consumption  $P_{sc}$  can be expressed as [13]:

$$P_{sc} = N_{nodes} \alpha_F \Delta t_{sc} I_{sc,avg} V_{dd} f_{clk} , \quad (1.4)$$

where  $I_{sc,avg}$  is an average short-circuit current. This equation can be transformed to obtain an expression similar to  $P_{sw}$  from Eq. (1.3) by denoting  $\Delta t_{sc} I_{sc,avg} = \frac{1}{2} C_{sc} V_{dd}$  and  $C_{sc} = \beta_{sc} C_L$ . Eq. (1.4) can thus be rewritten as:

$$P_{sc} = \frac{1}{2} N_{nodes} \alpha_F \beta_{sc} C_L V_{dd}^2 f_{clk} . \quad (1.5)$$

This is a first-order approximation because both  $\Delta t_{sc}$  and  $I_{sc,avg}$  non-linearly depend on  $V_{dd}$ , which means that  $\beta_{sc}$  is not independent on  $V_{dd}$ .

The total dynamic power  $P_{dyn}$  is thus expressed as:

$$\begin{aligned} P_{dyn} &= P_{sw} + P_{sc} \\ &= \frac{1}{2} N_{nodes} \alpha_F (1 + \beta_{sc}) C_L V_{dd}^2 f_{clk} . \end{aligned} \quad (1.6)$$

Provided that the circuit is properly sized i.e.  $\Delta t_{sc}$  is short,  $P_{sc}$  can be kept at 5-10% of  $P_{sw}$  [14]. Moreover, when  $V_{dd}$  is lower than the sum of NMOS and PMOS threshold voltages  $V_{t,n} + |V_{t,p}|$ ,  $I_{sc,avg}$  decreases fast so that  $\beta_{sc}$  is much lower than 1 and can be neglected. Therefrom, we do not explicitly address short-circuit power consumption in this dissertation, although it is naturally included in the simulation and measurement results.

### 1.3.2 Static power dissipation

The static power dissipation comes from leakage current flowing through the devices:

$$P_{stat} = I_{leak} V_{dd} . \quad (1.7)$$

In nanometer MOSFETs, there are three main components of leakage current: subthreshold, gate and junction leakage. In static CMOS logic, off-state devices have zero  $V_{gs}$  and  $V_{ds} = V_{dd}$ , as represented in Fig. 1.3. Leakage currents thus flow from the drain to the source, the gate and the substrate. In this section, we briefly review the physical reasons behind leakage currents and the device parameters that impact it. A more in-depth review can be found in [15].

#### Subthreshold leakage

A MOSFET is said to be in subthreshold regime when its gate-to-source voltage  $V_{gs}$  is lower than its threshold voltage  $V_t$ . This is also called the weak-inversion

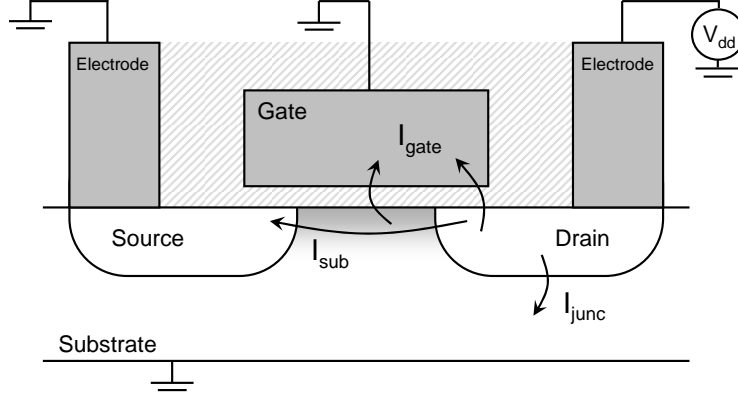


Fig. 1.3. Sketch of MOSFET leakage current components

regime as the channel is in weak inversion, i.e. the minority-carrier concentration is small but not zero. If a drain-to-source voltage  $V_{ds}$  is applied when in subthreshold regime, a diffusion current appears due to the different carrier concentrations in the inversion layer at source and drain terminals. This subthreshold current exponentially depends on  $V_{gs}$ ,  $V_{ds}$  and  $V_t$  through the carrier concentration and its value per device width unit can be expressed as [15]:

$$I_{sub} = \mu_0 C_{ox} \frac{1}{L_{eff}} (n-1) U_{th}^2 \times e^{\frac{V_{gs}-V_t}{nU_{th}}} \times \left(1 - e^{\frac{-V_{ds}}{U_{th}}}\right), \quad (1.8)$$

where  $\mu_0$  is the zero-bias mobility,  $C_{ox}$  the gate oxide capacitance,  $L_{eff}$  the effective channel length,  $n$  the body-effect factor and  $U_{th}$  the thermal voltage close to 26 mV at ambient temperature. Simulated NMOS drain current is plotted vs.  $V_{gs}$  in Fig. 1.4(a). The subthreshold or weak-inversion leakage contribution to  $I_{ds}$  is visible for  $-0.2V < V_{gs} < 0.4V$ , i.e. the linear  $I_{ds}$  region in logarithmic scale.

In off-state MOSFETs, zero- $V_{gs}$  subthreshold current is a leakage current component, which causes static power dissipation. The evolution of zero- $V_{gs}$   $I_{sub}$  with  $V_{ds} = V_{dd}$  is plotted in Fig. 1.4(b). The last term of Eq. (1.8) only impacts  $I_{sub}$  when  $V_{ds} < 0.1V$ . Nevertheless,  $I_{sub}$  still increases above 0.1V because of drain-induced barrier lowering (DIBL) effect. This effect only appears in short-channel devices where the drain depletion regions significantly penetrates the channel and lowers the barrier potentials [15]. DIBL can be modeled by a  $V_t$  reduction with linear dependence on  $V_{ds}$  [16]. Similarly, the body-to-source voltage  $V_{bs}$  also impact  $V_t$  through the body effect, which modifies the channel depletion width [16]. The threshold voltage can thus be expressed as:

$$V_t = V_{t0} - \gamma V_{bs} - \eta V_{ds}, \quad (1.9)$$

where  $V_{t0}$  is the zero-bias threshold voltage,  $\gamma$  the linearized body-effect coefficient and  $\eta$  the DIBL coefficient. Notice that  $V_t$  also depends on channel length through a short-channel effect known as  $V_t$  roll-off. Similarly to DIBL, source and drain depletion regions indeed deeply penetrate the channel when its length is scaled down, even at low drain bias, thereby lowering the barrier potential and thus  $V_t$  [15]. This effect can be mitigated by halo doping, which may in turn result in a reverse short-channel effect, i.e. an increase of  $V_t$  when scaling the channel length down [17].

Considering a MOSFET with body tied to source, the expression of  $I_{sub}$  per width unit from Eq. (1.8) can be rewritten by gathering all bias-independent multiplicative factors under an  $I_0$  term:

$$I_{sub} = I_0 \times 10^{\frac{V_{gs} + \eta V_{ds}}{S}} \times \left( 1 - e^{\frac{-V_{ds}}{U_{th}}} \right), \quad (1.10)$$

where  $S$  is the subthreshold swing equal to  $\ln(10) n U_{th}$ . Yet very simple, this expression accurately models  $I_{sub}$ , as shown in Fig. 1.4 (circle markers). Throughout this dissertation, we will therefore use  $I_{sub}$  expression from Eq. (1.10), when referring to subthreshold current.

Reduction of subthreshold leakage involves a  $V_t$  increase, which has a detrimental impact on circuit delay. This can be achieved by the use of high- $V_t$  devices in a multi- $V_t$  technology, which is common since 0.13  $\mu\text{m}$  node or by the application of a reverse body bias (RBB) to the devices. Many design techniques based on these features have been proposed to reduce subthreshold leakage, while maintaining circuit performances [15]. These techniques will be presented in details in Chapter 4.

#### Gate leakage

The thinning down of gate oxide results in carrier tunneling through the oxide, which causes the second leakage component: gate-tunneling leakage current  $I_{gate}$ . This current from/to the gate can flow from/to the source, the drain, the channel inversion layer and/or the substrate, depending on the bias conditions. Detailed modeling of gate leakage is beyond the scope of this brief discussion but can be found in [15]. Gate leakage exponentially depends on the electric field across the oxide and the height of the oxide potential barrier, and thus in turn the applied gate voltage and the oxide thickness. Notice that PMOS gate leakage is typically one order of magnitude lower than NMOS gate leakage because the energy required for hole tunneling in  $\text{SiO}_2$  is much higher than for electron tunneling [18].

Gate leakage of on- and off-state MOSFETs is plotted vs.  $V_{dd}$  in Fig. 1.4(b). On-state MOSFETs suffer from higher gate leakage because a bias voltage equal to  $V_{dd}$  is applied to  $V_{gs}$ ,  $V_{gd}$  and  $V_{gb}$  rather than only  $V_{gd}$  for off-state MOSFETs, resulting in higher energy-band bending. In 0.13  $\mu\text{m}$  technology, gate leakage is much smaller than subthreshold leakage. Moreover, when lowering  $V_{dd}$ , gate leakage becomes proportionally less important as it exhibits a higher dependence on  $V_{dd}$ .

Gate leakage mitigation requires an increase of the oxide thickness  $T_{ox}$ . In nanometer technologies, this is not feasible as a thicker  $T_{ox}$  implies a loss of gate capacitance and in turn channel control, which results in short-channel effects and variability increase. Today's main focus for gate leakage mitigation is thus the replacement of  $SiO_2$  gate dielectric by high- $\kappa$  dielectric material i.e. with a high dielectric permittivity. This enables having a thicker physical  $T_{ox}$  with less gate leakage as a result, at iso- $C_{ox}$  and thus channel control [19]. General adoption of high- $\kappa$  dielectric is predicted at 32nm node. In 2008, only Intel provides commercial 45nm chips with this feature, in which  $25\times$  gate leakage reduction is claimed as compared to standard 65nm technology [20].

#### *Junction leakage*

The last leakage component is the reverse-biased drain-to-substrate junction leakage  $I_{junc}$ . Junction leakage comes from two mechanisms: diffusion/drift of the minority carriers and thermal electron-hole pair generation in the depletion region. Moreover, in nanometer MOSFETs, heavily-doped shallow junctions and halo doping for short-channel effect control are often used, resulting in another junction leakage mechanism: band-to-band tunneling (BTBT) [15]. BTBT of electrons occurs from the valence band of the P-type region to the conduction band of the N-type region. Once again, detailed modeling of BTBT is beyond the scope of this discussion and can be found in [15].

Junction leakage is plotted vs.  $V_{dd}$  in Fig. 1.4(b) for a  $0.13\mu m$  technology. At ambient temperature, junction leakage is dominated by BTBT in this technology. Fig. 1.4 shows that in modern MOSFETs junction leakage is low as compared to subthreshold and gate leakage, at room temperature. Moreover, junction leakage can further be reduced [21] by using Silicon-on-insulator (SOI) technology. Little attention will thus be paid to junction leakage in this dissertation.

#### *Other leakage components*

In nanometer technologies, other leakage currents may also occur when the devices are used in non-standard configuration. Punchthrough occurs when the source and drain depletion regions merge under a too high  $V_{ds}$  for the considered channel length and doping [15]. This implies a large subthreshold current and degraded subthreshold swing. Punchthrough is avoided in properly-designed devices by additional implants. We thus do not consider punchthrough in this dissertation.

Gate-induced drain leakage (GIDL) occurs when the gate is reversely biased, as shown in Fig. 1.4. It comes from the high field induced in the drain junction and is more pronounced at high  $V_{ds}$  [15]. As the gate is never reversely biased in standard CMOS logic gates, GIDL does not occur under normal operation and can be neglected.

#### *Impact of the temperature*

The temperature has a strong impact on leakage currents, which makes static power even more important when the operating temperature increases. Table 1.1

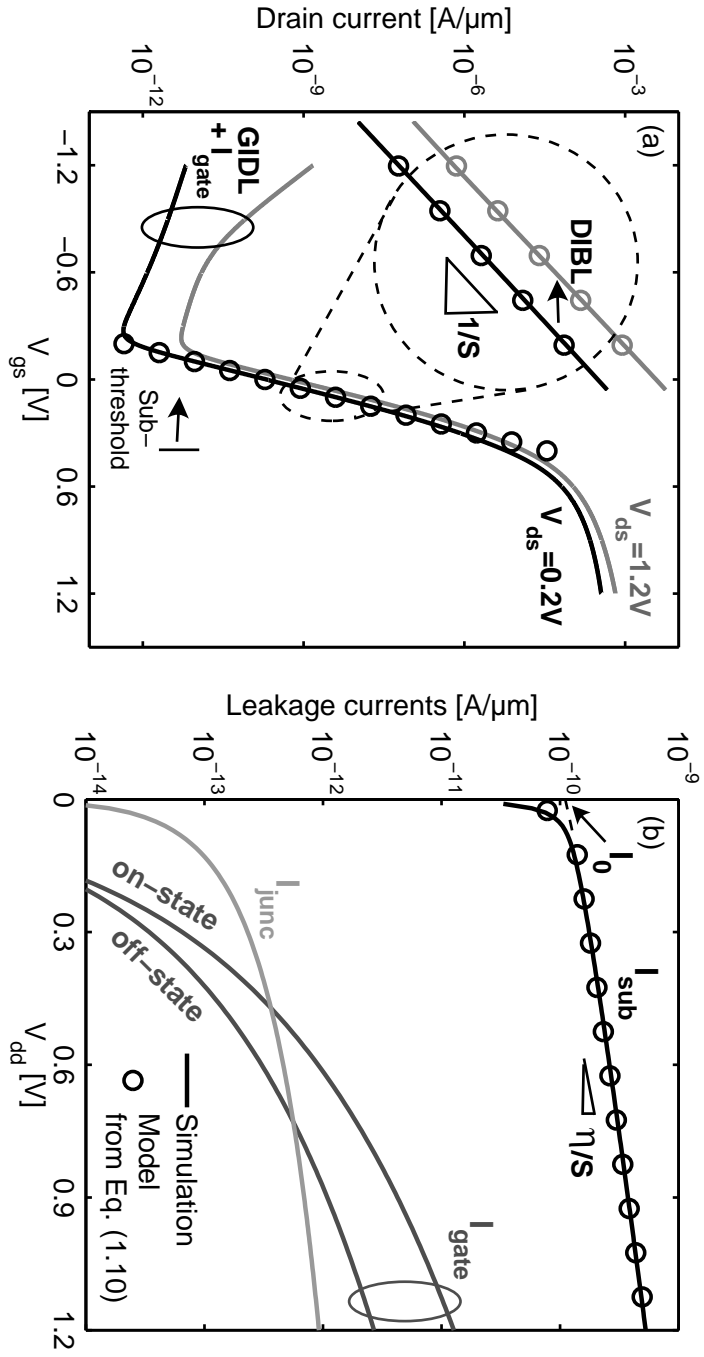


Fig. 1.4. NMOS drain current vs. gate voltage at 25°C (a) and leakage currents vs.  $V_{dd}$  (b) in standard bulk CMOS 0.13  $\mu\text{m}$  technology



**Table 1.1.** Temperature dependence of leakage components [22]

Leakage component	Temperature dependence
Gate leakage	$2 \times /100^\circ C$
Subthreshold leakage	$8 - 12 \times /100^\circ C$
Junction leakage	$50 - 100 \times /100^\circ C$

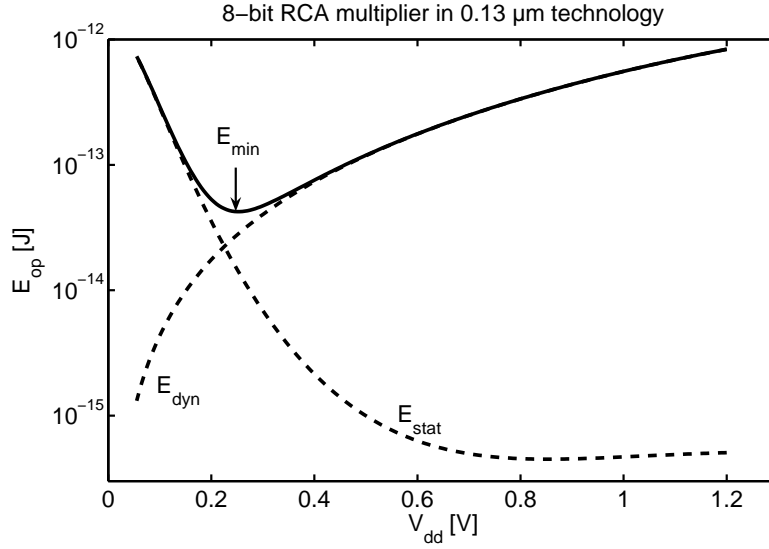
lists the temperature dependence of leakage components reported in [22]. Gate leakage weakly depends on the temperature. Although gate and subthreshold leakage components can be equally important at room temperature depending on the technology, at higher temperatures (75-150°C) subthreshold leakage completely dominates leakage currents. Indeed, subthreshold current increases fast as a temperature rise lowers  $V_t$  and degrades the subthreshold swing [23]. Above 150°C, junction leakage becomes troublesome. Notice that circuit operation at high temperature, i.e. above 150°C, often requires SOI technology for robustness concern and to prevent junction leakage from ruining power consumption [24]. High-temperature operation will be discussed into more details in Chapter 5.

### 1.3.3 Energy per operation

The energy per operation can be computed by integrating the power consumption over the time required to perform the operation  $T_{op} = 1/f_{op}$  :

$$\begin{aligned}
 E_{op} &= E_{dyn} + E_{stat} \\
 &= \int_0^{T_{op}} P_{dyn} dt + \int_0^{T_{op}} P_{stat} dt \\
 &= \frac{1}{2} N_{sw} C_L V_{dd}^2 + \frac{V_{dd} I_{leak}}{f_{op}}, \tag{1.11}
 \end{aligned}$$

where  $N_{sw} = \alpha_F N_{nodes}$  is the number of node switching to perform the operation. Recall that the target operation throughput is basically application-dependent and cannot be tuned by circuit designers, as explained in Section 1.2.2. However, if we make the assumption that there is no throughput constraint, the clock frequency  $f_{clk}$  can be freely assigned and replaces  $f_{op}$  in  $E_{stat}$  expression from Eq. (1.11). Consequently,  $E_{stat}$  is minimized by operating at the maximum  $f_{clk} = 1/T_{del}$ . Fig. 1.5 shows simulated  $E_{op}$  vs.  $V_{dd}$  for the benchmark multiplier in 0.13  $\mu m$  technology under this assumption. As reported in [25, 26], it leads to a minimum-energy point  $E_{min}$ , which lays in the subthreshold region. This comes from  $E_{dyn}$  quadratic reduction with  $V_{dd}$  scaling, whereas  $E_{stat} = V_{dd} I_{leak} T_{del}$  exponentially increases in subthreshold region as  $T_{del}$  does. This concept of minimum-energy point has been an important research direction since 2002, as it is useful for both pure ULP applications (niche market) and ULP-mode consumer portable applications (mass-production market) as detailed in Section I.3.



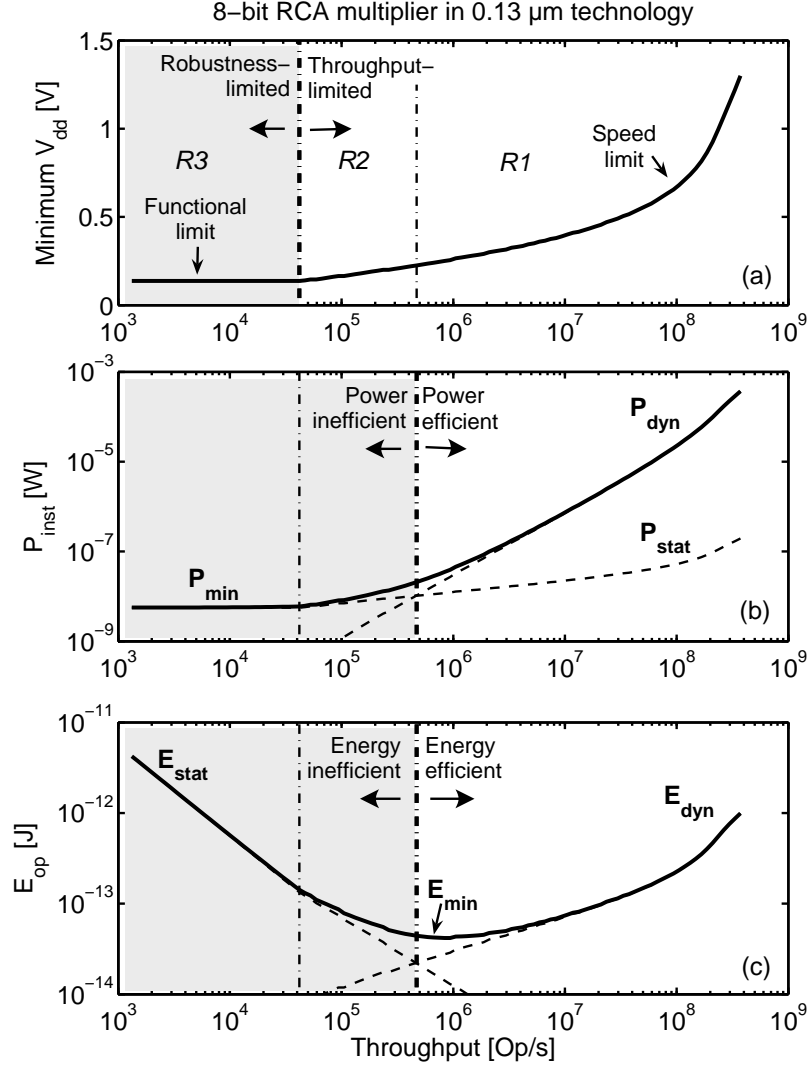
**Fig. 1.5.** Energy per operation vs.  $V_{dd}$  at maximum clock frequency  $f_{clk} = 1/T_{del}$  by assuming no throughput constraint (Spice simulation of an 8-bit RCA benchmark multiplier in 0.13  $\mu\text{m}$  standard bulk CMOS technology).

## 1.4 PRACTICAL POWER AND ENERGY CONSUMPTION UNDER ROBUSTNESS AND THROUGHPUT CONSTRAINTS

In this section, we present the evolution of power and energy consumption from high-performance to ULP applications under robustness and throughput constraints. We first examine the benefit of a static frequency/voltage-scaling (FVS) scheme and compares it to other low-power/energy operating schemes. We then show the impact of the operating temperature and the circuit/application parameters.

### 1.4.1 Frequency/voltage scaling scheme

Let us get back to the benchmark multiplier example in 0.13  $\mu\text{m}$  technology. Fig. 1.6(a) shows the minimum operating  $V_{dd}$  for meeting both the robustness constraint i.e. 99.9% functional yield, and the throughput constraint i.e.  $T_{del} \leq 1/f_{op}$ . When the target application throughput is relaxed from high-performance (several hundreds of MOp/s) to ULP (10 k to 10 MOp/s) range, minimum  $V_{dd}$  decreases monotonically, down into to MOSFET subthreshold region i.e. below 0.4V for the considered technology, and ultimately reaches the functional limit. From this figure, the throughput space can be divided into two regions depending on the constraint that gives the highest limit for minimum  $V_{dd}$ : robustness or throughput [CP6]. For the considered benchmark multiplier



**Fig. 1.6.** Minimum supply voltage  $V_{dd}$  (a) under throughput and robustness constraints with (b) corresponding instantaneous power and (c) energy per operation (Spice simulation of an 8-bit RCA benchmark multiplier in 0.13  $\mu\text{m}$  standard bulk CMOS technology). Throughput space can be divided into three regions: energy efficient (R1) and energy inefficient with either throughput (R2) or robustness (R3)  $V_{dd}$  limitation.

in 0.13  $\mu\text{m}$  technology, minimum  $V_{dd}$  is limited by robustness constraint when target throughput is lower than 40 kOp/s.

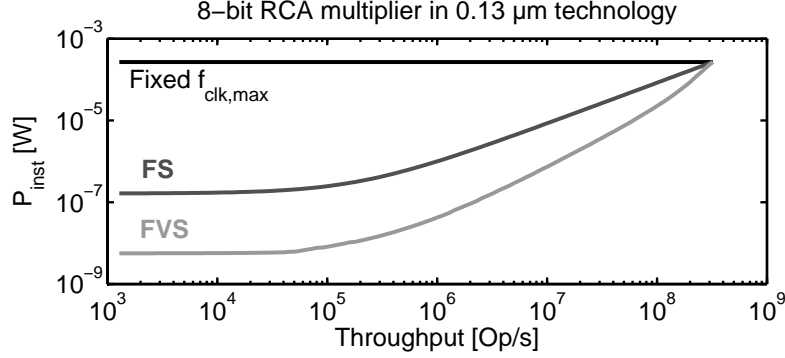
Fig. 1.6(b) shows the corresponding instantaneous power  $P_{inst}$  under such an FVS scheme i.e. at minimum  $V_{dd}$  and  $f_{clk} = f_{op}$ . Pseudo-random input pattern is considered. At high throughput,  $P_{inst}$  is dominated by dynamic power  $P_{dyn}$  and a relaxation of application throughput yields high  $P_{inst}$  benefits from both  $f_{clk} = f_{op}$  and  $V_{dd}^2$  lowering, according to Eq. (1.6). At lower throughputs, static power dissipation  $P_{stat}$  becomes dominant and the  $P_{inst}$  reduction thus slows down because  $P_{stat}$  does not depend on  $f_{clk}$ . The  $P_{stat}$  reduction thus relies on voltage scaling only, from  $V_{dd}$  term in Eq. (1.7) and from leakage current reduction as all leakage components are mitigated at low voltage. In particular, according to Eq. (1.10) the DIBL effect reduction at low voltage mitigates the subthreshold leakage, which is the dominant component in the considered technology. When minimum  $V_{dd}$  reaches its functional limit,  $P_{stat}$  is not reduced further and thus sets a lower bound  $P_{min}$  on total power. For the considered benchmark multiplier,  $P_{min}$  is 5.6 nW at 0.14V minimum functional  $V_{dd}$ , and is reached (within 10%) in applications with throughputs below 50 kOp/s.

As shown in Fig. 1.6(b), the throughput space can once more be divided into two new regions, depending on the dominating power/energy component [CP6]. For the considered technology, above 470 kOp/s, dynamic component dominates and the circuit is “power efficient” because consumed power/energy actually contributes to perform the operation. However, when throughput is lower than 470 kOp/s, static component dominates and the circuit is thus “power inefficient”.

Similarly, Fig. 1.6(c) shows the corresponding energy per operation  $E_{op}$  under such an FVS scheme i.e. with  $f_{clk} = f_{op}$ . As for  $P_{inst}$ ,  $E_{op}$  is dominated by dynamic component  $E_{dyn}$  in high-throughput applications and by static component  $E_{stat}$  at low throughputs.  $E_{dyn}$  reduction is slower than  $P_{dyn}$  as it is independent on  $f_{op}$  and only relies on  $V_{dd}$  reduction. Unlike  $P_{stat}$ ,  $E_{stat}$  depends on  $f_{op}$ . Indeed, when operating at low throughputs, the leakage currents are integrated over a longer operation period  $T_{op} = 1/f_{op}$  and  $E_{stat}$  increases, according to Eq. (1.11). This increase is somewhat mitigated by the minimum  $V_{dd}$  reduction down to the point where functional limit is reached. Below this throughput, the  $E_{stat}$  increase gets worst. As a result, the minimum-energy point  $E_{min}$  is obtained for one particular application throughput value, resulting from a balance between  $E_{dyn}$  and  $E_{stat}$ , when their slopes (in linear scale) have same amplitude with opposite sign. For the considered benchmark multiplier,  $E_{min}$  is 42 pJ at 0.26V  $V_{dd}$ , and is reached for applications with 940 kOp/s throughput.

As for  $P_{inst}$ , the throughput space can be divided in two regions, depending on the dominating energy component. Interestingly notice that  $E_{min}$  is situated in the “energy-efficient” region.

Let us summarize our observations. Static FVS scheme brings important power/energy saving when the application throughput is moderate. When looking at practical power/energy under robustness and throughput constraints, application throughput space can be divided into three regions [CP6]:



**Fig. 1.7.** Instantaneous power comparison between fixed frequency/voltage, fixed voltage and frequency scaling (FS) and frequency/voltage scaling (FVS) schemes (Spice simulation of an 8-bit RCA benchmark multiplier in 0.13  $\mu\text{m}$  technology).

- power/energy-efficient  $R1$  region where dynamic consumption dominates,
- power/energy-inefficient  $R2$  region where static consumption dominates and minimum  $V_{dd}$  is limited by throughput constraint,
- power/energy-inefficient  $R3$  region where minimum  $V_{dd}$  is limited by robustness constraint.

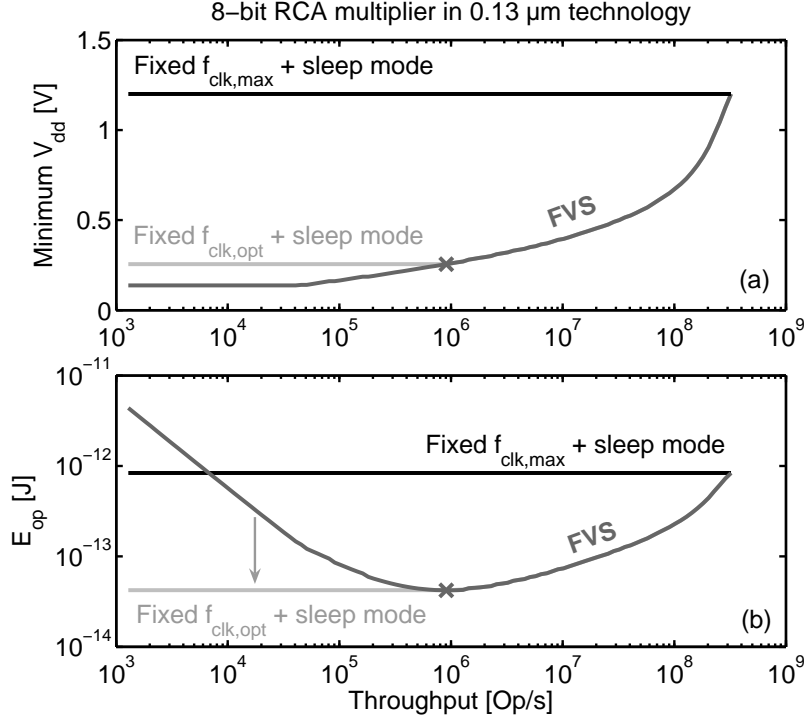
Instantaneous power consumption reaches a minimum value for a wide throughput range in  $R3$  region, while energy per operation features a minimum value at one particular target throughput located in  $R1$  region. Looking at the throughput region of interest for ULP applications ( $\approx 10$  kOp/s - 10 MOp/s), the circuit is either in  $R1$ ,  $R2$  or  $R3$  depending on the target throughput, the constant being a low  $V_{dd}$  down in the MOSFET subthreshold regime.

#### 1.4.2 Comparison with classic operating schemes

In order to evaluate the efficiency of FVS scheme, we compare it to other operating schemes. Fig. 1.7 shows the instantaneous power of the benchmark circuit:

1. under nominal supply voltage  $V_{dd,nom}$  with fixed clock frequency ( $f_{clk,max} = 1/T_{del}$  at  $V_{dd,nom}$ ),
2. under nominal supply voltage  $V_{dd,nom}$  with frequency-scaling (FS) scheme ( $f_{clk} = f_{op}$ ),
3. under FVS scheme.

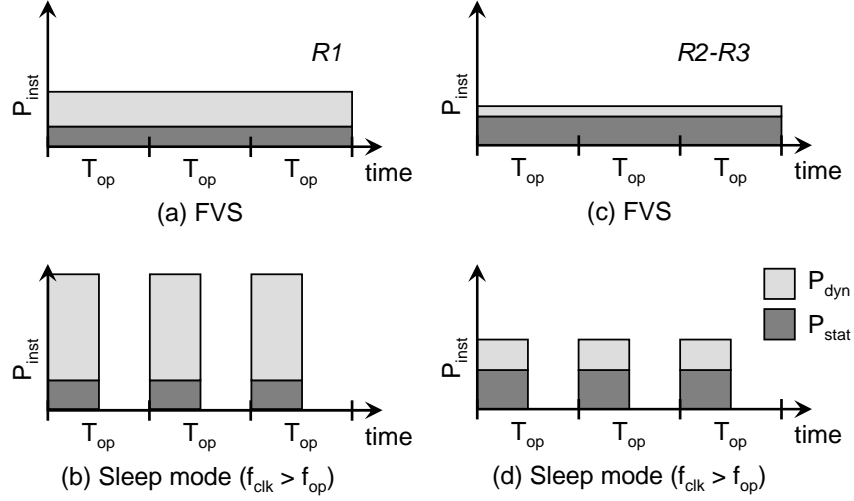
For low-throughput applications, FVS scheme reduces  $P_{inst}$  by 5 orders of magnitude as compared to fixed- $f_{clk}$  scheme and by 2 orders of magnitude as compared to FS scheme.



**Fig. 1.8.** Minimum supply voltage  $V_{dd}$  (a) under throughput and robustness constraints with corresponding energy per operation (b) under fixed frequency/voltage with sleep mode, and FVS with and without ideal sleep mode (Spice simulation of an 8-bit RCA benchmark multiplier in  $0.13\ \mu\text{m}$  technology).

#### 1.4.3 Comparison with sleep-mode operating scheme

Regarding energy per operation, circuit can be operated in a sleep-mode scheme i.e. running at fixed frequency and then being powered down if the operation is completed in advance. When in sleep mode, the target is to cut down the power consumption, which requires clock gating for suppressing  $P_{dyn}$  and leakage gating to suppress  $P_{stat}$ . Leakage-gating techniques such as multi-threshold CMOS (MTCMOS) power-gating or virtual-threshold CMOS (VTCMOS) dynamic everse body biasing will be presented into more details in Chapter 4. For simplicity concern, we make the assumption in this preliminary discussion of an ideal leakage-gating technique i.e. resulting in zero leakage when in sleep mode with no penalty on active mode and no energy overhead associated to mode transission. Moreover, in order to minimize the time over which static power is integrated and thus static energy,  $f_{clk}$  has to be set to its maximum value given by the circuit delay:  $f_{clk,max} = 1/T_{del}$  at  $V_{dd,nom}$ , as illustrated in Fig. 1.8(a).



**Fig. 1.9.** Sketch of  $P_{inst}$  under FVS scheme and sleep-mode scheme both in  $R1$  energy-efficient and  $R2 - R3$  energy inefficient regions.  $E_{op}$  results from the integration of  $P_{inst}$  over the operation period  $T_{op} = 1/f_{op}$ . Under sleep-mode scheme, the circuit is operated at a clock frequency higher than the application throughput  $f_{op}$ .

Under these conditions  $E_{op}$  is dominated by  $E_{dyn}$  and Fig. 1.8(b) shows it is kept constant over the whole throughput range according to the expression of  $E_{dyn}$  from Eq. (1.11). Sleep-mode scheme brings energy saving as compared to FVS scheme only for very-low-throughput applications (below 10 kOp/s) down in  $R3$  region, where static energy dominates.

Nevertheless, sleep mode can be combined with FVS in order to further decrease energy per operation. As sketched in Fig. 1.9, there is no point operating in FVS with a sleep-mode scheme when in  $R1$  region. In this region, the sleep-mode operation requires to raise somewhat  $V_{dd}$  to operate at a clock frequency higher than  $f_{op}$  in order to complete the operation in advance and then switch to sleep mode. The raise of  $V_{dd}$  increases  $E_{dyn}$ , which outweighs  $E_{stat}$  saving. However, when in  $R2-R3$  regions, the  $E_{dyn}$  overhead due to the raise of  $V_{dd}$  is negligible as compared to  $E_{stat}$  saving. As shown in Fig. 1.8, the optimum option for low-throughput applications is to minimize the active-mode energy consumption by operating at the minimum-energy point i.e. at the corresponding  $f_{clk,opt}$  and  $V_{dd,opt}$ , and then to enter sleep mode after completion of the operation. Considering an ideal sleep mode, this should provide  $E_{min}$  energy consumption regardless of the application throughput, provided that it enables operating at minimum-energy point ( $f_{op} < f_{clk,opt}$ ). The leakage-gating techniques will be addressed in Chapter 4 and we thus do not consider sleep-mode scheme in the remainder of this preliminary chapter.

#### 1.4.4 Impact of the temperature

ULP circuits are less subject to self-heating than high-performance circuits because they consume less power and thus generate less heat [27]. Nevertheless, they may be operated in an environment with a temperature different than room temperature. We have thus to impose that the circuit will still meet both robustness and throughput constraints when the temperature is changed. Fig. 1.10(a) shows minimum  $V_{dd,ind}$  that meets the constraints for an industrial -25/+85°C temperature range, as compared to  $V_{dd,RT}$  that meets the constraints at 25°C. Robustness-limited minimum  $V_{dd,ind}$  is slightly increased for 85°C operation because of degraded subthreshold swing when the temperature increases. Above 0.6V, throughput-limited minimum  $V_{dd,ind}$  is increased for 85°C operation because the channel mobility degradation with temperature implies a delay increase in superthreshold regime, while under 0.6V throughput-limited minimum  $V_{dd}$  is increased for -25°C operation because the subthreshold current is reduced at low temperatures as explained in Section 1.3.2. As shown in Fig. 1.10(b)(c), the operation at  $V_{dd,ind}$  instead of  $V_{dd,RT}$  has a weak impact on practical power and energy consumptions.

Fig. 1.10(b)(c) shows that operating at 85°C does not change the picture. It increases  $P_{stat}$  and  $E_{stat}$  through leakage currents, which shifts somewhat the limit between dynamic-dominated  $R1$  and static-dominated  $R2$  throughput regions and increases  $P_{min}$  and  $E_{min}$  levels. Nevertheless, observations remain valid. In this dissertation, we therefore mainly focus on room-temperature operation for the sake of simplicity, unless otherwise specified.

If the circuit has to operate in a real high-temperature environment (> 150°C), Fig. 1.10(b)(c) suggests that the power/energy can be completely dominated by static component and that FVS scheme will thus not be efficient anymore. This specific topic will be addressed in depth in Chapter 5.

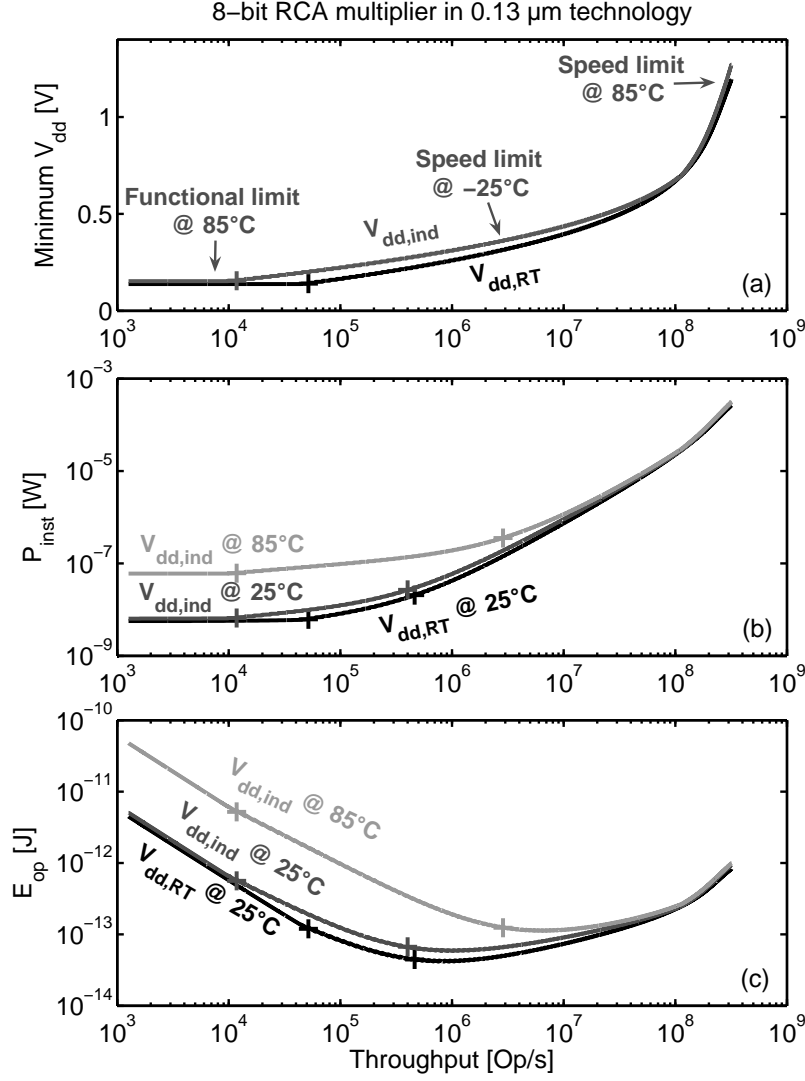
#### 1.4.5 Impact of circuit/application parameters

In order to validate previous observations, let us consider the cases of a different circuit architecture or a different application type leading to different parameters: activity factor and duty cycle.

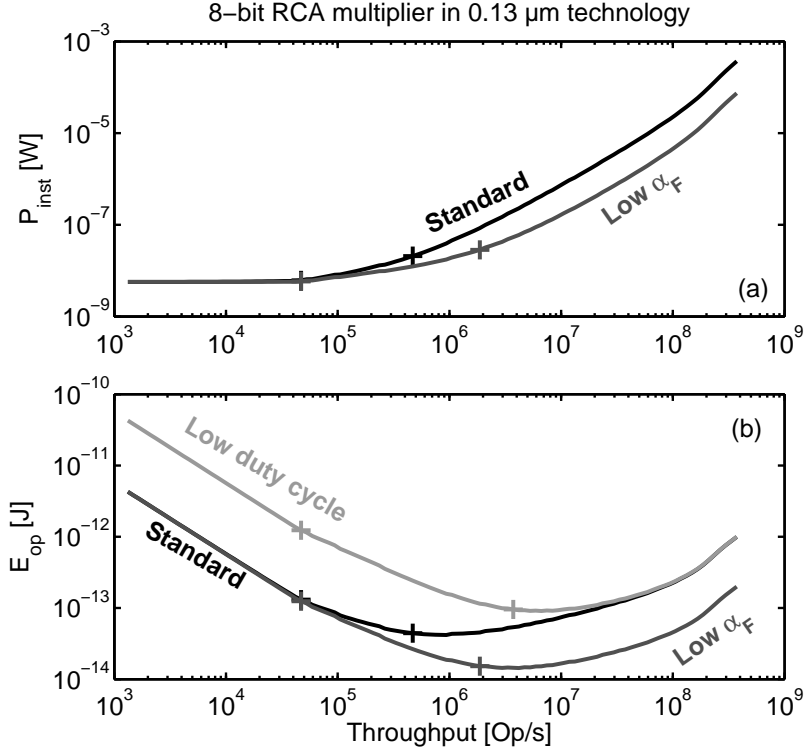
An RCA multiplier typically features a high activity factor  $\alpha_F$  due to unbalanced paths in the partial-product summation array, which results in parasitic activity (glitches) [28]. Let us see what happens in circuits with lower  $\alpha_F$ . To do so, we resimulate the benchmark multiplier with an artificial  $\alpha_F$  reduced by a factor 10 through slowing down of the input pattern. As expected, Fig. 1.11 shows that the resulting  $P_{dyn}$  and  $E_{dyn}$  are divided by 10. This does not impact  $P_{min}$ , which is dominated by  $P_{stat}$ . However, it results in a lower  $E_{min}$  and shifts the boundary between  $R1$  and  $R2$  to higher application throughputs.

Duty cycle is an application parameter: the time ratio between active and stand-by periods of the circuit. It does not affect maximum instantaneous power because maximum  $P_{inst}$  is reached in active periods. However, a low duty cy-





**Fig. 1.10.** Impact of the temperature: minimum  $V_{dd}$  (a) under throughput and robustness constraints at 25°C room temperature ( $V_{dd,RT}$ ), and for an industrial -25/+85°C temperature range ( $V_{dd,ind}$ ) with corresponding instantaneous power (b) and energy per operation (c) at 25°C under  $V_{dd,RT}$  and  $V_{dd,ind}$ , and at 85°C under  $V_{dd,ind}$  (Spice simulation of an 8-bit RCA benchmark multiplier in 0.13  $\mu\text{m}$  standard bulk technology). Cross markers indicate the frontiers between  $R1$ ,  $R2$  and  $R3$  throughput regions.



**Fig. 1.11.** Impact of circuit/application parameters: instantaneous power (a) and energy per operation (b) under FVS scheme with low activity factor  $\alpha_F/10$ , and low duty cycle i.e. circuit operation required 10% of the time (Spice simulation of an 8-bit RCA benchmark multiplier in 0.13  $\mu\text{m}$  standard bulk technology). Cross markers indicate the frontiers between  $R1$ ,  $R2$  and  $R3$  throughput regions.

cle increases  $E_{\text{op}}$  through  $E_{\text{stat}}$  [29]. It results in a higher  $E_{\text{min}}$  and a shift of the  $R1$ - $R2$  boundary to higher application throughputs. Notice that the use of an ideal sleep-mode scheme fully suppress the  $E_{\text{stat}}$  overhead of low-duty-cycle applications, bringing the same  $E_{\text{min}}$  as in an application with unity duty cycle.

Similarly to operating temperature, circuit parameters quantitatively modify the power and energy levels as well as the boundaries between throughput regions but they do not change the qualitative picture. This shows that the proposed analysis scheme remains valid for a wide range of circuit architectures and application types. Without loss of generality, we therefore stick in this dissertation to the standard case of nominal  $\alpha_F$  and unity duty cycle, unless otherwise specified.

## 1.5 CONCLUSION

In this preliminary chapter, we reviewed power/energy consumption of digital circuits. We confirmed the important benefit brought by static FVS scheme when moving from high-performance to ULP applications. Moreover, minimum energy per operation makes subthreshold operation very promising for applications beyond the niche market of pure ULP applications. Indeed, operation at minimum-energy point is desirable for DFVS circuits or highly-parallelized architectures for more general low-power/wireless applications.

We proposed a new analysis framework of energy efficiency under robustness and throughput constraints, which divides application throughput range in three regions: energy-efficient *R1* region, energy-inefficient throughput-limited *R2* region and energy-inefficient robustness-limited *R3* region. In  $0.13\ \mu\text{m}$  technology, a subthreshold circuit for ULP applications lies either in *R1*, *R2* or *R3* regions depending on the application throughput from medium (1-10 MOp/s) to low values (10-100 kOp/s). In Chapter 2, we investigate the impact of technology scaling in the nanometer era within the proposed framework and point out 2 issues: increase of minimum-energy level and degradation of energy efficiency, that we try to fix in 45 nm technology in Chapters 3 and 4.

Finally, we suggested that high-temperature operation ( $> 150^\circ\text{C}$ ) pushes circuits in *R2/R3* regions for the whole throughput range of ULP applications, with a dramatic impact on power/energy consumption. In Chapter 5, we analyze high-temperature operation in depth with an SOI technology and propose a new ULP logic style to mitigate static power/energy at high temperature.

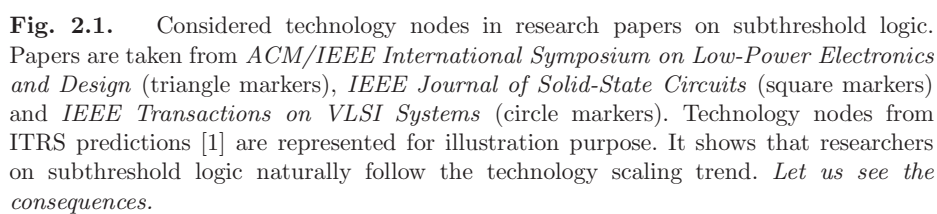
## REFERENCES

1. A.P. Chandrakasan, S. Sheng and R. W. Brodersen: "Low-power CMOS digital design", in *IEEE J. Solid-State Circuits*, vol. 27 (4), pp. 473-484, Apr. 1992.
2. H. Soeleman and K. Roy, "Ultra-low power digital subthreshold logic circuits", in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 94-96, 1999.
3. M. Hempstead, G.-Y. Wei and D. Brooks: "Architecture and circuit techniques for low-throughput energy-constrained systems across technology generations", in *Proc. Int. Conf. Compilers, Architecture and Synthesis for Embedded Systems*, pp. 368-378, 2006.
4. C. H. Stapper, F. M. Armstrong and K. Saji, "Integrated circuit yield statistics", in *Proc. IEEE*, vol. 71, no. 4, pp. 45-470, Apr. 1983.
5. T. S. Barnett, M. Grady, K. G. Purdy and A. D. Singh, "Combining negative binomial and weibull distributions for yield and reliability prediction", in *IEEE Des. Test Computers*, vol. 23, no. 2, pp. 110-116, Mar.-Apr. 2006.
6. B. Zhai, S. Hanson, D. Blauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 20-25, 2005.
7. B. H. Calhoun, A. Wang and A. P. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits", in *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778-1786, Sep. 2005.
8. J. Kwong and A. P. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits", in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 8-13, 2006.
9. Y. Pu, J. Pineda de Gyvez, H. Corporaal and Y. Ha, "Statistical noise margin estimation for sub-threshold combinatorial circuits", in *Proc. IEEE/ACM Asia South Pacific Des. Autom. Conf.*, pp. 176-179, 2008.
10. Y. S. Dhillon, A. U. Diril, A. Chatterjee and A. D. Singh, "Analysis and optimization of nanometer CMOS circuits for soft-error tolerance", in *IEEE Trans. VLSI Syst.*, vol. 14, no. 5, pp. 514-524, May 2006.
11. T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas", in *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584-594, Apr. 1990.
12. K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang and J. D. Meindl "A physical alpha-power law MOSFET model", in *Proc. IEEE/ACM Int. Symp. Low-Power Electronics Des.*, pp. 218-222, 1999.
13. D. A. Hodges, H. G. Jackson and R. A. Saleh, *Analysis and design of digital integrated circuits*, Mc Graw-Hill, 2003.
14. H. J. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits", in *IEEE J. Solid-State Circuits*, vol. 19, no. 4, pp. 468-473, Aug. 1983.
15. K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits", in *Proc. IEEE*, vol. 91, no 2, pp. 305-327, Feb. 2003.

16. Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko and Y. C. Cheng, "Threshold voltage model for deep-submicrometer MOSFET's", in *IEEE Trans. Electron Dev.*, vol. 40, no. 1, pp. 86-95, Jan. 1993.
17. B. Yu, E. Nowak, K. Noda and C. Hu, "Reverse short-channel effects and channel-engineering indeep-submicron MOSFETs: modeling and optimization", in *Dig. Tech. Papers VLSI Technology*, pp. 162-163, 1996.
18. B. Yu, H. Wang, C. Riccobene, Q. Xiang and M. R. Lin, "Limits of gate oxide scaling in nano-transistors", in *Dig. Tech. Papers VLSI Technology*, pp. 90-91, 2000.
19. B. H. Lee, S. C. Song, R. Choi and P. Kirsch, "Metal electrode/high-k dielectric gate-stack technology for power management," in *IEEE Trans. Electron Dev.*, vol. 55, no. 1, pp. 8-20, Jan. 2008.
20. K. Mistry *et al.*, "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging", in *Dig. IEEE Int. Electron Dev. Meeting*, pp. 247-250, 2007.
21. J.-P. Colinge, *Silicon-On-Insulator: Materials to VLSI, 3rd Edition*, Kluwer Academic Publishers, 2004.
22. L. T. Clark, R. Patel and T. S. Beatty, "Managing standby and active mode leakage power in deep sub-micron design", in *Proc. IEEE/ACM Int. Symp. Low-Power Electronics Des.*, pp. 274-279, 2004.
23. D. Flandre, "Silicon-on-insulator technology for high temperature metal oxide semiconductor devices and circuits", in *High-Temperature Electronics*, IEEE Press, Ed. R. Kirschman, pp. 303-308, 1998.
24. D. Flandre *et al.*, "Fully depleted SOI CMOS technology for heterogeneous micropower, high-temperature or RF microsystems", in *Solid-State Electronics*, vol. 45, no. 4, pp. 541-549, Apr. 2001.
25. J. T. Kao, M. Masayuki and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture", in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545-1554, Nov. 2002.
26. A. Wang, S. Kosonocky and A. P. Chandrakasan, "Optimal supply and threshold scaling for subthreshold CMOS circuits", in *Proc. IEEE Comp. Soc. Annual Symp. VLSI*, pp. 5-9, 2002.
27. R. Kumar and V. Kursun, "Temperature-adaptive energy reduction for ultra-low power-supply-voltage subthreshold logic circuits", in *Proc. IEEE Int. Conf. Electron. Circ. Syst.*, pp. 1280-1283, 2007.
28. K. S. Chong, B.-H. Gwee and J. S. Chang, "A micropower low-voltage multiplier with reduced spurious switching", in *IEEE Trans. VLSI Syst.*, vol. 13, no. 2, pp. 255-265, Feb. 2005.
29. M. Seok, S. Hanson, D. Sylvester and D. Blaauw, "Analysis and optimization of sleep modes in subthreshold circuit design", in *Proc. ACM/IEEE Des. Autom. Conf.*, pp. 694-699, 2007.



# IMPACT OF TECHNOLOGY SCALING ON SUBTHRESHOLD LOGIC



## Abstract

---

We focus on subthreshold logic and analyze the impact of CMOS technology scaling from  $0.25\ \mu\text{m}$  to 32 nm node [JP2][CP3]. The analysis is first carried out at device level. It shows that worst-case subthreshold  $I_{on}$  increases with constant-field scaling trend until 90 nm node and then saturates because of subthreshold swing, drain-induced barrier lowering (DIBL) and variability increase. Fringing capacitances due to slow scaling of gate-stack height also exhibit a worrying increase.

At circuit level, the analysis shows that minimum supply voltage  $V_{dd}$  of subthreshold circuits jumps from speed to robustness limitation when migrating to smaller technology nodes. Instantaneous power consumption in low-throughput applications suffer from an extension of the minimum-power range and the increase of minimum-power level. Regarding energy per operation, we first report that minimum-energy level is reduced when migrating to 90 nm node thanks to dynamic energy reduction. It then increases as static energy does. Second, we show that technology scaling shifts the minimum-energy point towards higher throughput values. This shift combined with the reduction of minimum-energy level enables considerable practical energy savings at medium throughputs when migrating to 90/65 nm nodes. However, at 45/32 nm nodes, this benefit is outweighed by static energy. Moreover, for low-throughput applications, practical energy increases by 2 orders of magnitude when migrating from 180/90 nm to 45/32 nm node.

## Contents

---

2.1	Introduction	29
2.2	Technology scaling	30
2.3	Impact on MOSFET subthreshold operation	32
2.4	Impact on subthreshold logic	40
2.5	Results validation	54
2.6	Conclusion	54

---



## 2.1 INTRODUCTION

Based on a  $0.13\mu\text{m}$  technology, we showed in Chapter 1 that frequency/voltage scaling down to the subthreshold region yields important power/energy reduction for ULP applications. These application fields such as sensor networks, RFID tags and biomedical devices, typically require low-cost robust circuits with low-to-medium data throughputs. Therefore, one can imagine that ULP circuit designers would not focus on the most advanced technologies. However, as shown in Fig. 2.1, researchers on subthreshold logic tend to naturally follow the technology scaling trend by considering more and more advanced technology nodes. At the *2008 IEEE International Solid-State Circuits Conference (ISSCC)*, the very first industrial subthreshold circuits were presented: a motion estimation accelerator for video processing from Intel [2] and a general-purpose microcontroller for biomedical applications from a collaboration between the Massachusetts Institute of Technology and Texas Instruments [3]. Both circuits are implemented in recent 65 nm technologies, thereby confirming the trend observed in the academic research world from Fig. 2.1.

Let us recall that subthreshold logic does not only target pure ULP applications. Indeed, as explained in Section I.3, the minimum-energy property makes subthreshold operation also very promising for dynamic frequency/voltage-scaling (DFVS) scheme [4, 5] and highly-parallelized architectures [6, 7], both in mass-production low-power/wireless applications such as laptop computers and cell phones. In these consumer portable applications, the attraction for technology scaling is clearer than in the niche market of pure ULP applications: cost. Indeed, in mass production, the raw material cost i.e. the Silicon wafers is very important. The increased density of scaled technologies thus comes with cost reduction, which is assumed to motivate the trend observed in Fig. 2.1.

From a circuit-performance point of view, on one hand, technology scaling decreases circuit capacitances and thus the power consumption under fixed clock frequency and supply voltage. Moreover, scaled technologies feature an increased MOSFET subthreshold current. As suggested in [8], this translates into a lower subthreshold  $V_{dd}$  value to get a fixed large delay, thereby considerably lowering power consumption.

However, technology scaling also comes with some drawbacks: increase of short-channel effects, leakage currents and variability. Amongst them, variability has been shown to be a severe limitation for subthreshold logic [9]. It is reported in [10] that not only delay and leakage variability has to be taken into account for subthreshold logic, as it is normally the case at nominal (super-threshold) supply voltage  $V_{dd}$ . Indeed, as briefly reported in Chapter 1,  $V_t$  variations exponentially modify  $I_{on}$  and  $I_{off}$  currents, which can induce bad output logic levels. This can in turn imply functional breakdown of some gates, resulting in a bad functional yield of the circuit. It is thus not clear whether technology scaling is actually desirable for subthreshold logic.

In this chapter, the interests and limitations of technology scaling for subthreshold logic are investigated from  $0.25\mu\text{m}$  to 32 nm nodes, first by extensive

subthreshold device modeling and then by thorough circuit-level simulation of a benchmark multiplier. An enhanced version of Predictive Technology Model (PTM) from Arizona State University [11], is used to get smooth scaled MOSFET characteristics over the whole technology range. Results are further validated with production models from an industrial foundry.

In Section 2.2, we briefly review MOSFET scaling theory, the issues in nanometer technologies and the CMOS technologies we consider for this study. In Section 2.3, the impact of technology scaling on the subthreshold operation of MOSFETs is investigated. In Section 2.4, we extend this study to circuit-level investigation and present the interests and limitations of technology scaling for subthreshold logic. We then validate these results in Section 2.5.

## 2.2 TECHNOLOGY SCALING

The scaling of CMOS technologies driven by Moore’s law has been the major driving force of the IC market. Since the beginning of IC industry in the late 1950’s, several scaling models and theories have occurred. In this section, we briefly review the main scaling trend and the scaling issues that appear when reaching the nanometer era. We then present the device models we consider for this study.

### 2.2.1 Scaling theory

The traditional scaling trend well known as “constant-field scaling” was introduced by Dennard *et al.* in 1974 [12]. It is based on the conservation of the electric field in the MOSFET channel. As presented in Table 2.1, this scaling trend assumes that all the physical dimensions of the MOSFETs (channel length  $L$ , width  $W$  and oxide thickness  $T_{ox}$ ) are reduced by a common factor  $\alpha$ . Therefore, the MOSFET area is divided by  $\alpha^2$  and the gate capacitance per device by  $\alpha$ . To keep the electric field in the channel constant, the channel doping  $N_{ch}$  is increased by  $\alpha$  whereas the power supply  $V_{dd}$  and the threshold voltage  $V_t$  are divided by  $\alpha$ . The on-state current  $I_{on}$  per width unit is kept constant, leading to a division by  $\alpha$  of the  $I_{on}$  per device. The gate delay is thus divided by  $\alpha$ , as it is proportional to the intrinsic logic gate delay  $C_g V_{dd} / I_{on}$ . This scaling trend is very efficient in reducing the circuit area and improving its performances. In particular, the dynamic energy per operation, which is proportional to  $C_g V_{dd}^2$ , is divided by  $\alpha^3$ .

### 2.2.2 Technology scaling in the nanometer era

In submicron technologies, the actual scaling trend is different from these simple relationships because of two major historic side-effects: the saturation of carrier velocity and the increasing subthreshold leakage current [13]. Amongst these effects, increasing subthreshold leakage is the main limitation to constant-field

**Table 2.1.** Scaling factors for the basic scaling trends

Parameters and figures of merit	Constant-field scaling [12]	Generalized scaling [14]
$W, L, T_{ox}$	$1/\alpha$	$1/\alpha$
Electric field	1	$\epsilon$
$N_{ch}$	$\alpha$	$\epsilon\alpha$
$V_{dd}, V_t$	$1/\alpha$	$\epsilon/\alpha$
$I_{on}/W$	1	$\epsilon$
$C_g/W$	1	1
Area	$1/\alpha^2$	$1/\alpha^2$
Gate delay	$1/\alpha$	$1/\alpha$
$E_{dyn}$	$1/\alpha^3$	$\epsilon^2/\alpha^3$

scaling in today's nanometer technologies. As subthreshold leakage exponentially depends on  $V_t$  according to Eq. (1.8), static power consumption issues prevent from scaling  $V_t$  by  $\alpha$ . In order to maintain sufficient gate overdrive for speed concern,  $V_{dd}$  is neither scaled by  $\alpha$ . This trend can be modeled by the introduction of an electrical scaling factor  $\epsilon$ . As shown in Table 2.1, the generalized scaling theory [14] tolerates an increase of the electric field in the channel to compensate for the slower scaling of  $V_{dd}$ . Scaling into the nanometer range has a detrimental impact on power consumption. First, even if  $I_{off}$  is limited by the slow scaling of  $V_t$ , it still increases considerably, leading to higher static power consumption. Second, as shown in Table 2.1, the slow scaling of  $V_{dd}$  makes the scaling of the dynamic energy per operation less efficient, being limited by the  $\epsilon$  factor.

With the constant shrinking of the channel length, short-channel effects have appeared such as drain-induced barrier lowering (DIBL) effect and  $V_t$ -roll-off [15]. Moreover, when reaching the sub-100nm nodes, the shrinking of the oxide thickness  $T_{ox}$  leads to an exponential growth of the gate-tunneling leakage current. As gate leakage has to be limited for power consumption concern,  $T_{ox}$  is no longer scaled by  $\alpha$  [13]. The consequence is the decreased control of the channel potential by the gate. These effects have a detrimental impact on the subthreshold swing, which is degraded at nanometer technology nodes [16, 17].

### 2.2.3 Considered device models

In order to get smooth transition between the features of the different technology nodes, we use generic models rather than foundry models. Moreover, for circuit-level simulation time issues, we use compact device models with Spice simulator. The models are Predictive Technology Models<sup>1</sup> (PTM) from Arizona

<sup>1</sup>Models are available on-line at <http://www.eas.asu.edu/~ptm>.

**Table 2.2.** Device parameters from the considered CMOS technologies and  $I_{on}/I_{off}$  currents under nominal supply voltage (high-performance flavor)

Node [nm]	$L_{eff}$ [nm]	$T_{ox}$ [nm]	$V_{dd,nom}$ [V]	$V_{t,nom}$ [V]	$I_{off,nom}$ [nA/ $\mu m$ ]	$I_{on,nom}$ [ $\mu A/\mu m$ ]
250	120	4.0	2.5	0.63	0.002	820
180	70	2.3	1.8	0.49	0.11	840
130	49	1.6	1.3	0.36	4.5	890
90	35	1.4	1.2	0.32	19	1030
65	24.5	1.2	1.1	0.30	62	1150
45	17.5	1.1	1.0	0.27	200	1250
32	12.6	1.0	0.9	0.27	350	1290

State University [11]. These are full model cards for BSIM4 compact MOSFET model [18]. For the sake of generality, we consider high-performance technology trend from ITRS predictions, as it is the main technology driver [1]. The impact of a low-power trend will be discussed in Section 2.5.

PTM models can be customized by the definition of some key parameters for each node:  $L_{eff}$ ,  $T_{ox}$ ,  $V_{dd}$  and  $V_t$ . This has been shown to accommodate a wide range of technologies with different parameter values [11]. We keep default PTM values of  $L_{eff}$  and  $T_{ox}$ . We modify  $V_{dd}$  values at 0.25 and 0.18  $\mu m$  nodes to reflect constant-field scaling trend from 0.25 and 0.13  $\mu m$  node. We then tune  $V_t$  for the whole technology range to get a constant 30% improvement per generation of the intrinsic gate delay  $C_g V_{dd}/I_{on}$ . However, at 32 nm node,  $V_t$  is kept constant for static power issues. The resulting  $C_g V_{dd}/I_{on}$  improvement between 45 and 32 nm nodes is 25%. Device parameters are summarized in Table 2.2. In the remainder of this chapter, we refer to device/circuit characteristics under nominal superthreshold supply voltage  $V_{dd,nom}$  by indexing them with a “nom” subscript (e.g.  $I_{on,nom}$ ). Subthreshold characteristics are referred to without this subscript.

## 2.3 IMPACT ON MOSFET SUBTHRESHOLD OPERATION

In this section, the evolution of the main parameters of MOSFET subthreshold operation is investigated: region of operation, drain current, capacitances and variability.

### 2.3.1 Subthreshold region of operation

As shown in Table 2.2,  $V_t$  is decreased with technology scaling. Subthreshold region of operation gets thus smaller, i.e. the maximum subthreshold  $V_{dd}$  decreases. However, as we are more concerned about achieving ultra-low power/energy

rather than having purely subthreshold operation, we consider the same  $V_{dd}$  range for all the nodes: from 0.2 to 0.5V.

### 2.3.2 Subthreshold drain current

Let us recall the subthreshold drain current expression per width unit from Eq. (1.10):

$$I_{sub} = I_0 \times 10^{\frac{V_{gs} + \eta V_{ds}}{S}} \times \left(1 - e^{\frac{-V_{ds}}{U_{th}}}\right) \quad (2.1)$$

where  $S$  is the subthreshold swing,  $\eta$  the DIBL coefficient and  $I_0$  a reference current per width unit, which exponentially depends on  $V_t$ . At a given temperature,  $I_{sub}$  only depends on three parameters:  $I_0$ ,  $S$  and  $\eta$ . If  $I_0$  value is known, notice that the exact  $V_t$  value is not required for drain current modeling, provided that the devices actually stay in subthreshold regime. The evolution of  $I_0$ ,  $S$  and  $\eta$  with technology scaling extracted from PTM BSIM4 models is shown in Fig. 2.2 and discussed below.

#### Subthreshold swing

For long-channel devices, the subthreshold swing can be expressed as [18]:

$$S = \ln(10) \times U_{th} \times \left(1 + \frac{\epsilon_{Si}}{\epsilon_{ox}} \frac{T_{ox}}{X_{dep}}\right), \quad (2.2)$$

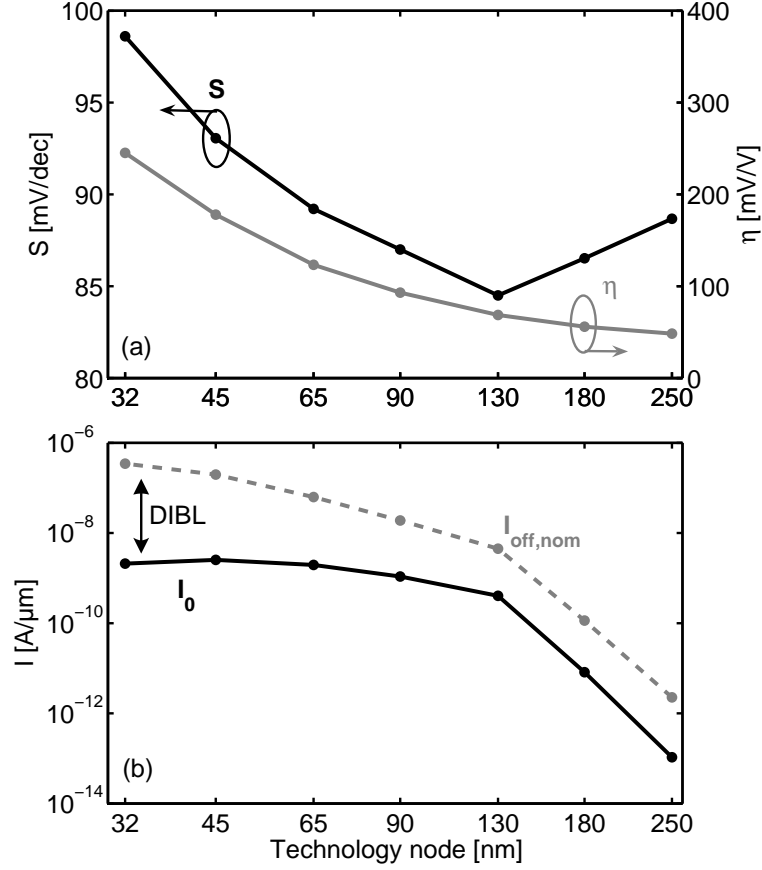
where  $X_{dep}$  is the channel depletion depth,  $\epsilon_{Si}$  and  $\epsilon_{ox}$  the dielectric permittivity of Silicon and gate oxide, respectively.

In PTM BSIM4 model, the short-channel degradation of  $S$  is taken into account by a multiplicative factor named **Nfactor** ( $> 1$ ). **Nfactor** has been fitted vs.  $L_{eff}$  with published data from numerous technologies [11]: **Nfactor** increases smoothly with reducing  $L_{eff}$ , thereby degrading  $S$ . However, as depicted in Fig. 2.2,  $S$  improves from 0.25 to 0.13  $\mu\text{m}$  nodes, even with the increase of **Nfactor**. This is due to the aggressive scaling of  $T_{ox}$ , according to constant-field scaling trend, which is faster than the scaling of  $X_{dep}$  proportional to  $1/\sqrt{N_{ch}}$ . As shown in Table 2.2,  $T_{ox}$  scaling slows down from 0.13  $\mu\text{m}$  node to limit gate leakage. Therefrom,  $S$  deteriorates fast, which has a bad impact on subthreshold operation by lowering  $I_{on}/I_{off}$  ratio as shown with device simulations in [17]. Notice that this scenario could be pessimistic as the introduction of high- $\kappa$ /metal-gate devices at 45 or 32 nm node may lead to smaller equivalent  $T_{ox}$  and thus improved  $S$ . However, for the sake of generality, we only consider standard oxide material in this dissertation.

#### Drain-induced barrier lowering

The DIBL coefficient can be expressed as [15]:

$$\eta = \frac{1}{2 \cosh(L_{eff}/2l_t)} \text{ with } l_t = \sqrt{\frac{\epsilon_{Si} T_{ox} X_{dep}}{\epsilon_{ox} \beta}}, \quad (2.3)$$



**Fig. 2.2.** Evolution of the parameters of subthreshold current from Eq. (2.1): subthreshold swing  $S$ , DIBL coefficient  $\eta$  (a) and reference subthreshold current  $I_0$  (b). Starting at 90 nm node,  $S$  deteriorates due to the relatively slow scaling of  $T_{ox}$ . DIBL increases severely with the scaling of  $L_{eff}$ .  $I_0$  is compared to off-state current  $I_{off,nom}$  at nominal superthreshold  $V_{dd,nom}$ . DIBL effect implies an increasing difference between  $I_{off,nom}$  and  $I_0$ .

where  $l_t$  is a characteristic length and  $\beta$  a fitting parameter. For the considered 45 nm node, the value of  $l_t$  is close to 10 nm. When  $L_{eff} \ll l_t$ , Eq. (2.3) can be simplified as:

$$\eta = e^{\frac{-L_{eff}}{2l_t}} + 2e^{\frac{-L_{eff}}{l_t}}, \quad (2.4)$$

which clearly shows a negative exponential dependence on  $L_{eff}$ . In practice,  $L_{eff} > l_t$  and Eq. (2.3) cannot be simplified but the dependence remains close to exponentiality.

In BSIM4 model, another coefficient namely **Eta0** ( $< 1$ ) is added to take into account channel and source/drain engineering such as pocket/halo doping to mitigate the DIBL effect [18]. **Eta0** simply multiplies the expression of  $\eta$  from Eq. (2.3). In PTM model cards, **Eta0** has been fitted vs.  $L_{eff}$  with published data from numerous technologies [11]. **Eta0** decreases strongly for each generation. However, Fig. 2.2 shows that the increasing device-engineering level is not sufficient to mitigate the DIBL effect increase with technology scaling, leading to  $\eta$  values higher than 150 mV/V as in [19] or even higher than 200 mV/V as in [20]. This is due to the relatively slow scaling of  $T_{ox}$  and  $X_{dep}$  and thus  $l_t$  as compared to  $L_{eff}$ .

#### Subthreshold reference current

When rewriting Eq. (2.1) with  $V_{gs} = 0$  and  $V_{ds} = V_{dd,nom}$ , the subthreshold reference current  $I_0$  per width unit is related to  $I_{off,nom}$  at  $V_{dd,nom}$  from Table 2.2, which is a device design target:

$$I_0 = I_{off,nom} \times 10^{\frac{-\eta V_{dd,nom}}{S}}. \quad (2.5)$$

Therefrom,  $I_0$  considerably increases with constant-field scaling between 0.25 and 0.13  $\mu\text{m}$  nodes as  $I_{off,nom}$  does, when  $\eta$  remains small. It then reaches a plateau as DIBL induces an increasing difference between  $I_0$  and  $I_{off,nom}$ . At 32 nm node,  $I_0$  even starts decreasing to compensate for the considerable increase of  $\eta$ , while keeping  $I_{off,nom}$  reasonable, for static power consumption issues.

#### 2.3.3 Capacitances in subthreshold regime

The capacitances strongly influence gate delay and dynamic energy consumption. The intrinsic superthreshold gate capacitance per width unit  $C_{g,nom}$  is represented in Fig. 2.4, as extracted from PTM BSIM4 models at  $V_{gs} = V_{ds} = V_{dd,nom}$ . Its evolution follows the one of  $C_{ox} = \epsilon_{ox} L_{eff} / T_{ox}$ : it remains roughly constant from 0.25 to 0.13  $\mu\text{m}$  node with constant-field scaling because  $T_{ox}$  scales at the same pace as  $L_{eff}$ , and starting at 90 nm node, it strongly decreases because of slower  $T_{ox}$  scaling. In subthreshold regime, the intrinsic gate capacitance  $C_{g,sub}$  is the series connection of  $C_{ox}$  and the channel depletion capacitance  $C_{dep}$ .  $C_{dep}$  is lower than  $C_{ox}$  and resulting  $C_{g,sub}$  thus mainly depends on  $C_{dep}$ . As shown in Fig. 2.4,  $C_{g,sub}$  extracted at  $V_{gs} = V_{ds} = 0.2V$  is lower than  $C_{g,nom}$  and decreases smoothly with technology scaling.

As  $C_{g,sub}$  is quite low, parasitic capacitances are even more important in subthreshold than in superthreshold regime. They have thus to be considered carefully. The device parasitic capacitances are depicted in Fig. 2.3 and discussed below:

- junction capacitance  $C_j$ ,
- overlap gate capacitance  $C_{ov}$ ,
- inner fringing gate capacitance  $C_{if}$ ,
- outer fringing gate capacitances  $C_{of,side}$ ,  $C_{of,top}$  and  $C_{of,dif}$ .

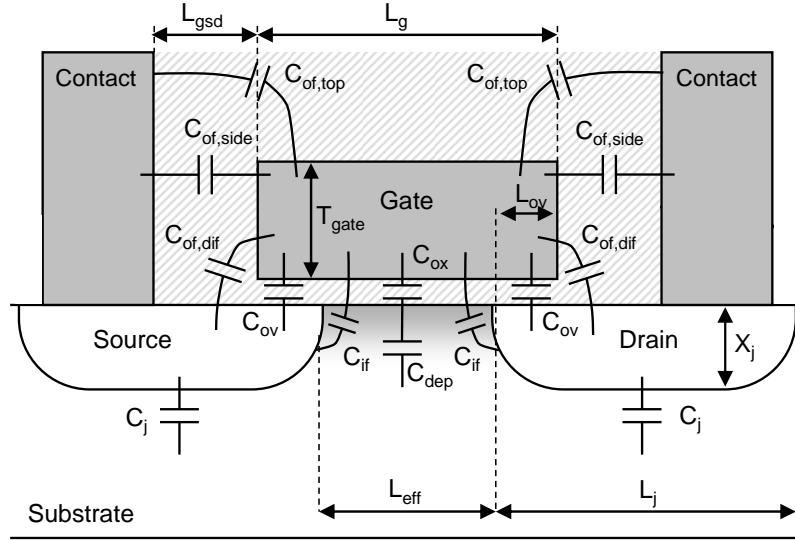


Fig. 2.3. MOSFET capacitances in subthreshold regime

The overlap and fringing capacitances are grouped under the  $C_{g,par}$  term of parasitic gate capacitances. In PTM model cards, parasitic capacitances are roughly modeled with default BSIM parameter values, which is fine when considering devices in superthreshold regime. Because of the greater importance of parasitic capacitances in subthreshold regime, we introduce new parameter values in PTM model cards to accurately predict them. The evolution of their values per width unit is represented in Fig. 2.4.

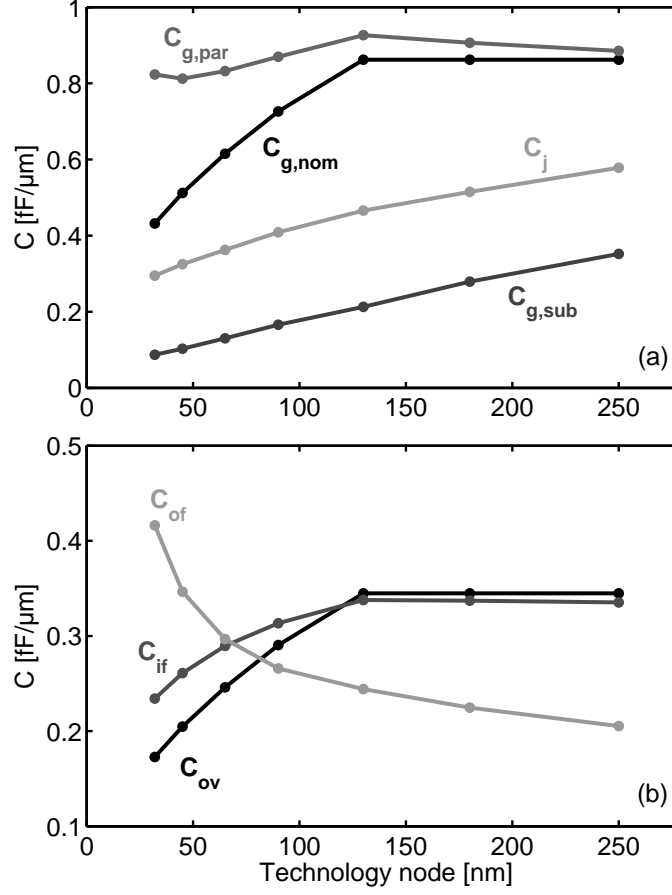
#### Junction capacitance

Source/drain junction capacitances are composed of bottom and side-wall parts. Capacitance values depend on the doping level of the substrate and of the source/drain diffusion. In [21], junction capacitance per area  $C_{js}$  is shown to grow with technology scaling because of higher doping levels. In our model, we choose a typical value of  $0.0007 F/m^2$  for  $C_{js}$  at  $0.25 \mu m$  node [21] and assume a linear growth of 30% per generation. This is a worst-case choice because substrate doping level is hardly increased that fast in practice. The junction depth  $X_j$  is taken from PTM model cards and the junction length  $L_j$  is set at  $2.5 \times \lambda$ , where  $\lambda$  is half the metall pitch. As shown in Fig. 2.4(a), the reduction of  $L_j$  and  $X_j$  results in a decrease of  $C_j$  per width unit, even if  $C_{js}$  per area unit increases.

#### Overlap capacitance

The overlap capacitance per width unit  $C_{ov}$  is given by  $\epsilon_{ox} L_{ov}/T_{ox}$ , with  $L_{ov}$  the overlap length. In [21],  $C_{ov}$  per side has been shown to be roughly equivalent





**Fig. 2.4.** Evolution of the capacitances per width unit: (a) intrinsic superthreshold  $C_{g,nom}$  at  $V_{dd,nom}$ , intrinsic subthreshold  $C_{g,sub}$  at 0.2V, extrinsic  $C_{g,par}$  gate capacitances and junction capacitance  $C_j$ , (b) components of  $C_{g,par}$ : overlap capacitance  $C_{ov}$ , inner  $C_{if}$  and outer  $C_{of}$  fringing capacitances. For  $C_{g,par}$  and its components, the sum of both drain and source sides is considered.

to 20% of the intrinsic gate capacitance  $C_{g,sup}$ , for minimum-length devices for a wide range of technology nodes. Therefore, we set  $L_{ov}$  to 20% of minimum channel length. The resulting  $C_{ov}$  follows the evolution of  $C_{ox}$  and  $C_{g,nom}$ .

#### Fringing capacitance

The last parasitic gate capacitance category is fringing capacitance, which becomes more and more important with technology scaling because of the proximity of source/drain contacts [23]. It is composed of the inner fringing capacitance through the channel  $C_{if}$  and the outer fringing capacitances: from gate side to

the contacts  $C_{of,side}$ , from gate top to the contacts  $C_{of,top}$  and from gate side to the diffusions  $C_{of,dif}$ . From [23] and [22], these capacitances per width unit are modeled by:

$$C_{if} = \frac{2 \epsilon_{Si}}{\pi} \ln \left( 1 + \frac{X_j}{2 T_{ox}} \right), \quad (2.6)$$

$$C_{of,side} = \epsilon_{ox} \left( \frac{T_{ox} + T_{gate}}{L_{gsd}} - 0.55 \right), \quad (2.7)$$

$$C_{of,top} = \frac{2 \times 0.8 \epsilon_{ox}}{\pi} \ln \left( 1 + \frac{L_g}{2 L_{gsd}} \right), \quad (2.8)$$

$$C_{of,dif} = \frac{0.8 \epsilon_{ox}}{\pi} \ln \left[ (M^2 - 1) \left( \frac{M^2}{M^2 - 1} \right)^{M^2} \right] \quad (2.9)$$

where  $X_j$  is the source/drain diffusion depth,  $T_{gate}$  the gate electrode height,  $L_g$  the printed gate length  $L_{gsd}$  the distance between the gate and the source/drain contacts and  $M$  is defined as  $L_{gsd}/T_{ox}$ . Again,  $X_j$  values come from PTM model cards and we set  $L_{gsd}$  to  $1 \times \lambda$  (half the metal pitch). In [23],  $T_{gate}$  is presented as constant. However, in ITRS prediction  $T_{gate}$  is supposed to scale by 30% per generation, which is the ideal case. As the total gate stack height is hardly scaled that fast in practice, we rather make an intermediate assumption considering that  $T_{gate}$  is reduced by 15% per generation. The resulting total  $C_{of}$  per width unit considerably increases with technology scaling, whereas  $C_{if}$  roughly follows  $C_{ox}$  trend, as shown in Fig. 2.4(b).

From this figure, it is clear that the important role of parasitic capacitances is exacerbated in subthreshold regime. At smallest technology nodes, outer fringing capacitance is dominant because of the aggressive reduction of  $L_{gsd}$  gate-to-source/drain distance. It is therefore fundamental to consider it carefully when making subthreshold logic simulations.

### 2.3.4 Variability in subthreshold regime

Device variability affects on- and off-state drain currents. It can be divided into two categories: intrinsic and extrinsic variability, depending on the origin of the parameter fluctuation. Notice that capacitances may also fluctuate with device variations but this is beyond the scope of this discussion.

#### *Intrinsic variability*

Intrinsic variations come from the atomistic nature of nanometer devices. The main sources are random dopant fluctuation (RDF), line edge roughness and  $T_{ox}$  variations, all of them implying  $V_t$  variability [24]. Amongst them, RDF has the strongest impact on  $V_t$  variability. It is shown in [26] that, following ITRS guidelines, the  $\sigma_{V_t}$  due to line edge roughness remains small as compared to  $\sigma_{V_t}$  due to RDF, although in practice it is expected to play an important role in sub-32 nm nodes. It is also shown that  $\sigma_{V_t}$  due to  $T_{ox}$  variation is less than the half

of  $\sigma_{V_t}$  due to RDF. As these sources are not correlated, they add in quadrature. The resulting  $\sigma_{V_t}$  due to both RDF and  $T_{ox}$  variations is thus only 10% higher than  $\sigma_{V_t}$  due only to RDF. Therefore, in order to keep the discussion simple, we consider RDF as the lower bound of intrinsic variability. From [24], we model the impact of RDF through a normal distribution of  $V_t$  with:

$$\sigma_{V_t, RDF} = \frac{A_{V_t, RDF}}{\sqrt{W} L_{eff}} = 3.19 \times 10^{-8} \frac{T_{ox} N_{ch}^{0.4}}{\sqrt{W} L_{eff}}. \quad (2.10)$$

As RDF are random by definition, there is no spatial correlation as experimentally verified in [25]. We thus consider the  $V_t$  of each device as an independent normally-distributed variable. The resulting values of  $\sigma_{V_t, RDF}$  for minimum-sized devices are enclosed in Fig. 2.5. Variability increases for the whole technology range due to channel area scaling.

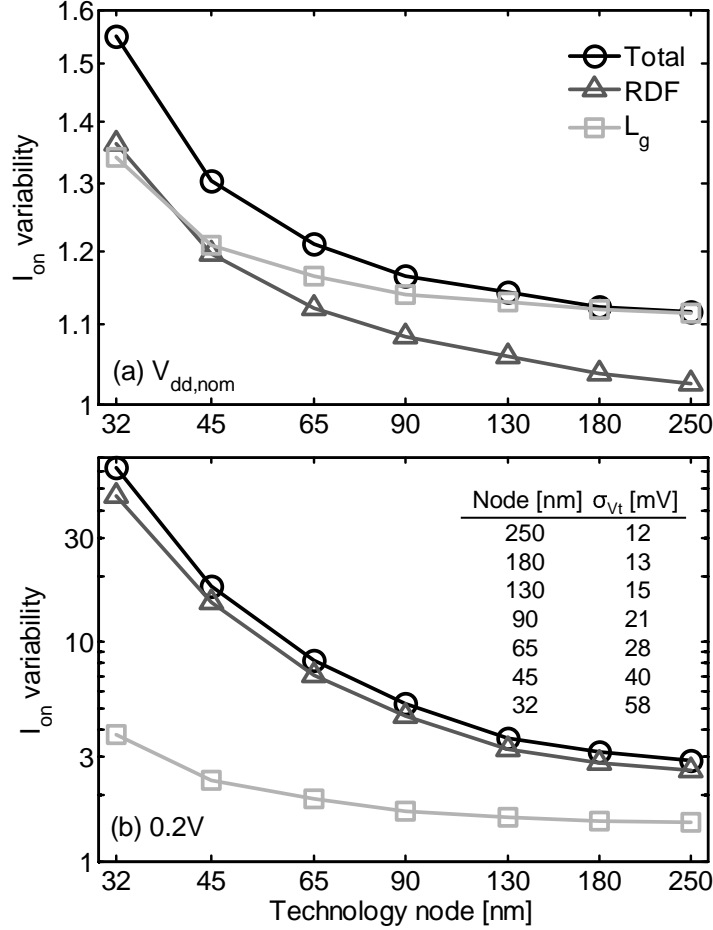
#### *Extrinsic variability*

The manufacturing process implies global  $V_t$  variations and fluctuations of critical dimensions (CD).  $V_t$  variations severely affect subthreshold circuits. However, owing to their global behavior, they can be efficiently compensated through adaptive body biasing (ABB) [27, 28]. ABB has become a common technique for subthreshold design and we therefore do not consider extrinsic global  $V_t$  variations by making the assumption that the circuit relies on ABB. The efficiency of ABB in nanometer subthreshold circuits is further analyzed in Chapter 4. Amongst CD variations, variations in printed gate length  $L_g$  have the strongest impact on  $I_{on}$  through  $L_{eff}$  and thus  $V_t$  modulation because of  $V_t$  roll-off and DIBL effects. Therefore, we consider  $L_g$  variability as extrinsic variability source, as it is usually the case [29]. We model  $L_g$  variations as a normal distribution with  $3\sigma/\mu$  value of 20%. Notice that  $L_g$  variations directly affect  $L_{eff}$  and thus  $\sigma_{L_g} = \sigma_{L_{eff}}$ . Variations of  $L_g$  have a strong spatial correlation [29] and the benchmark circuit from Section 2.4 is quite small. Therefore, we consider only one normally-distributed  $L_g$  variable, common to all the devices of the circuit.

#### *Subthreshold-current variability*

Monte-Carlo simulations with 10k runs have been performed at both nominal superthreshold  $V_{dd, nom}$  and subthreshold  $V_{dd}$ . The impact of RDF and  $L_g$  variations on  $I_{on}$  is shown in Fig. 2.5 for minimum-sized devices, through the ratio between typical and worst-case  $I_{on}$ . Worst-case condition is defined with a confidence interval of 99.9% throughout this dissertation. Notice that superthreshold current is modeled by a normal distribution, whereas subthreshold current is modeled by a lognormal distribution as it exponentially depends on normally-distributed  $V_t$  [9]. The 99.9% confidence interval corresponds to  $\mu \pm 3\sigma$  value for normal distributions and  $\mu \times \sigma^{\pm 3}$  for lognormal distributions.

At  $V_{dd, nom}$ ,  $L_g$  variations are dominant in largest technology nodes, whereas with technology scaling the impact of RDF becomes comparable to  $L_g$  variations. However, at subthreshold 0.2V  $V_{dd}$ , RDF is dominant for the whole technology

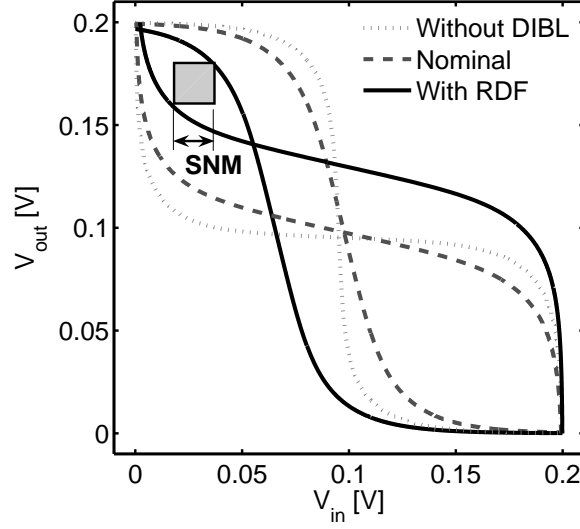


**Fig. 2.5.** Current variability of minimum-sized devices (drawn  $L_g = L_{min}$ ,  $W = 1.5 L_{min}$ ) under nominal superthreshold  $V_{dd}$  (a) and subthreshold  $0.2V$   $V_{dd}$  (b) with enclosed absolute  $\sigma_{V_t, RDF}$  resulting from Eq. (2.10). Variability is expressed as the ratio between typical and  $3\sigma$  worst-case  $I_{on}$ . In subthreshold regime, variability is dominated by random dopant fluctuation.

range. It implies a subthreshold  $I_{on}$  variability up to a factor 60 at 32 nm node. This is much larger than at  $V_{dd,nom}$  because of exponential dependence of drain current on  $V_t$  in subthreshold regime, as suggested in [9].

## 2.4 IMPACT ON SUBTHRESHOLD LOGIC

Based on the observations we made at device level in Section 2.3, we investigate the impact of technology scaling on subthreshold logic through simulation of



**Fig. 2.6.** Inverter voltage transfer curve at 45 nm node (minimum-sized devices,  $V_{dd} = 0.2V$ ). SNM is reduced by DIBL and mismatch between pull-up NMOS and pull-down MOS due to RDF.

the 8-bit RCA multiplier as benchmark circuit. We use the same investigation scheme as in Chapter 1: first robustness and throughput constraints given by functional yield and delay, then components of instantaneous power, minimum energy per operation, and finally practical power and energy under robustness and throughput constraints.

#### 2.4.1 Static noise margins and functional yield

Static noise margin (SNM) of a logic gate is defined as the voltage margin between its output low logic level  $V_{OL}$  and the highest input voltage interpreted as low logic level  $V_{IL}$ :  $SNM = V_{IL} - V_{OL}$ , and vice-versa with  $V_{OH}$  and  $V_{IH}$ . A negative SNM means a functional breakdown of the gate, the output being stuck at high or low logic level, whatever the input level (comprised between 0 and  $V_{dd}$ ), as explained in Section 1.2.1. We have seen in this section that the low  $I_{on}/I_{off}$  ratio of subthreshold logic combined with variability can imply bad output logic levels, which would not be recognized as the correct logic level by next gates, leading to functional failure. Let us first examine the effects that impact SNM.

##### *Limitations from nanometer MOSFET effects*

In order to ease the discussion, we consider the case of an inverter driving another inverter. The graphical representation of the inverter SNM is shown in Fig. 2.6 at 45 nm node and 0.2V subthreshold  $V_{dd}$ . The SNM is the side of the largest square inscribed between the normal and the mirrored voltage transfer curve.

In this figure, the SNM simulated with typical devices is compared with the SNM simulated first without DIBL effect, by setting BSIM `Eta0` parameter to 0, and second considering variability due to RDF,  $V_t$  being shifted by  $\sigma_{V_t}$  for NMOS and PMOS in opposite ways. From this figure, we see that both DIBL and variability severely degrades the SNM. Notice, that gate leakage can also harm logic gates SNM by worsening the  $I_{on}/I_{off}$  ratio. Nevertheless, in the standard high-performance technologies we consider in this chapter, subthreshold  $I_{off}$  is high and masks gate leakage. Moreover, as gate leakage is less sensitive to variability, it does not significantly impact the SNM.

The impact of variability on SNM is well known and is studied in the case of subthreshold logic in [30], for example. However, to the author's knowledge, the impact of DIBL on SNM has never been investigated up to now. In order to get a qualitative insight of this impact, let us consider an inverter in DC operation under subthreshold  $V_{dd}$ . Let us assume that the input has an ideal high level  $V_{dd}$ . The output is  $V_{OL} \approx 0$ . The pull-down NMOS in on-state has  $V_{gs} = V_{dd}$  and  $V_{ds} \approx 0$ . The pull-up PMOS in off-state has  $|V_{gs}| = 0$  and  $|V_{ds}| \approx V_{dd}$ . The DIBL effect implies an unevenness between  $V_t$  of the on-state device with low  $V_{ds}$  and  $V_t$  of the off-state device with  $V_{ds} \approx V_{dd}$ , similar to a systematic detrimental  $V_t$  mismatch of  $\eta V_{dd}$  between on- and off-state devices. The output voltage  $V_{OL}$  suffers from this and increases somewhat. We also consider variability through a detrimental  $V_t$  mismatch of  $1\sigma$ .  $V_{OL}$  can be found by equating the currents of the pull-down and pull-up devices in Eq. (2.1) with corresponding  $V_{gs}$  and  $V_{ds}$  values:

$$I_{sub,NMOS}|_{on} = I_{sub,PMOS}|_{off}$$

$$I_0 \times 10^{\frac{V_{dd} - \sigma_{V_t}/2}{S}} \times \left( 1 - e^{\frac{-V_{OL}}{U_{th}}} \right) = I_0 \times 10^{\frac{\eta V_{dd} + \sigma_{V_t}/2}{S}}$$

$$1 - e^{\frac{-V_{OL}}{U_{th}}} = 10^{\frac{-((1-\eta)V_{dd} - \sigma_{V_t})}{S}}$$

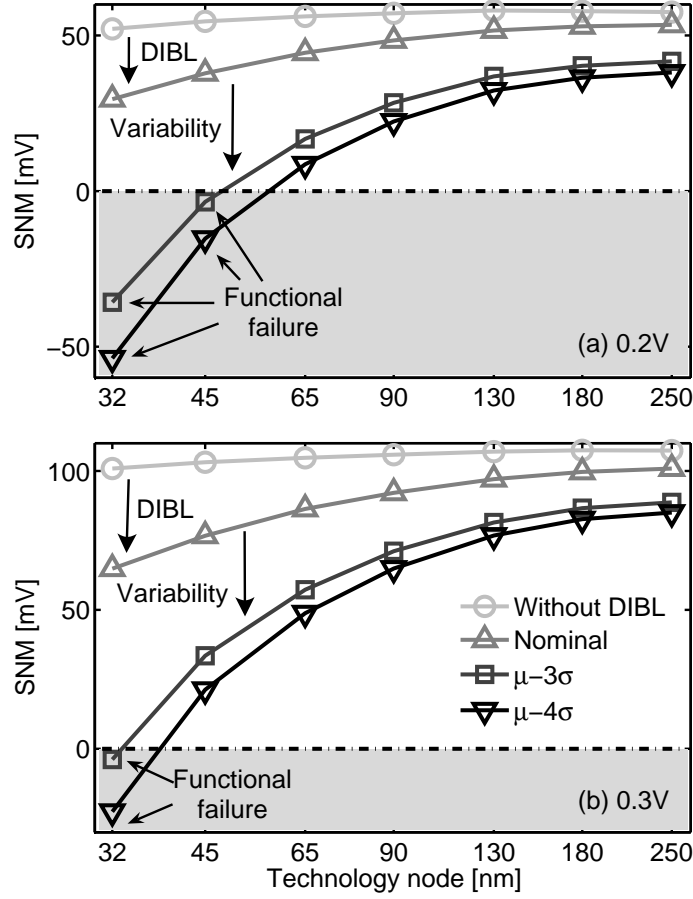
$$V_{OL} = -U_{th} \times \ln \left( 1 - 10^{-F_{SNM}} \right), \quad (2.11)$$

$$\text{with } F_{SNM} = \frac{(1-\eta)V_{dd} - \sigma_{V_t}}{S}. \quad (2.12)$$

The factor  $F_{SNM}$  accounts for the deviation of  $V_{OL}$  from its ideal 0V logic level. For very low  $V_{dd}$  values,  $V_{OL}$  depends on  $F_{SNM}$  as  $10^{-F_{SNM}}$  is no longer negligible as compared to 1. Therefrom, an increase in  $S$  implies a deterioration of  $V_{OL}$  and thus SNM, as suggested in [17]. Moreover, Eq. (2.12) shows that the impact of the DIBL effect is important, through  $\eta$  that further lowers  $F_{SNM}$ . An  $\eta$  value of 200 mV/V has the same impact on the SNM as an increase of  $S$  by 25%.

#### Evolution of static noise margins

As in Section 1.2.1, we use the method from [30] to simulate the SNM of logic gates for the considered technology nodes at subthreshold  $V_{dd}$  values. Simulated SNM is shown in Fig. 2.7(a) at  $V_{dd} = 0.2V$ . Devices are minimum sized (drawn



**Fig. 2.7.** Evolution of the SNM at 0.2V (a) and 0.3V (b) subthreshold  $V_{dd}$ . At 32 nm node, DIBL and variability implies a functional failure (negative SNM) for much more than 0.1% of the gates at 0.2V.

$L_g = L_{min}$ ,  $W_N = 1.5 L_{min}$  and  $W_P = 3 W_N$ ). Although it is shown in [31] that speed-optimum stack sizing in subthreshold logic implies a larger device closest to the supply rail, we consider for the sake of generality a straightforward equal width upsize by a factor  $N$  in stacks made of  $N$  devices.

Without considering the DIBL effect, the SNM decreases by 10% between  $0.13 \mu\text{m}$  and 32 nm nodes because of  $S$  degradation, as observed in [17]. However, the  $S$  degradation has a small impact on SNM as compared to the DIBL effect, which degrades the SNM by up to 45 % between  $0.25 \mu\text{m}$  and 32 nm nodes. Moreover, from 1k-run Monte-Carlo simulations with both RDF and  $L_g$  variations, the impact of variability increases with technology scaling. It is shown in

[32] that the number of logic gates in a subthreshold logic circuit impacts functional yield as larger circuits are more likely to feature logic gates with worst-case SNM. We therefore consider in Fig. 2.7 both the 3 and 4 $\sigma$  worst-case SNM. They decrease dramatically with technology scaling. They even become negative at 45 and 32 nm nodes, leading to an unacceptable failure rate of 15% i.e. a functional yield of 85% under 0.2V at 32 nm node. In order to increase the SNM and thus the functional yield,  $V_{dd}$  has to be raised. As shown in Fig. 2.7(b), raising  $V_{dd}$  to 0.3V allows worst-case SNM at 45 nm node to become positive, thereby ensuring sufficient functional yield. From this, it is clear that technology scaling implies an increasing minimum  $V_{dd}$  to keep a fixed functional yield. This also shows that sub-0.2V functional  $V_{dd}$  values, previously reported for manufactured circuits in [28, 35], are no longer reachable at 45 and 32 nm nodes, unless devices are considerably upsized to mitigate variability.

Notice that SNM is further degraded with process imbalance coming either from different typical NMOS/PMOS  $V_t$  values in subthreshold regime or from cross-corners global variations (fast NMOS/slow PMOS and vice-versa) [33, 34]. Nevertheless, this is highly process-related and varies from a foundry to another [34]. Moreover, it can efficiently be compensated by ABB technique [33] and we therefore do not consider process imbalance in this investigation.

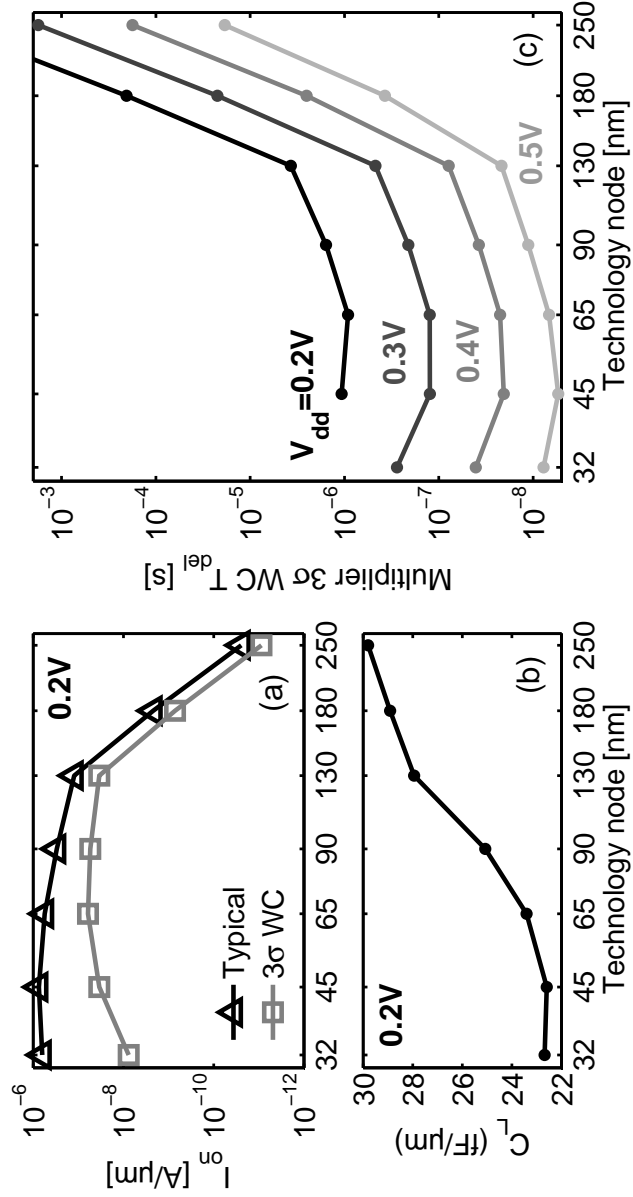
## 2.4.2 Subthreshold circuit delay

Circuit delay  $T_{del}$  is related to the gate delay  $C_L V_{dd} / I_{on}$ , where  $C_L$  is the total subthreshold load capacitance of a logic gate including junction capacitance, intrinsic and extrinsic gate capacitances. The evolution of subthreshold  $I_{on}$  and  $C_L = C_{g,sub} + C_{g,ext} + C_j$  per device width unit is represented in Fig. 2.8(a) at  $V_{dd} = 0.2V$ .  $C_L$  per width unit is reduced by 25% between 0.25  $\mu m$  and 32 nm nodes because of slow  $T_{ox}$  scaling as explained in Section 2.3.3. This reduction is negligible as compared to the increase of  $I_{on}$  per width unit. Therefore, the evolution of the delay mainly depends on the evolution of  $I_{on}$ . From Eq. (2.1) and (2.5), typical subthreshold  $I_{on}$  per width unit is given by:

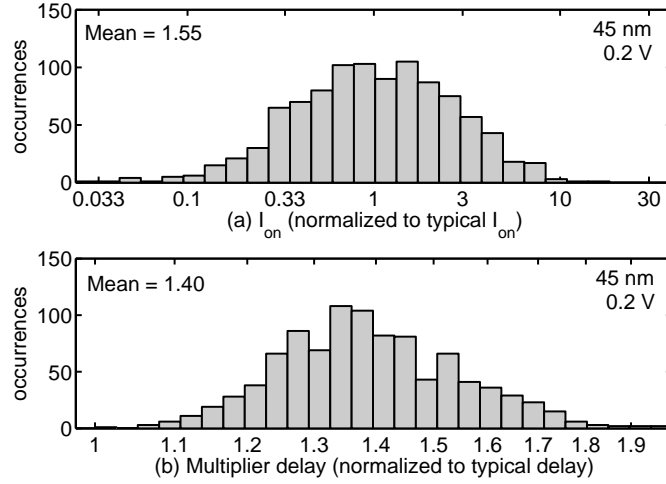
$$\begin{aligned} I_{on} &= I_0 \times 10^{\frac{(1+\eta)V_{dd}}{S}} \\ &= I_{off,nom} \times 10^{\frac{V_{dd} - \eta(V_{dd,nom} - V_{dd})}{S}}, \end{aligned} \quad (2.13)$$

where  $I_{off,nom}$  is the off-state current under nominal superthreshold  $V_{dd,nom}$  from Table 2.2. From 0.25 to 0.13  $\mu m$  node,  $I_{on}$  increases considerably with constant-field scaling as  $I_{off,nom}$  does. However, starting at 90 nm node, the pace slows down and  $I_{on}$  even starts to decrease at 32 nm node. This is due to the limitation of the  $I_{off,nom}$  increase by technology designers for static power issues, whereas  $S$  and  $\eta$  dramatically increase. Notice that the slight  $I_{on}$  decrease at 32 nm node is not significant. Indeed, this is strongly related to the  $I_{off,nom}$  scaling trend, which may differ from a foundry to another, thereby resulting in a slight increase or decrease of typical subthreshold  $I_{on}$ .





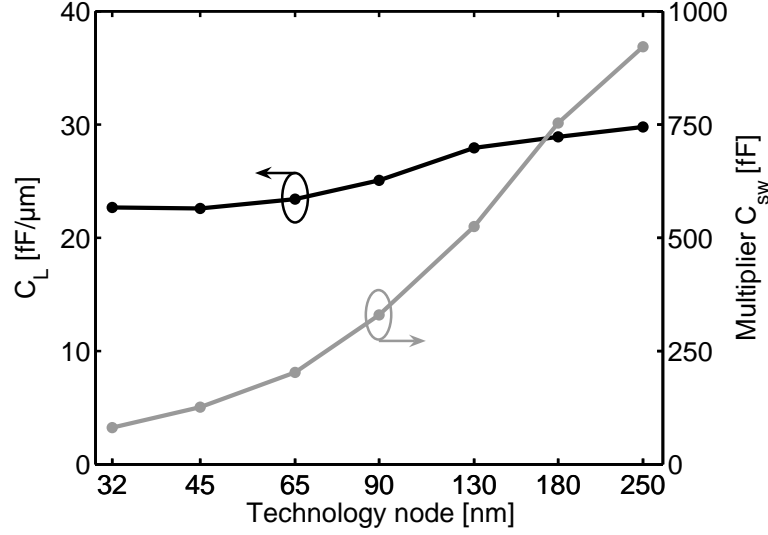
**Fig. 2.8.** Evolution of  $I_{on}$  (a) and total load capacitance  $C_L$  of an FO4 inverter (b) per width unit under 0.2V and multiplier worst-case delay  $T_{del}$  (c) for  $V_{dd}$  ranging from 0.2 to 0.5V (b). At 90 nm node, the delay reduction slows down because  $I_{on}$  increase is slower. At 32 nm node the worst-case delay increases because of high variability and under 0.2V the multiplier does not correctly perform the operation because of bad functional yield.



**Fig. 2.9.** Distribution of NMOS  $I_{on}$  (a) and multiplier delay (b) at 45 nm node under 0.2V subthreshold  $V_{dd}$ .

As shown in Fig. 2.8, variability mainly due to RDF implies a significant reduction of the worst-case  $I_{on}$ . Nevertheless, the impact of  $I_{on}$  variability on circuit delay is reduced through averaging, first inside the gates because of stacked devices, secondly in a critical path because of circuit logic depth and thirdly amongst all the critical paths. This is illustrated in Fig. 2.9 with the distribution of the NMOS  $I_{on}$  and delay of the 8-bit benchmark multiplier at 45 nm node under 0.2V. Mean  $I_{on}$  and mean delay are both higher than the typical one because they are lognormal distributions (exponential dependence on normally distributed  $V_t$ ) but delay variance is considerably lower than  $I_{on}$  variance, thanks to averaging. Notice that averaging only applies to intrinsic  $V_t$  variations such as RDF, whose distributions are spatially uncorrelated. Extrinsic global  $V_t$  variations are not averaged because they are locally correlated. However, they can be compensated by adaptive body biasing, as explained in Section 2.3.4. Again, we thus do not consider them in this investigation.

The evolution of the multiplier  $3\sigma$  worst-case delay, extracted from Monte-Carlo simulations, is shown in Fig. 2.8(b). The multiplier delay is considerably reduced with constant-field scaling from 0.25  $\mu\text{m}$  to 0.13  $\mu\text{m}$  node as  $I_{on}$  increases but the reduction is less important at smallest nodes. At 32 nm node, the delay even increases because of RDF. From the evolution of the delay, it is clear that migrating from 0.25 to 0.13  $\mu\text{m}$  node allows a considerable reduction of  $V_{dd}$  for a fixed throughput, as suggested in [8]. However, migrating to 65 nm node allows only a small further  $V_{dd}$  reduction and migrating to 32 nm node is even detrimental, thereby making 32 nm node a questionable candidate for subthreshold logic.



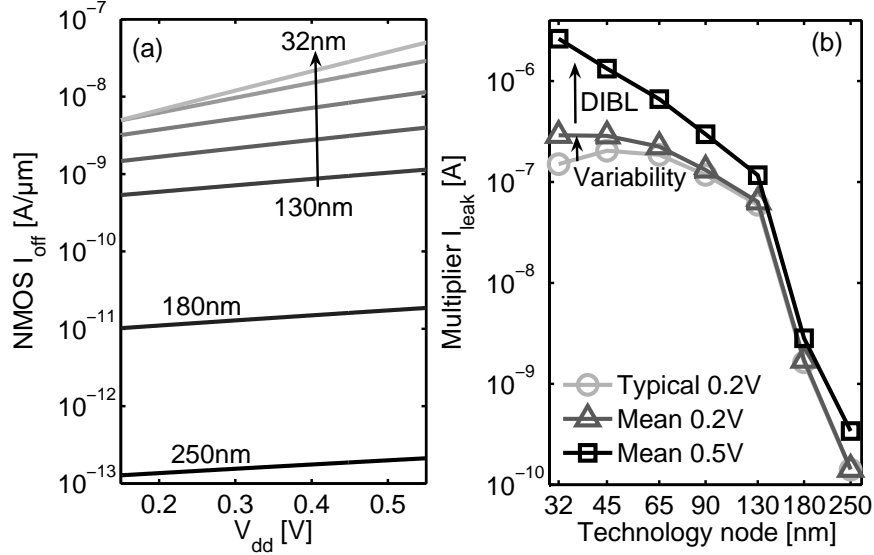
**Fig. 2.10.** Evolution of total  $C_L$  load capacitance of FO4 inverter per width unit and of total switched  $C_{sw}$  capacitance of the multiplier at 0.2V

### 2.4.3 Dynamic power consumption

As reported in Section 1.3.1, dynamic power consumption  $P_{dyn}$  is dominated by capacitance switching. Technology scaling only impacts  $C_L$  parameter of  $P_{dyn}$  expression from Eq. (1.6). Fig. 2.10 shows the evolution of  $C_L$  expressed per device width unit, which exhibits a reduction by 30% between 0.25  $\mu\text{m}$  and 32 nm nodes because of slow  $T_{ox}$  scaling. However, the total switched capacitance  $C_{sw}$  of a circuit is drastically reduced as device minimum width is scaled too. Fig. 2.10 shows that  $C_{sw}$  of the multiplier is reduced by a factor 10 between 0.25  $\mu\text{m}$  and 32 nm nodes, resulting in identical  $P_{dyn}$  reduction at a given  $V_{dd}$  and  $f_{clk}$ .

### 2.4.4 Static power consumption

Static power consumption  $P_{stat}$  depends on the MOSFET off-state currents. Fig. 2.11(a) shows  $I_{off}$  per device width unit vs.  $V_{dd}$  for the whole technology range, which is dominated by subthreshold leakage. The increase of  $I_{off}$  is very important between 0.25 and 0.13  $\mu\text{m}$  nodes as  $V_t$  is scaled linearly with constant-field scaling trend. Notice that at 0.25  $\mu\text{m}$  node,  $I_{leak}$  is dominated by junction leakage. The pace then slows down as  $V_t$  scaling saturates for static-power concern. Interestingly, below 0.2V  $I_{off}$  is not increased between 45 and 32 nm nodes. This suggests that, at nominal  $V_{dd,nom}$ , the  $I_{off,nom}$  increase between these nodes and the underlying  $V_t$  reduction from Table 2.2 mainly come from DIBL effect, which is reduced at low  $V_{dd}$ .

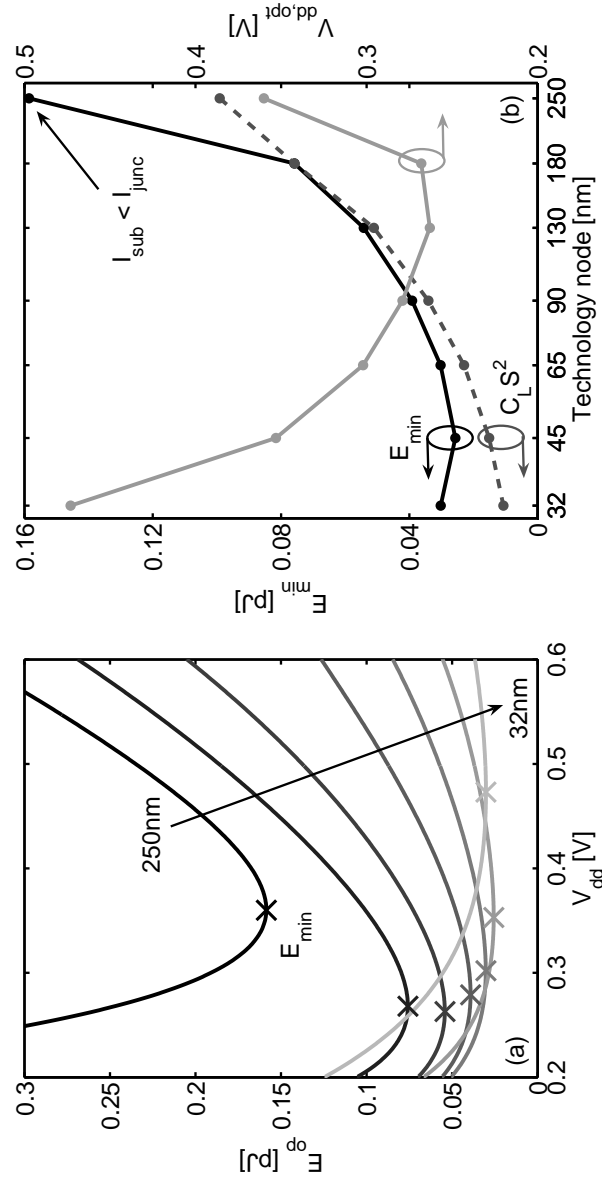


**Fig. 2.11.** Static power consumption (typical) of the multiplier vs.  $V_{dd}$  (a) and its evolution at 0.2V  $V_{dd}$  (b). Variability implies an increase of mean  $P_{stat}$  through subthreshold leakage, which is a lognormal distribution.

Fig. 2.11(b) shows the evolution of total  $I_{leak}$  of the multiplier. The  $I_{leak}$  increase is slightly mitigated as compared to the  $I_{off}$  increase per width unit by the width scaling. At 0.2V, typical  $I_{off}$  even decreases between 45 and 32 nm nodes. Again, this slight decrease is not significant as it depends on the  $I_{off,nom}$  scaling trend and may vary from one foundry to another. As shown in Fig. 2.11(b), variability worsens the picture. Indeed, subthreshold current is a lognormal distribution due to its exponential dependence on normally-distributed  $V_t$ . The mean subthreshold current and thus  $I_{leak}$  is higher than typical values. Moreover at 0.5V, DIBL effect is more important and makes  $I_{leak}$  significantly increase for the whole technology range.

#### 2.4.5 Minimum-energy point

As explained in Section 1.3.3, minimum-energy point results from balancing dynamic and static energies per operation at a particular  $V_{dd}$  and clock frequency. The impact of technology scaling on minimum-energy point is investigated in [17] from 90 to 32 nm node, by considering an ITRS-recommended Low-Standby-Power (LSTP) technology trend [1] and neglecting variability. Hanson *et al.* show that minimum energy level  $E_{min}$  is proportional to  $C_L S^2$  and scales monotonically for the considered technology range. They also show that the corresponding optimum supply voltage  $V_{dd,opt}$  increases as  $S$  does.



**Fig. 2.12.** Energy per operation  $E_{op}$  of the multiplier vs.  $V_{dd}$  (a) and evolution of the minimum-energy level  $E_{min}$  with corresponding  $V_{dd,opt}$  (b). At 32 nm node, variability implies an increased  $E_{min}$ .

We carried out simulations of the benchmark multiplier considering variability, with the high-performance PTM device models from 0.25  $\mu\text{m}$  to 32 nm node. As shown in Fig. 2.12, the trend for  $V_{dd,opt}$  is confirmed from 0.18  $\mu\text{m}$  node. The 0.25  $\mu\text{m}$  technology does not fit the trend because in the considered device model, subthreshold leakage is so low at this node that junction leakage dominates. As  $V_{dd,opt}$  increases, the optimum clock frequency  $f_{clk,opt}$  at minimum-energy point increases from 2.3 kHz in 0.25  $\mu\text{m}$  technology to 1.6 MHz in 32 nm technology to compensate the increased  $I_{leak}$  by reducing the time over which it is integrated. This is particularly interesting for general-purpose low-power applications with DFVS scheme or highly-parallelized architectures, where speed performances do matter unlike in pure ULP applications.

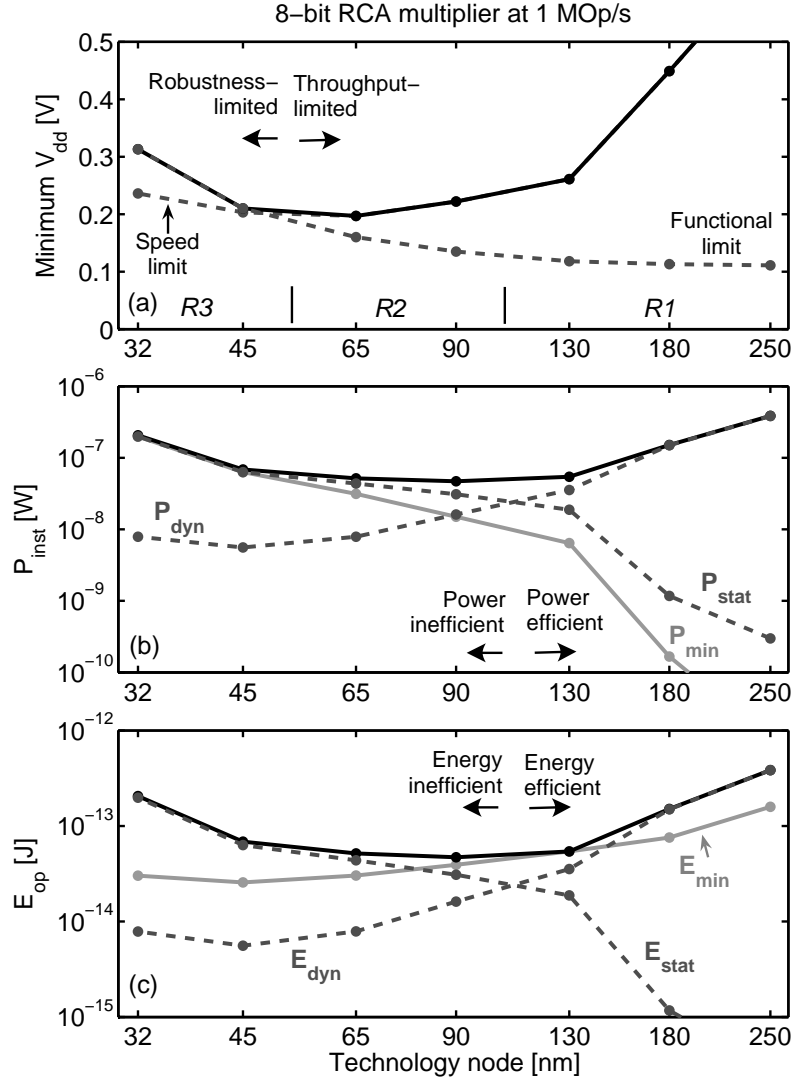
However, the  $E_{min}$  trend from [17] is only verified before 90 nm node. Starting at 65 nm node, the  $E_{min}$  reduction saturates and even increases at 32 nm node, whereas the  $C_L S^2$  decreases for the whole technology range. It shows that some effects of nanometer MOSFETs have a new impact on  $E_{min}$ . These new impacts are quite complex and will be fully investigated in Chapter 3.

#### 2.4.6 Practical power and energy under robustness and throughput constraints

Let us examine the impact of robustness and throughput constraints on practical power/energy consumption. In Fig. 2.13(a), the  $V_{dd}$  functional limit for fixed 99.9% functional yield (positive  $3\sigma$  worst-case SNM) is represented. As shown in Section 2.4.1, it increases dramatically with technology scaling because of increasing  $S$ , DIBL and variability. Minimum  $V_{dd}$  for 1 MOp/s throughput is plotted too. From this figure, we see that the assumption from [8] of lowering  $V_{dd}$  to achieve a fixed throughput is clearly limited by functional-yield degradation because minimum  $V_{dd}$  jumps from throughput to robustness limitation. For the considered benchmark circuit, the jump occurs between 65 and 45 nm nodes.

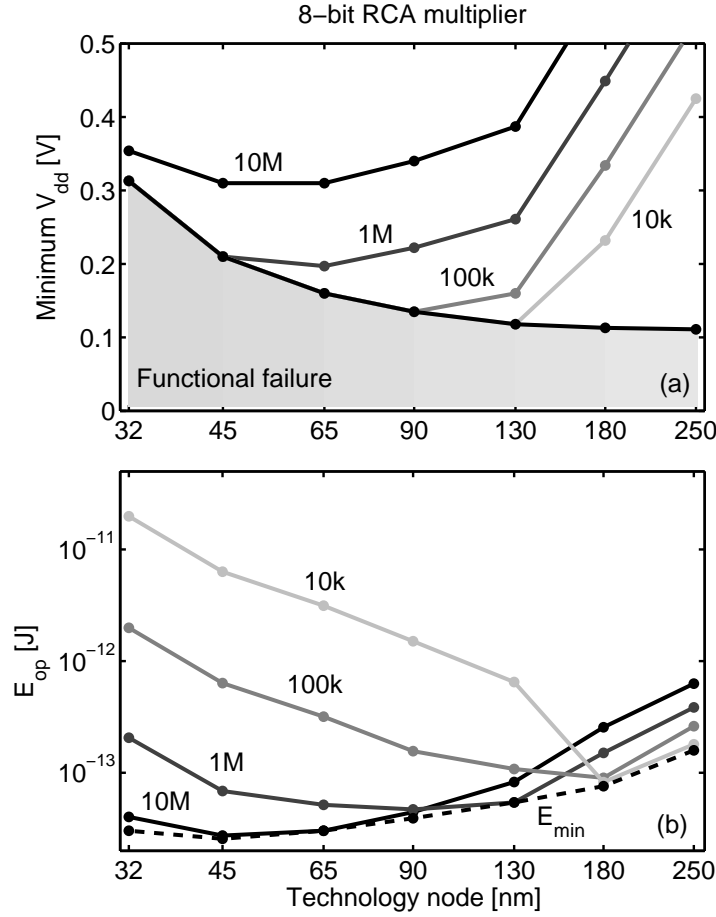
Instantaneous power is plotted in 2.13(b) with dynamic  $P_{dyn}$  and static  $P_{stat}$  components.  $P_{dyn}$  decreases with scaling thanks to the reduction of total switched capacitance  $C_{sw}$ , according to Section 2.4.3, and minimum  $V_{dd}$  lowering. However, when reaching  $V_{dd}$  functional limit,  $P_{dyn}$  starts to increase. According to Section 2.4.4,  $P_{stat}$  dramatically increases with constant-field scaling as  $I_{leak}$  does and then saturates. Consequently,  $P_{inst}$  is first reduced with technology scaling and then increases when  $P_{stat}$  dominates. It is ultimately limited by the increasing  $P_{min}$  level, i.e. static power at  $V_{dd}$  functional limit as defined in Section 1.4.1.

It shows that a circuit that was in power-efficient  $R1$  region in old technology nodes jumps to power-inefficient  $R2$  and ultimately to robustness-limited  $R3$  regions in scaled technologies. Similar observations are drawn from 2.13(c) for energy per operation  $E_{op}$ . It is first improved and tends to minimum energy level  $E_{min}$  as the optimum clock frequency  $f_{clk,opt}$  at minimum-energy point increases and gets closer to the considered 1 MOp/s application throughput. It then increases far above  $E_{min}$  as the circuit enters  $R2$  and  $R3$  regions and



**Fig. 2.13.** Evolution of minimum  $V_{dd}$  (a) under robustness and 1 MOp/s throughput constraints with corresponding instantaneous power  $P_{inst}$  (b) and energy per operation  $E_{op}$  (c) (Spice simulation of a benchmark multiplier). Minimum power  $P_{min}$  and energy  $E_{min}$  levels, i.e. without throughput constraints, are represented too.

$f_{clk,opt}$  gets higher than application throughput. As a consequence, there is an optimum node that minimizes  $E_{op}$  under robustness and throughput constraints as it was recently confirmed by Seok *et al.* in [36]. Let us investigate the impact of application throughput, circuit parameters and temperature on the optimum node.



**Fig. 2.14.** Evolution of minimum  $V_{dd}$  (a) under robustness and various throughput constraints with corresponding energy per operation  $E_{op}$  (b) (Spice simulation of a benchmark multiplier). Minimum energy level  $E_{min}$ , i.e. without target throughput constraint, is represented too (dashed line).

#### Impact of the application throughput

As  $P_{inst}$  and  $E_{op}$  follow the same evolution, we restrict the discussion to  $E_{op}$  and examine the impact of having a different application throughput. Fig. 2.14 shows minimum  $V_{dd}$  and corresponding  $E_{op}$  with throughputs from 10 k to 10 MOp/s. It shows that the jump from throughput to robustness limitation is especially important at low-throughput constraints. Regarding energy per operation, a lower (resp. higher) throughput shifts the optimum node for minimizing practical  $E_{op}$  to an earlier (resp. later) node.



**Table 2.3.** Impact of circuit/application parameters and temperature on optimum node for 8-bit RCA multiplier with 1 MOp/s throughput

	$R1/R2$ trans.	$R2/R3$ trans.	Opt. node	$E_{op}$ [fJ]	Min. $V_{dd}$ [V]	WC SNM [mV]
Baseline	130/90	65/45	90	47.0	0.22	37.6
$\alpha_F/10$	180/130	65/45	180	16.1	0.45	155
Duty cycle = 0.1	180/130	65/45	180	161	0.45	155
$6\sigma$ robustness	130/90	90/65	90	47.0	0.22	19.2
$T = 85^\circ\text{C}$	180/130	90/65	130	108	0.23	40.9

*Impact of circuit/application parameters*

Similarly, let us examine the impact of circuit/application parameters on the optimum node. Table 2.3 shows the nodes of both transitions from energy-efficient  $R1$  to energy-inefficient  $R2$  and from throughput-limited  $R2$  to robustness-limited  $R3$  regions, as well as the optimum node with 1 MOp/s application throughput. It also shows  $E_{op}$ , minimum  $V_{dd}$  and worst-case SNM at the optimum node. An activity factor or a duty cycle reduction shifts the  $R1/R2$  transition by one generation and consequently the optimum node from 90 nm to 180 nm. As minimum  $V_{dd}$  at 180 nm node is higher, worst-case SNM is improved. As large circuits require more than 99.9% functional yield as detailed in Section 2.4.1, in this case the robustness constraint is tighter. Imposing a  $6\sigma$  SNM robustness shifts the  $R2/R3$  transition by one generation and the worst-case SNM at optimum 90 nm node is reduced.

*Impact of the temperature*

A temperature increase shifts the both  $R1/R2$  and  $R2/R3$  transitions by one generation.  $R2/R3$  transition occurs earlier because the temperature-induced subthreshold swing degradation decreases functional yield and the  $V_t$  reduction lowers subthreshold delay through  $I_0$  increase, according to Section 1.4.4.  $R1/R2$  transition occurs earlier because temperature increases  $E_{stat}$ . The optimum node becomes 130 nm and  $E_{op}$  is increased.

From this discussion, we conclude that technology scaling until 65/45 nm nodes is desirable for medium-throughput applications because it yields an energy reduction by more than one order of magnitude from  $E_{dyn}$  reduction. However, the use of sub-100nm nodes for low-throughput applications suffers from dramatically-low energy efficiency due to increasing minimum  $V_{dd}$  to achieve functional yield and high  $E_{stat}$ . Moreover, a higher environment temperature, a low activity factor or a low duty cycle worsens the picture and make 45/32 nm nodes very questionable candidates for subthreshold logic in both low- and medium-throughput ULP applications.

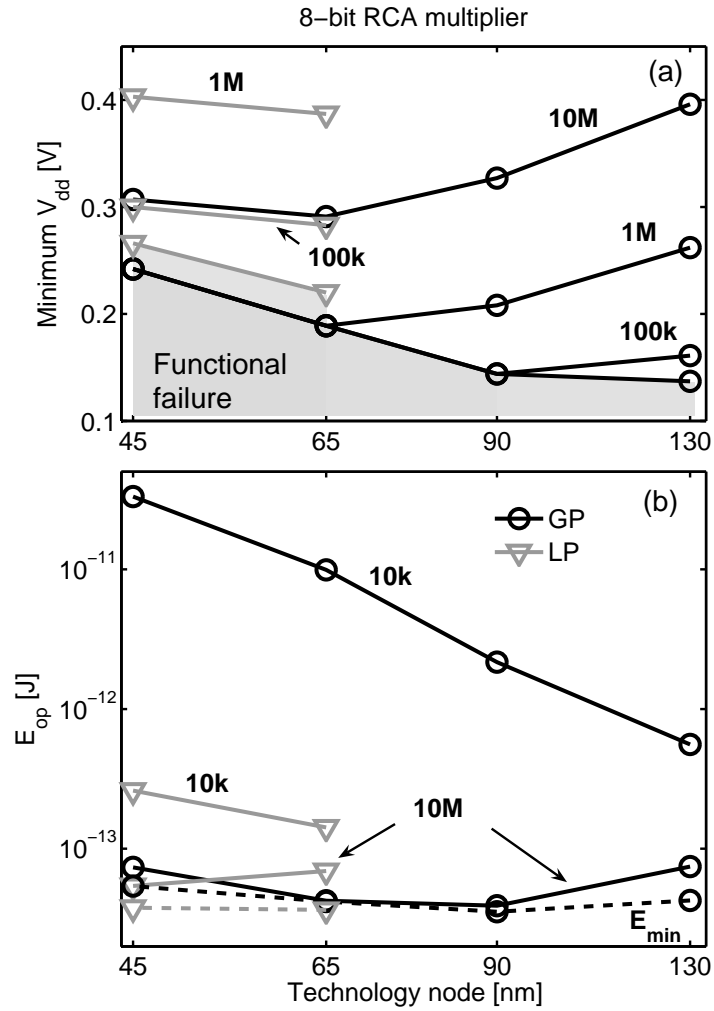
## 2.5 RESULTS VALIDATION

In order to validate these results from predictive models, we simulate the same benchmark multiplier with production models of General-Purpose (GP) industrial technologies from  $0.13\mu\text{m}$  to 45 nm node. Intrinsic random variability (RDF) is considered through Monte-Carlo extraction of  $3\sigma$  worst-case delay and mean  $I_{stat}$ . Again, process-induced global  $V_t$  variations are assumed to be compensated by adaptive body biasing, as explained in Section 2.3.4. First, Fig. 2.15(a) shows that the jump from throughput to functional-yield limitation for minimum  $V_{dd}$  is verified. Corresponding practical  $E_{op}$  is shown in Fig. 2.15(b) for 10 k and 10 MOp/s. The industrial GP trend at both medium and low application throughputs is similar to the one obtained with predictive models, yet the increase of static energy is somewhat faster. The reason is that this foundry introduced at 65 nm node a special flavor of their technologies dedicated to low-power (LP), thereby relaxing the leakage constraints on their GP technologies. This results for the GP trend in a faster increase of practical  $E_{op}$  at low throughputs. At medium throughputs, practical  $E_{op}$  is minimized at 90/65 nm instead of 65/45 nm with PTM models because the GP trend features a faster subthreshold swing degradation as well as a more pronounced fringing capacitance increase.

Let us consider the low-power trend for 65/45 nm nodes in Fig. 2.15. This trend features longer printed gate length  $L_g$ , thicker  $T_{ox}$  and higher channel doping to achieve lower leakage currents. This implies higher RDF confirmed by Eq. (2.10). However, DIBL is mitigated ( $\eta_{LP} = 110\text{mV/V} < \eta_{GP} = 160\text{mV/V}$  at 45 nm node) and subthreshold reference current  $I_0$  is reduced by 2 orders of magnitude. This results in a higher minimum  $V_{dd}$  to achieve both functional yield and required throughput. The corresponding practical  $E_{op}$  under robustness and throughput constraints is reduced at low throughputs. At medium throughputs, energy for LP technologies is dominated by  $E_{dyn}$  and migrating from 65 to 45 nm LP nodes thus lowers  $E_{op}$ . At 45 nm node, a low-power technology is thus an interesting option for subthreshold logic for both low- and medium-throughput applications. Nevertheless, the increase of both practical  $E_{op}$  at 10 kOp/s and  $E_{min}$  between 65 and 45 nm LP nodes suggests that LP technologies follow the same evolution than the GP ones but delayed by one or two generations. At 32 nm node, LP technology is thus expected to suffer from the same energy increase than GP technologies.

## 2.6 CONCLUSION

In this chapter, we investigated the impact of technology scaling on subthreshold logic from  $0.25\mu\text{m}$  to 32 nm nodes. Through extensive device modeling, we showed that subthreshold  $I_{on}$ , which was exponentially increasing with constant-field scaling until 90 nm node, saturates in nanometer technologies due to degraded subthreshold swing and high DIBL effect. This effect is worsened by variability increase, leading to reduced worst-case  $I_{on}$  at 45 and 32 nm nodes. We



**Fig. 2.15.** Validation with industrial models for general-purpose (GP) and low-power (LP) technology trends: (a) minimum  $V_{dd}$  under robustness and various throughput constraints with (b) corresponding energy per operation  $E_{op}$  (Spice simulation of a benchmark multiplier, dashed lines represent minimum energy level  $E_{min}$ ). GP trend results are similar to Fig. 2.14. Thanks to important  $E_{stat}$  reduction, the LP trend extends the benefit of technology scaling to 45 nm node. However, the increasing  $E_{min}$  and  $E_{op}$  at 10 kOp/s suggest that LP technologies will also suffer from an energy increase at 32 nm node.

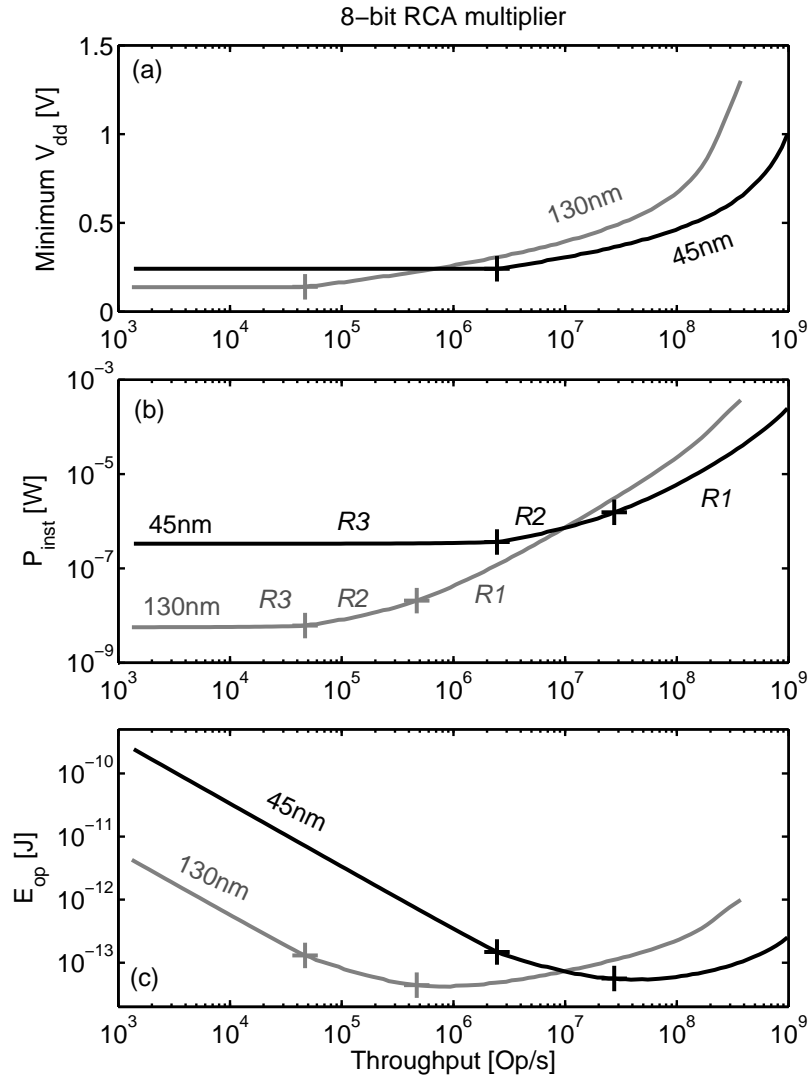
also show that fringing capacitances dominate in subthreshold regime and that, amongst them, capacitances between gate and source/drain contacts exhibit a worrying increase due to aggressive scaling of gate-to-source/drain spacing.

At circuit level, we showed that technology scaling dramatically degrades static noise margins of subthreshold logic gates, due to high subthreshold swing, DIBL and variability. This implies an increasing minimum  $V_{dd}$  to meet robustness constraint. Thorough Monte-Carlo simulations of a benchmark multiplier under subthreshold supply voltages showed that subthreshold delay is improved until 90 nm node and then saturates as a direct consequence of the worst-case subthreshold  $I_{on}$  beyond this node. As a result of these facts, minimum  $V_{dd}$  jumps from throughput to robustness limitation when scaling to nanometer technologies and the minimum-power range is extended to higher throughputs.

Technology scaling has an important impact on minimum-energy point. It shifts minimum-energy point to higher optimum supply voltage and clock frequency. This makes minimum-energy operation more and more feasible for DFVS circuits and highly-parallelized architectures in consumer low-power/wireless applications. Nevertheless, starting at 65 nm node, minimum-energy level increases, suggesting that nanometer MOSFET effects put new limitations on minimum energy. In Chapter 3, we study these new limitations and investigate both technology option and optimum MOSFET selection to fix this increase of minimum-energy level.

Regarding energy per operation under robustness and throughput constraints, migrating from 0.25  $\mu\text{m}$  to 90/65 nm nodes provides an important interest for medium-throughput applications ( $\approx 1\text{-}10$  MOp/s) thanks to dynamic energy reduction by more than one order of magnitude. However, at 45/32 nm nodes this benefit is limited by high static energy due to degraded subthreshold swing and high delay variability. Furthermore, low-throughput applications ( $\approx 10\text{-}100$  kOp/s) suffer from an extra static energy overhead. This is due to high leakage currents and high minimum  $V_{dd}$  to achieve sufficient functional yield, coming from degraded subthreshold swing, high DIBL and variability.

According to the analysis framework proposed in Chapter 1, a subthreshold circuit, which lies in  $R1$  energy-efficient submicron technology at a given application throughput will be shifted to energy-inefficient  $R2$  and ultimately to robustness-limited  $R3$  regions, when migrating to nanometer technologies. As illustrated in Fig. 2.16, direct porting of an FVS circuit from 0.13  $\mu\text{m}$  to 45 nm technology only benefits applications with throughputs higher than 10 MOp/s. For ULP application throughputs, subthreshold circuit designers face a new paradigm in 45 nm technology, as circuits lie in  $R3$  region at the limit of robustness with high static power/energy component. In Chapter 4, we revisit the design choices for minimizing practical energy per operation under robustness and throughput constraints in this light.



**Fig. 2.16.** Comparison of 0.13  $\mu\text{m}$  and 45 nm technologies: (a) minimum  $V_{dd}$  under robustness and throughput constraints with (b) corresponding instantaneous power and (c) energy per operation (Spice simulation of the 8-bit RCA benchmark multiplier in industrial standard bulk technologies). For the throughput range of ULP applications ( $\approx 10$  k-10 MOp/s), circuits in 45 nm technology mainly lie in robustness-limited energy-inefficient R3 throughput region.

## REFERENCES

1. Semiconductor Industry Association, "Executive summary", in *The International Technology Roadmap for Semiconductors 1999-2007 Editions*, Tech. Rep., Semiconductor Industry Association, 1999-2007.
2. H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, R. Krishnamurthy and S. Borkar, "A 320mV 65 $\mu$ W 411GOPS/Watt ultra-low voltage motion estimator accelerator in 65 nm CMOS", in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 316-317, 2008.
3. J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann and A. Chandrakasan, "A 65 nm sub- $V_t$  microcontroller with integrated SRAM and switched-capacitor DC-DC converter", in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 318-319, 2008.
4. B. Zhai, D. Blaauw, D. Sylvester and K. Flautner, "The limit of dynamic voltage scaling and insomnia dynamic voltage scaling", in *IEEE Trans. VLSI Syst.*, vol. 13, no. 11, pp. 1239-1252, Nov. 2005.
5. B. H. Calhoun and A. P. Chandrakasan, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering", in *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 238-245, Jan. 2006.
6. B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge and D. Sylvester, "Energy efficient near-threshold chip multi-processing", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 32-37, 2007.
7. V. Sze and A. P. Chandrakasan, "A 0.4-V UWB baseband processor", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 262-267, 2007.
8. A. Raychowdhury, B. C. Paul and K. Roy, "Computing with subthreshold leakage: device/circuit/architecture co-design for ultralow-power subthreshold operation", in *IEEE Trans. VLSI Syst.*, vol. 13, no. 11, pp. 1213-1224, Feb. 2005.
9. B. Zhai, S. Hanson, D. Blaauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 20-25, 2005.
10. B. H. Calhoun, A. Wang and A. P. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits", in *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778-1786, Sep. 2005.
11. W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration", in *IEEE Trans. Electron Dev.*, vol. 53, no. 11, pp. 2816-2823, Nov. 2006.
12. R. Dennard, F. Gaensslen, V. Rideout, E. Bassous and A. Leblanc, "Design of ion-implanted MOSFET's with very small physical dimensions", in *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256-268, Oct. 1974.
13. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur and H.-S. Wong, "Device scaling limits of Si MOSFETs and their application dependencies", in *Proc. IEEE*, vol. 89, no. 3, pp. 259-288, Mar. 2001.
14. G. Baccarani, M. Wordeman and R. Dennard, "Generalized scaling theory and its application to a 1/4 micron MOSFET design", in *IEEE Trans. Electron Dev.*, vol. 31, no. 4, pp. 452-462, Apr. 1984.

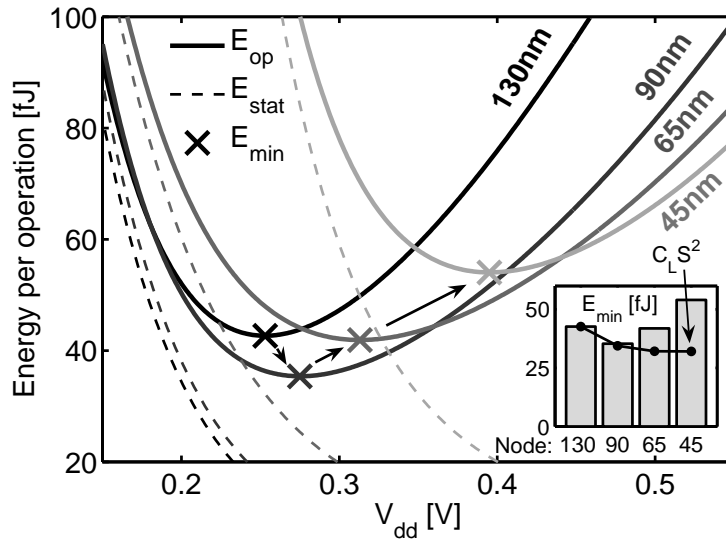
15. Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. K. Ko and Y. C. Cheng, "Threshold voltage model for deep-submicrometer MOSFET's", in *IEEE Trans. Electron Dev.*, vol. 40, no. 1, pp. 86-95, Jan. 1993.
16. R. Gwoziecki and T. Skotnicki, "Physics of the subthreshold slope - initial improvement and final degradation in short CMOS devices", in *Proc. European Solid-State Dev. Research Conf.*, pp. 639-642, 2002.
17. S. Hanson, M. Seok, D. Sylvester and D. Blaauw, "Nanometer device scaling in subthreshold logic and SRAM", in *IEEE Trans. Electron Dev.*, vol. 55, no. 1, pp. 175-185, Jan. 2008.
18. M. V. Dunga *et al.*, "BSIM4.6.1 MOSFET model", available on-line at [www-device.eecs.berkeley.edu/bsim3/bsim4.html](http://www-device.eecs.berkeley.edu/bsim3/bsim4.html).
19. H. Ohta *et al.*, "High performance 30 nm gate bulk CMOS for 45 nm node with  $\Sigma$ -shaped SiGe-SD", in *Proc. IEEE Int. Electron Dev. Meeting*, 2005.
20. K. Goto *et al.*, "High performance 25 nm gate CMOSFETs for 65 nm node high speed MPUs", in *Proc. IEEE Int. Electron Dev. Meeting*, 2003.
21. Q. Huang, F. Piazza, P. Orsatti and T. Ohguro, "The impact of scaling down to deep submicron on CMOS RF circuits", in *IEEE J. Solid-State Circuits*, vol. 33, no. 7, pp. 1023-1036, July. 1998.
22. B. C. Paul, A. Raychowdhury and K. Roy, "Device optimization for digital subthreshold logic operation", in *IEEE Trans. Electron Dev.*, vol. 52, no. 2., pp. 237-247, Feb. 2005.
23. N. R. Mohapatra, M. P. Desai, S. G. Narendra and V. Ramgopal Rao, "Modeling of parasitic capacitances in deep submicrometer conventional and high- $\kappa$  dielectric MOS transistors", in *IEEE Trans. Electron Dev.*, vol. 50, no. 4, pp. 959-966, Apr. 2003.
24. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decanometer and nanometer-scale MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.
25. D. Levacq, T. Minakawa, M. Takamiya and T. Sakurai, "A wide range spatial frequency analysis of intra-die variations with 4-mm x 1 transistor arrays in 90nm CMOS", in *Proc. IEEE Int. Custom Integrated Circuits Conf.*, pp. 257-260, 2007.
26. G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs", in *Solid-State Electronics*, vol. 44, no. 6, pp. 1105-1109, Jun. 2006.
27. J. T. Kao, M. Masayuki and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545-1554, Nov. 2002.
28. S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singha, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester and D. Blaauw, "Exploring variability and performance in a sub-200-mV processor", in *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 881-891, Apr. 2008.
29. Y. Cao and L. T. Seok, "Mapping statistical process variations toward circuit performance variability: an analytical modeling approach", in *Proc. ACM/IEEE Des. Autom. Conf.*, pp. 658-663, 2005.

30. J. Kwong and A. P. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 8-13, 2006.
31. J. Keane, H. Eom, T.-H. Kim, S. Sapatnekar and C. H. Kim, "Stack sizing for optimal current drivability in subthreshold circuits", in *IEEE Trans. VLSI Syst.*, vol. 16, no. 5, pp. 598-602, May 2008.
32. T. Niiyama, Z. Piao, K. Ishida, M. Murakata, M. Takamiya and T. Sakurai, "Increasing minimum operating voltage  $V_{DDmin}$  with number of cmos logic gates and experimental verification with up to 1 Mega-stage ting oscillators", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 117-122, 2008.
33. Y. Pu, J. Pineda de Gyvez, H. Corporaal and Y. Ha, " $V_t$  balancing and device sizing towards high yield of sub-threshold static logic gates", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 355-358, 2007.
34. J. F. Ryan, J. Wang and B. H. Calhoun, "Analyzing and modeling process balance for sub-threshold circuit design", in *Proc. Great Lakes Symp. VLSI*, pp. 275-280, 2007.
35. A. Wang and A. P. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology", in *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310-319, Jan. 2005.
36. M. Seok, D. Sylvester and D. Blaauw, "Optimal technology selection for minimizing energy and variability in low voltage applications", in *IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 9-14, 2008.



## CHAPTER 3

# OPTIMUM NANOMETER CMOS DEVICES AND TECHNOLOGY FOR MINIMUM-ENERGY SUBTHRESHOLD CIRCUITS



**Fig. 3.1.** Total  $E_{op}$  and static  $E_{stat}$  energies per operation with minimum-energy point  $E_{min}$  in recent general-purpose technologies from an industrial foundry (variability-aware Spice simulation results of a benchmark 8-bit multiplier from Chapter 2). The insert shows that  $E_{min}$  deviates from its  $C_L S^2$  figure of merit reported in [1]. *Let us see why.*

## Abstract

---

As shown in Chapter 2, minimum energy in subthreshold circuits increases from 90 nm node, whereas its previously-reported  $C_L S^2$  figure of merit decreases. In this chapter, we first explain the new effects that make minimum energy rise in nanometer technology: DIBL, gate leakage and device variability. We then study the impact of nanometer MOSFET parameters on minimum energy. We show that traditional technology flavors are not adapted to minimum-energy subthreshold circuits and we propose an optimum device selection to improve energy efficiency, at circuit level, i.e. without any process modification. At 45 nm node, we show that the use of thin-oxide low- $V_t$  devices in a high-performance technology flavor with gate length upsized by 15 to 25 nm reduces minimum-energy level by 35-40%, with mitigation of delay variability as an extra benefit. This study draws a new route for device optimization towards ultimate subthreshold circuits, indicating that efforts should be devoted to minimizing subthreshold swing, DIBL and variability, while gate leakage increase can be tolerated provided that it remains below the subthreshold leakage level.

Finally, we investigate the potential of ultra-thin-body fully-depleted (FD) Silicon-on-insulator (SOI) technology to reduce minimum energy [CP5]. In standard 45 nm high-performance technology, FD SOI brings 45% minimum-energy reduction at minimum gate length, thanks to subthreshold swing improvement, capacitance reduction and variability mitigation. The combination of an undoped channel with a metal gate further increases this improvement, yielding a 60% minimum-energy reduction as compared to bulk, thanks to outstanding variability mitigation.

## Contents

---

<b>3.1</b>	<b>Introduction</b>	<b>63</b>
<b>3.2</b>	<b>Background and related work</b>	<b>64</b>
<b>3.3</b>	<b>Pre-Silicon bulk MOSFET compact models for sub-threshold circuit simulation</b>	<b>66</b>
<b>3.4</b>	<b>Limitations from nanometer MOSFET effects</b>	<b>69</b>
<b>3.5</b>	<b>Impact of nanometer MOSFET parameters</b>	<b>71</b>
<b>3.6</b>	<b>Optimum technology and device selection</b>	<b>74</b>
<b>3.7</b>	<b>Fully-depleted SOI technology</b>	<b>80</b>
<b>3.8</b>	<b>Conclusion</b>	<b>88</b>

---

### 3.1 INTRODUCTION

As first reported in [2, 3], we showed in Chapter 1 that energy per operation without robustness nor throughput constraints can be minimized by operating at a particular supply voltage  $V_{dd,opt}$  and clock frequency  $f_{clk,opt}$ , i.e. the so-called “minimum-energy point”, which often lies in MOSFET subthreshold region. The corresponding minimum-energy level  $E_{min}$  results from a balance between dynamic and static energy components. When lowering  $V_{dd}$ , dynamic component is quadratically reduced whereas static component exponentially increases in the subthreshold region as the delay does and thus the execution time of the operation, over which leakage currents are integrated. Minimum energy per operation is achieved at the expense of speed performances, which limits it to pure ULP applications, where speed requirements are not stringent.

We showed in Chapter 2 that technology scaling increases  $V_{dd,opt}$  and  $f_{clk,opt}$ , which makes operation at minimum-energy point feasible in the low MHz-frequency range, thereby extending the application spectrum. Indeed, both DFVS general-purpose microprocessors [4, 5] and high-performance multi-processing architectures [6, 7, 8] operating at minimum-energy point have recently been proposed for general low-power/wireless applications.

Regarding minimum-energy level, technology scaling yields interesting dynamic energy reduction through capacitance reduction at the expense of increased leakage currents, short-channel effects and device variability. In [1], Hanson *et al.* shows that minimum-energy level  $E_{min}$  is reduced when migrating from 90 nm to 32 nm node, when considering a Low-Stanby Power (LSTP) technology trend and neglecting device variability. They also show that  $E_{min}$  is proportional to the switched capacitance multiplied by the square of the subthreshold swing  $C_L S^2$ , a factor that decreases with technology scaling. Nevertheless, when considering an 8-bit benchmark multiplier in high-performance (HP) CMOS technologies, we showed in Chapter 2 that  $E_{min}$  reduction saturates at 65 nm node and even increases at 32 nm node. In order to validate this observation, we carry out variability-aware statistical Monte-Carlo Spice simulations of the same benchmark multiplier with production MOSFET models from the industrial general-purpose technologies of Section 2.5. Fig. 3.1 illustrates the resulting energy per operation  $E_{op}$  vs.  $V_{dd}$  and its static  $E_{stat}$  component. Although technology scaling reduces the dynamic energy ( $E_{op}$  curves above 0.4V  $V_{dd}$ ) and the  $C_L S^2$  factor as shown in the insert of Fig. 3.1,  $E_{min}$  increases from 90 nm node because of dramatical static-energy increase. It shows that new effects in nanometer technologies make  $E_{min}$  deviate from previously-reported  $C_L S^2$  factor. This suggests a major challenge to keep  $E_{min}$  reasonable when migrating to nanometer technologies for meeting area and cost constraints.

The leakage current increase due to technology scaling has also lead to a diversification of the technologies to trade off speed for low power. Multi- $V_t$  option is widespread since 0.13  $\mu\text{m}$  node and starting at 65 nm node, many industrial foundries offer a versatile technology menu with 2 or 3 flavors i.e. with different  $V_t$ , oxide thicknesses  $T_{ox}$  and minimum gate length  $L_g$  to further optimize this

trade off. However, as none of these flavors explicitly targets minimum-energy subthreshold operation, circuit designers face a completely new issue: selecting the optimum technology and devices at a given node amongst all these opportunities. Moreover, new technologies such as fully-depleted (FD) ultra-thin-body (UTB) Silicon-on-insulator (SOI) are predicted to be indispensable beyond 32 nm node to keep short-channel effects and variability under control [9]. It is thus worth evaluating the potential of such an FD SOI technology for mitigating the minimum-energy increase in nanometer technologies.

In this chapter, we therefore first analyze the effects that make  $E_{min}$  rise in nanometer technologies, as well as the impact of the basic device parameters on  $E_{min}$ . We then address the issue of selecting the best technology at 45 nm node for minimum energy, from a circuit designer point of view i.e. within a practical set of available technologies. We consider the ITRS recommended technology flavors in planar bulk technology: high performance (HP), low operating power (LOP) and low stand-by power (LSTP) [9]. We then investigate the interest of FD SOI technology for minimum-energy subthreshold circuits.

This chapter is organized as follows. Section 3.2 briefly reviews the concept of minimum-energy point as well as previous work on device optimization for subthreshold circuits. In Section 3.3, we propose a pre-Silicon MOSFET compact modeling approach for realistic subthreshold circuit simulation. We then use generated models to show in Sections 3.4 and 3.5 the impact of nanometer MOSFET effects and device parameters on minimum energy. The problem of optimum bulk technology/device selection is addressed in Section 3.6 and finally, the benefits brought by FD SOI technology are investigated in Section 3.7.

## 3.2 BACKGROUND AND RELATED WORK

### 3.2.1 Minimum-energy point modeling

As explained in Chapter 1, the energy per operation of a circuit is the sum of dynamic energy  $E_{dyn}$  due to the capacitance switching and static energy  $E_{stat}$  due to leakage currents  $I_{leak}$  flowing through the devices. Static energy results from the integration of  $I_{leak}$  during the actual execution of the operation i.e. over a time period equal to the delay  $T_{del}$  of the circuit critical path. According to Section 1.3.3, energy is thus expressed as:

$$\begin{aligned} E_{op} &= E_{dyn} + E_{stat} \\ E_{dyn} &= \frac{1}{2} N_{sw} C_L V_{dd}^2 \\ E_{stat} &= V_{dd} I_{leak} T_{del} \end{aligned} \quad (3.1)$$

where  $N_{sw}$  is the number of switched nodes to perform the operation and  $C_L$  the typical node capacitance. Minimum energy is often achieved when operating the circuit in MOSFET subthreshold region [2]. Subthreshold drain current is

expressed as:

$$I_{sub} = I_0 \times 10^{\frac{V_{gs} + \eta V_{ds}}{S}} \times \left( 1 - e^{\frac{-V_{ds}}{U_{th}}} \right) \quad (3.2)$$

In previous works on minimum-energy point [1, 10], leakage currents are assumed to be dominated by subthreshold leakage and static energy is expressed as:

$$\begin{aligned} E_{stat} &= V_{dd} \times I_{leak} \times T_{del} \\ &\propto V_{dd} \times I_{sub,off} \times \frac{L_D C_L V_{dd}}{I_{sub,on}} \\ &\propto V_{dd} \times I_0 10^{\frac{\eta V_{dd}}{S}} \times \frac{L_D C_L V_{dd}}{I_0 10^{\frac{(1+\eta)V_{dd}}{S}}} \\ &\propto L_D C_L 10^{\frac{-V_{dd}}{S}} V_{dd}^2, \end{aligned} \quad (3.3)$$

where  $L_D$  is the logic depth, i.e. the number of logic gates in the critical path.

Under these conditions, when neglecting device variability, minimum energy  $E_{min}$  is shown in [1] to be proportional to  $C_L S^2$ . In [12], variability is shown to worsen  $E_{min}$  because worst-case delay has to be considered in order for all manufactured chips to correctly operate with the same clock frequency under a given  $V_{dd}$ .

### 3.2.2 Device optimization for minimum energy

Several works investigate the interests of new MOSFET architectures for subthreshold circuits such as double-gate SOI [13] and underlap devices [14] or post-Si devices such as super cut-off transistors [15]. Nevertheless, in this chapter we rather focus on planar Si devices as it is today's mainstream for circuit designers. For optimum subthreshold operation of such standard devices, it has been shown that halo doping can be reduced [1],[16] and that a high-to-low vertical channel doping profile is preferable [16]. However, to the authors' knowledge no industrial foundry offers a technology that specifically targets minimum-energy subthreshold logic. Moreover, process modifications such as doping profile optimizations are hardly available to circuit designers. In this chapter, we therefore focus on the 3 main device parameters that circuit designers can usually choose within a set in a versatile yet standard technology menu:  $V_t$ ,  $T_{ox}$  and  $L_g$ .

In [4, 10], it is suggested that  $V_t$  has no impact on  $E_{min}$ , provided that the devices actually remain in subthreshold regime. This is confirmed by  $E_{stat}$  expression from Eq. (3.3) where  $V_t$  impact through  $I_0$  parameter is simplified when multiplying  $I_{leak}$  by the delay. However, if  $V_t$  is too low,  $I_{on}$  is no longer a subthreshold current and Eq. (3.2) does not hold. The delay is thus no longer exponentially-dependent on  $V_{dd}$  and the  $I_{leak} \times T_{del}$  product increases.

In [11], Kim *et al.* propose to upsize the channel length to benefit from reverse-short-channel effects (RSCE) in bulk technology for improving subthreshold operation. Indeed, if devices are dominated by RSCE i.e. positive  $V_t$  roll-off because of high halo doping, a channel length upsize results in an increased subthreshold

$I_{on}$  [11]. However, as RSCE also increases subthreshold  $I_{off}$ , it does not provide energy reduction. On another hand, it is shown in [1] that, independently of positive or negative  $V_t$  roll-off, there is an optimum  $L_g$  to minimize  $E_{min}$ , which results from a trade off between  $S$  improvement from short-channel behavior mitigation and  $C_L$  increase at longer  $L_g$ . Notice that the optimum  $L_g$  is quite long as the impact of intrinsic gate capacitance  $C_g$  on  $C_L$  is small. This comes from the reduced value of  $C_g$  in subthreshold regime as compared to parasitic capacitances because subthreshold  $C_g$  is dominated by the depletion capacitance in the channel, as explained in Chapter 2.

Finally, it is shown in [17] that an optimum  $T_{ox}$  also results from a trade-off between  $S$  improvement from improved channel control and both intrinsic and extrinsic (fringing and overlap) gate capacitance increase, at thinner  $T_{ox}$ . In Section 3.5, we will show that these trade-offs remain valid in nanometer technologies but may be outweighed by DIBL, gate leakage and variability.

### 3.3 PRE-SILICON BULK MOSFET COMPACT MODELS FOR SUBTHRESHOLD CIRCUIT SIMULATION

In this chapter, we investigate circuit-level implications of nanometer MOSFET characteristics by targeting 45 nm CMOS technologies. We consider the 8-bit RCA multiplier as a benchmark circuit and for simulation time issues, we thus use Spice simulations based on BSIM4 bulk MOSFET compact models [18]. In order to investigate the impact of device parameters, we need a generic yet accurate model card to define BSIM4 parameters. In this section, we describe the methodology we use to generate these models as illustrated in Fig. 3.2. Further details on the generated model cards can be found in Appendix C.

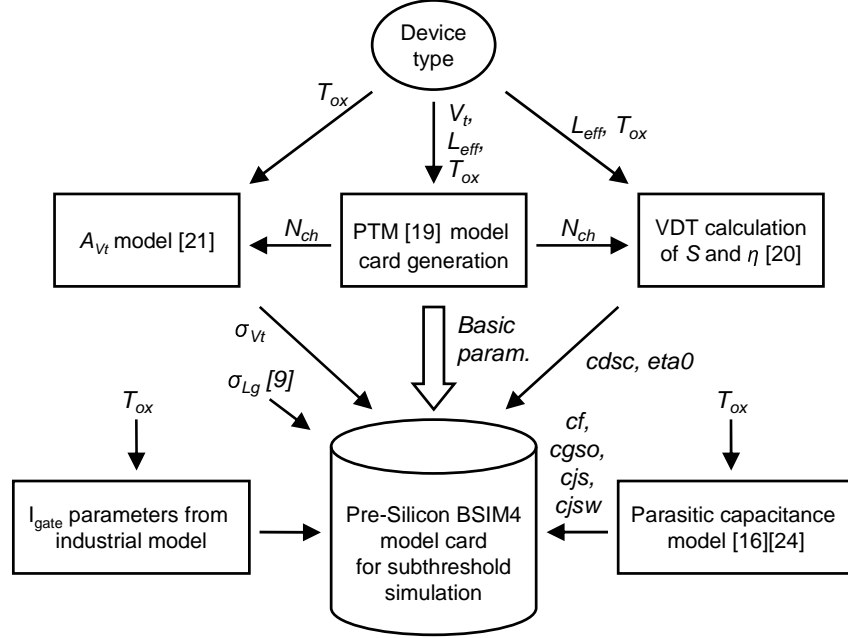
#### *Basic parameters*

First, we use 45 nm Predictive Technology Models<sup>1</sup> (PTM) from Arizona State University [19] as basic BSIM4 model card. PTM can be customized by defining effective channel length  $L_{eff}$ ,  $T_{ox}$  and  $V_t$ . We rely on the generated model card for second-order parameters as well as for the channel doping  $N_{ch}$ , the mobility and the body effect. Nevertheless, PTM model has been calibrated to only model high-performance technologies. We therefore refine several parameters in the generated model card to get a realistic subthreshold behavior for different technology flavors.

#### *Subthreshold current parameters*

We use Skotnicki's voltage-doping transformation (VDT) equations of electrostatic integrity [20] for empirical calculation of the subthreshold swing  $S$  and the DIBL factor  $\eta$ , given the considered  $T_{ox}$ ,  $L_g$  and  $N_{ch}$  values:

<sup>1</sup>Models are available on-line at [www.eas.asu.edu/~ptm](http://www.eas.asu.edu/~ptm).



**Fig. 3.2.** Methodology to generate BSIM4 MOSFET model cards for pre-Silicon subthreshold circuit simulation

$$S = U_t \times \ln(10) \times \left( 1 + \frac{\epsilon_{Si}}{\epsilon_{ox}} \frac{T_{ox,el}}{X_{dep}} + \frac{\epsilon_{Si}}{\epsilon_{ox}} EIS \sqrt{1 + 2 \frac{V_{ds}}{\Phi_s}} \right) \quad (3.4)$$

$$\text{with } EIS = \frac{T_{ox,el}}{L_{g,el}} \frac{X_j}{L_{g,el}} \times \left( 1 + \frac{3}{4} \frac{X_{dep}}{L_{g,el}} \right),$$

$$\eta = 0.8 \times \frac{\epsilon_{Si}}{\epsilon_{ox}} \times EI \quad (3.5)$$

$$\text{with } EI = \frac{T_{ox,el}}{L_{g,el}} \frac{X_{dep}}{L_{g,el}} \times \left( 1 + \frac{X_j^2}{L_{g,el}^2} \right).$$

In these equations, the electrical gate length  $L_{g,el}$  is estimated at  $\frac{2}{3}L_g$  (printed) and  $\Phi_s$  is the surface potential close to 1V for the considered technologies [20]. The electrical  $T_{ox,el}$  is calculated from the physical  $T_{ox}$  with addition the so-called “dark space” and poly-depletion thicknesses estimated at 0.65 nm together [19]. The junction depth  $X_j$  value is 14 nm according to 45 nm PTM model [19]. We then modify the BSIM4 model card to get the predicted  $S$  and  $\eta$ . The DIBL effect is easily tuned through **Eta0** parameter. In default PTM model cards, the short-channel degradation of  $S$  is modeled by **Nfactor** parameter. Al-

though the resulting  $S$  can fit the short-channel value, it is independent on  $L_g$ . As we consider statistical device simulations with  $L_g$  variations, we rather keep the long-channel  $S$  to its nominal value by setting **Nfactor** at 1 and then tune BSIM **cdsc** parameter (short-channel degradation of  $S$  from charge sharing) to adjust  $S$  to the predicted short-channel value. Thereby,  $S$  is affected by variability, which is very important for accuracy of Monte-Carlo statistical simulations, especially in subthreshold regime as recently confirmed by ring-oscillator measurements in [23]. Indeed a device with printed  $L_g$  of 30.8 nm instead of 35 nm ( $3\sigma$  deviation) features an  $S$  value degraded by 15%.

#### *Parasitic capacitances*

As shown in Chapter 2, parasitic capacitances in PTM models are roughly calibrated, which is fine when considering standard above-threshold operation. However, the gate capacitance is smaller in subthreshold regime because of the addition of the channel depletion capacitance in series with the oxide capacitance. Parasitic capacitances are thus proportionally more important and have to be considered carefully. We use the same modeling as in Section 2.3.3: models from [16] and [24] for fringing capacitances and parallel-plate approximation for overlap capacitances.

#### *Gate leakage*

We calibrate  $I_{gate}$  gate leakage vs. the 45 nm industrial model considered for the simulations of Fig. 3.1 at 1.1 and 1.7 nm  $T_{ox}$ . For extrapolation of  $I_{gate}$  to other  $T_{ox}$  values, we rely on the default BSIM  $I_{gate}$  vs.  $T_{ox}$  relationship, by only modifying  $T_{ox}$  parameter.

#### *Variability*

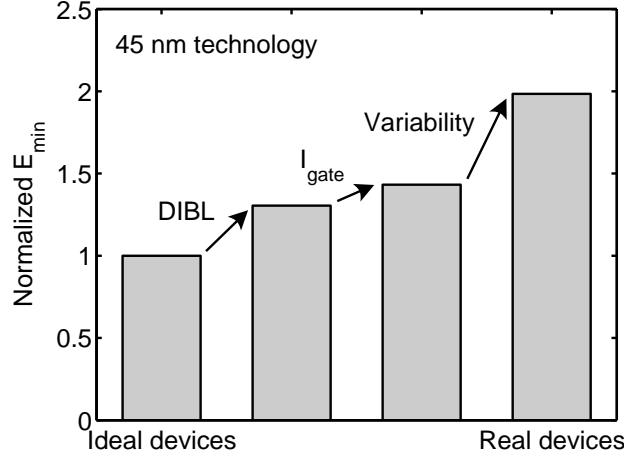
Similarly to Section 2.3.4, we consider two main variability sources: Random Doping Fluctuations (RDF) and Critical Dimension (CD) variations. RDF are modeled by a normally-distributed *vth0* BSIM parameter for each device, with standard deviation  $\sigma_{V_t}$  given by Asenov's empirical model [21]:

$$\sigma_{V_t} = \frac{A_{V_t, RDF}}{\sqrt{W} L_{eff}} = 3.19 \times 10^{-8} \frac{T_{ox} N_{ch}^{0.4}}{\sqrt{W} L_{eff}}. \quad (3.6)$$

CD variations are modeled by a normally-distributed  $L_g$  value with  $3\sigma_{L_g} = 12\% L_{g, min}$ , according to ITRS roadmaps [9]. We consider a single  $L_g$  variable common to all devices because CD variations exhibit a strong spatial correlation [22] while the benchmark circuit is quite small. Notice that when changing  $L_g$ , we keep  $\sigma_{L_g}$  constant, by assuming identical process resolution at a given technology node.

Finally, notice that variability is again addressed throughout this chapter by considering  $3\sigma$  worst-case delay and mean  $I_{leak}$ , which are statistically extracted from Monte-Carlo Spice simulations of the benchmark circuit.





**Fig. 3.3.** Contribution of nanometer MOSFET effects to minimum-energy level  $E_{min}$  for a benchmark 8-bit multiplier in 45 nm technology ( $V_{dd}$  is implicitly adapted to  $V_{dd,opt}$  to minimize the energy level for each device configuration).

### 3.4 LIMITATIONS FROM NANOMETER MOSFET EFFECTS

As shown in Fig. 3.1, new effects in nanometer technologies make  $E_{min}$  increase and deviate from  $C_L S^2$  trend. In order to investigate these effects, we consider the benchmark multiplier with nominal devices in a 45 nm High-Performance technology (device parameters can be found in first row of Table 3.1), with BSIM4 models generated according to the methodology presented in Section 3.3. The circuit is simulated first with “ideal devices”, i.e. without variability, gate leakage nor DIBL. The same simulation is then carried out by successively adding DIBL, gate leakage and variability. The resulting  $E_{min}$  values are plotted in Fig. 3.3 showing the high  $E_{min}$  overhead of these effects. Let us analyze the reasons of this overhead.

#### *Drain-induced barrier lowering*

The DIBL effect, which increases with technology scaling (Section 2.3.2), implies an exponential dependence of  $I_{sub}$  on  $V_{ds}$  as shown in Eq. (3.2). According to Eq. (3.3), DIBL should not impact  $E_{stat}$  as the  $I_{leak}$  increase from DIBL ( $\eta V_{dd}$  factor) is compensated by an equal delay reduction. In this equation, the delay is assumed to be inversely proportional to the maximum  $I_{on}$ , i.e. with  $V_{ds} = V_{dd}$ . However, during a transition, the current to charge/discharge the load capacitance is not constant. In particular as  $V_{ds}$  varies, the delay depends on the

integral of  $I_{sub}$  current over the transition  $Sw$ , that we model as:

$$\begin{aligned}
 T_{del} &\propto \frac{C_L V_{dd}}{\int^{Sw} I_{sub}} \approx \frac{C_L V_{dd}}{I_0 10^{\frac{V_{dd}}{S}} \times \int^{Sw} 10^{\frac{\eta V_{ds}}{S}}} \\
 &\approx \frac{C_L V_{dd}}{I_0 10^{\frac{V_{dd}}{S}} \times 10^{\frac{(1-k_{DIBL})\eta V_{dd}}{S}}} \\
 &= \frac{C_L V_{dd}}{I_0 10^{\frac{(1+\eta)V_{dd}}{S}}} \times 10^{\frac{k_{DIBL}\eta V_{dd}}{S}} \\
 &= \frac{C_L V_{dd}}{I_{sub,on}} \times 10^{\frac{k_{DIBL}\eta V_{dd}}{S}}, \tag{3.7}
 \end{aligned}$$

where  $k_{DIBL}$  is a fitting parameter, whose value depends on  $V_{dd}$  with  $0 < k_{DIBL} < 1$ . It accounts for the DIBL-induced  $I_{sub}$  reduction during a transition, the ideal case being  $k_{DIBL} = 0$ . The value of  $k_{DIBL}$  can be empirically extracted from simulation of the delay with and without the DIBL effect. Simulations of the benchmark circuit have been carried out with a large set of device parameters ( $L_g$ ,  $V_t$  and  $T_{ox}$ ) and show that the value of  $k_{DIBL}$  is pretty much independent on these parameters, provided that the devices actually stay in subthreshold regime, i.e.  $V_{dd} \leq V_t$ . The value of  $k_{DIBL}$  is 0.65 at 0.2V and 0.75 at 0.5V.

This shows that the DIBL effect has a delay overhead. When injecting the DIBL-aware delay expression from Eq. (3.7) into  $E_{stat}$  formula from Eq. (3.1), the impact of DIBL effect on  $E_{stat}$  clearly appears:

$$\begin{aligned}
 E_{stat} &= V_{dd} \times I_{leak} \times T_{del} \\
 &\propto V_{dd} \times I_0 10^{\frac{\eta V_{dd}}{S}} \times \frac{L_D C_L V_{dd}}{I_0 10^{\frac{(1+\eta)V_{dd}}{S}}} \times 10^{\frac{k_{DIBL}\eta V_{dd}}{S}} \\
 &\propto L_D C_L 10^{\frac{-V_{dd}}{S}} V_{dd}^2 \times 10^{\frac{k_{DIBL}\eta V_{dd}}{S}}. \tag{3.8}
 \end{aligned}$$

This DIBL-induced  $E_{stat}$  overhead is higher than 100% at 0.3V. Although the impact of DIBL on  $E_{min}$  is lower than on  $E_{stat}$  (30% as shown in Fig. 3.3) because dynamic energy is not affected by the DIBL effect, it is still very important to consider DIBL effect when making a technology choice for minimum-energy subthreshold circuits.

### Gate leakage

When shrinking  $T_{ox}$ , gate leakage exponentially increases and becomes comparable to subthreshold leakage in nanometer technologies. Therefore,  $E_{stat}$  expres-

sion must include  $I_{gate}$  contribution to total  $I_{leak}$ , i.e. back from Eq. (3.3):

$$\begin{aligned}
 E_{stat} &= V_{dd} \times I_{leak} \times T_{del} \\
 &\propto V_{dd} \times (I_0 10^{\frac{\eta V_{dd}}{S}} + I_{gate}) \times \frac{L_D C_L V_{dd}}{I_0 10^{\frac{(1+\eta)V_{dd}}{S}}} \\
 &\propto L_D C_L 10^{\frac{-V_{dd}}{S}} V_{dd}^2 \times (1 + \frac{I_{gate}}{I_0 10^{\frac{\eta V_{dd}}{S}}}) \\
 &\propto L_D C_L 10^{\frac{-V_{dd}}{S}} V_{dd}^2 \times (1 + \frac{I_{gate}}{I_{sub,off}}), \tag{3.9}
 \end{aligned}$$

which clearly shows that  $I_{gate}$  worsens  $E_{stat}$  unless it is much lower than subthreshold leakage. Rather than having a low absolute  $I_{gate}$ , the important target to minimize  $E_{min}$  is to keep the  $I_{gate}/I_{sub}$  ratio lower than 1.

### Variability

Device variability has a strong impact on subthreshold circuit delay as  $I_{sub}$  exponentially depends on  $V_t$  [12]. A detailed model of the device variability impact on subthreshold worst-case delay is presented in [25]. As the analytical relationship is quite complex, we do not recall it here. Nevertheless, we do consider variability in circuit simulations because the worst-case delay overhead results in  $E_{stat}$  and in turn  $E_{min}$  overheads (40% shown in Fig. 3.3). It is thus important to limit variability when considering optimum devices for minimum-energy circuits, as will be discussed in next sections.

This discussion show that there are new device targets in nanometer technologies to design/select optimum device and technology for minimum-energy subthreshold circuits:

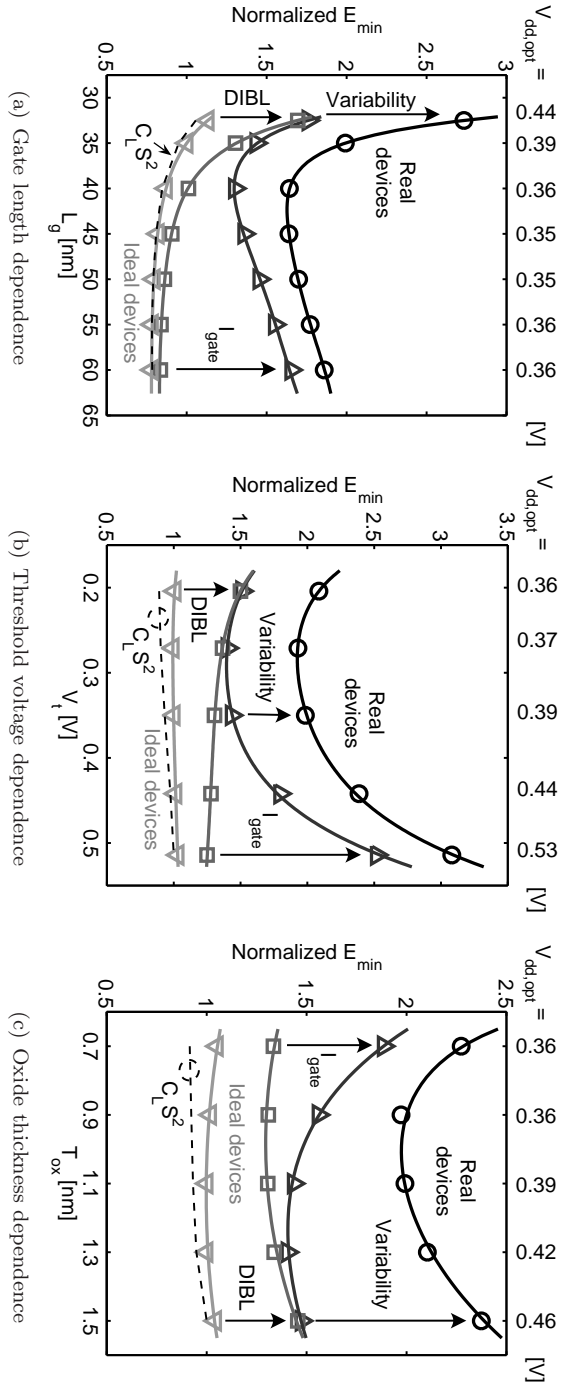
- low  $C_L S^2$  factor [1],
- low DIBL effect,
- $I_{gate}/I_{sub}$  ratio lower than 1,
- low variability [12].

## 3.5 IMPACT OF NANOMETER MOSFET PARAMETERS

Let us now investigate the impact of basic  $L_g$ ,  $V_t$ , and  $T_{ox}$  parameters on  $E_{min}$ , given the nanometer MOSFET effects.

### 3.5.1 Gate length impact

Fig. 3.4(a) shows  $E_{min}$  vs.  $L_g$  for ideal to real devices by successively adding DIBL, gate leakage and variability. The  $C_L S^2$  factor is plotted too for comparison purpose. For ideal devices,  $E_{min}$  exhibits the same dependence vs.  $L_g$  as



**Fig. 3.4.** Impact of individual device parameters on  $E_{min}$  (all other parameters being kept constant at nominal values for HP 45nm technology given in Table 3.1).  $V_{dd}$  is implicitly adapted to get the lowest energy level for each device configuration, leading to  $V_{dd,opt}$  values given at the top of each graph. Solid lines represent  $E_{min}$  from ideal devices (downward triangle markers) to real devices (circle markers) by successively enabling DIBL, gate leakage and variability. Previously-reported  $C_L S^2$  figure of merit [1] is plotted with dashed line for comparison purpose.  $E_{min}$  significantly deviates from  $C_L S^2$  trend because of DIBL, variability and gate leakage.

$C_L S^2$ : it decreases until 55 nm  $L_g$  thanks to  $S$  improvement from mitigation of short-channel behavior, whereas  $C_L$  increases slowly because it is dominated by parasitic capacitances [1].

High DIBL effect of short devices implies an important energy overhead. Long devices are affected by gate leakage due to a high  $I_{gate}/I_{sub}$  ratio. This not only comes from the slight  $I_{gate}$  increase with  $L_g$  but also from strong  $I_{sub}$  reduction from DIBL mitigation. Finally, variability worsens the picture for short devices. It comes from higher random doping fluctuations due to smaller channel area, as well as magnified current sensitivity against  $L_g$  variations because of high DIBL.

There is thus an optimum  $L_g$  to minimize  $E_{min}$ . In nanometer technologies, this optimum  $L_g$  results from a trade-off between variability and short-channel effect mitigation for long devices and low  $I_{gate}/I_{sub}$  for short devices. This trade-off clearly outweighs previously-reported  $C_L$  vs.  $S^2$  trade-off.

### 3.5.2 Threshold voltage impact

Minimum energy level and  $C_L S^2$  factor are plotted vs.  $V_t$  in Fig. 3.4(b). Lowering  $V_t$  implies reducing the channel doping  $N_{ch}$ , which results in a lower channel depletion capacitance  $C_{dep}$  and in turn a better subthreshold swing. Low- $V_t$  devices thus feature a low  $C_L S^2$  factor. Nevertheless, lowering  $V_t$  does not improve  $E_{min}$  of ideal devices because the  $S$  improvement is compensated by the fact the devices leave the subthreshold regime, which results in lower  $I_{on}/I_{off}$  ratio at a given  $V_{dd}$  and in turn  $E_{stat}$  overhead.

With a reduced  $N_{ch}$  for low- $V_t$  devices, the short-channel effects are increased and DIBL thus degrades  $E_{min}$ . On the other hand, higher  $V_t$  lowers  $I_{sub}$ , thereby degrading  $I_{gate}/I_{sub}$  ratio. Variability has an important impact on  $E_{min}$  for both high- and low- $V_t$  devices. High- $V_t$  devices have high  $\sigma_{V_t}$  because of high channel doping. Low- $V_t$  devices suffer from an important current sensitivity against  $\sigma_{L_g}$  due to their high DIBL. Optimum  $V_t$  selection thus results from a trade-off between DIBL mitigation and  $I_{gate}/I_{sub}$  ratio reduction.

### 3.5.3 Oxide thickness impact

Finally, the impact of  $T_{ox}$  on  $E_{min}$  is shown in Fig. 3.4(c). Notice that, for illustration purpose, we keep  $N_{ch}$  constant rather than  $V_t$  when varying  $T_{ox}$ , because of the direct  $S$ ,  $\eta$  and  $\sigma_{V_t}$  dependence on  $N_{ch}$ .

A  $T_{ox}$  reduction yields on one hand a better channel control by the gate, which results in an improved  $S$ . On the other hand, it increases the load capacitance [17]. The resulting  $C_L S^2$  factor is slightly improved for thin-oxide devices. No minimum  $C_L S^2$  is seen for the considered  $T_{ox}$  range because  $C_L$  increases slowly with  $T_{ox}$  reduction. This is a surprising observation because, although  $C_L$  is dominated by parasitic capacitances, one could expect a strong  $C_L$  reduction from fringing and overlap capacitance mitigation when increasing  $T_{ox}$ . Nevertheless, in nanometer technologies the outer fringing capacitance from gate to source/drain contacts is very important [24] as the gate electrode thickness is

high and the spacer width is small. This capacitance component hardly depends on  $T_{ox}$  and thus sets, together with junction capacitance, a lower bound on the achievable  $C_L$  reduction.

For ideal devices, a thinner  $T_{ox}$  implies a lower  $V_t$  at iso- $N_{ch}$ , which translates into an  $E_{min}$  increase as the devices leave the subthreshold regime, as explained in previous section. The DIBL-induced energy overhead is somewhat lower for thin-oxide devices, whereas the  $I_{gate}$  overhead is only important for thin-oxide devices, which feature a high  $I_{gate}/I_{sub}$  ratio despite their low  $V_t$ . Finally, variability makes  $E_{min}$  dramatically rise for thick  $T_{ox}$  because of low channel control and thus important RDF-induced  $\sigma_{V_t}$ , according to Eq. (3.6). Again, the trade-off between variability/short-channel behavior mitigation and low  $I_{gate}/I_{sub}$  ratio implies an optimum  $T_{ox}$  value.

### 3.6 OPTIMUM TECHNOLOGY AND DEVICE SELECTION

As shown in Fig. 3.1, direct porting of an ultra-low-power circuit from 90 nm to 45 nm general-purpose technologies results in 50%  $E_{min}$  overhead. However at 45 nm node, circuit designers have new opportunities as technologies are versatile. They offer a menu of several flavors, with various speed/power trade-offs resulting from different device configurations ( $L_g$ ,  $T_{ox}$  and  $V_t$ ). Moreover, each technology flavor often features dual or triple- $V_t$  devices and one can choose to use nominal (minimum) or longer  $L_g$ . Circuit designers thus face the complex problem of optimum technology/device selection with multiple degrees of freedom. In this section, we address this selection problem in two steps, considering planar bulk technologies. We first investigate the technology flavor the most adapted to minimum-energy subthreshold circuits. We then propose to optimize the device selection within a particular flavor to get the lowest  $E_{min}$ .

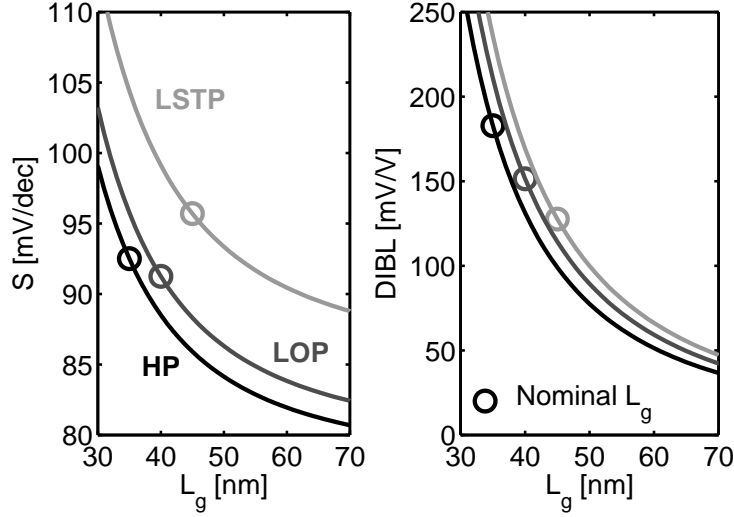
#### 3.6.1 Technology flavor comparison

For generality purpose, we consider the technology flavors recommended by the ITRS [9]: High-Performance (HP), low operating power (LOP) and Low-Stand-by Power (LSTP). The HP flavor is dedicated to super computers and servers. It features very low intrinsic gate delay  $C_g V_{dd}/I_{on}$ , which involves short  $L_g$ , low  $V_t$  and thin  $T_{ox}$ . The value of HP parameters we consider in this investigation is given in Table 3.1, assuming a 45 nm node. We use the methodology from Section 3.3 to generate corresponding BSIM4 models. The resulting device characteristics under nominal  $V_{dd}$  are also given.

The LSTP flavor targets portable devices such as cell phones. The goal is to reduce static power from leakage currents by having longer  $L_g$ , higher  $V_t$  and thicker  $T_{ox}$ . In order to maintain sufficient speed, nominal  $V_{dd}$  is higher than for HP flavor. LOP flavor is intended for portable yet fast devices such as laptops. The target is to operate at a reduced  $V_{dd}$  for dynamic power concern with intermediate static power. This leads to intermediate device parameter values.

**Table 3.1.** Considered device features and resulting characteristics under nominal  $V_{dd}$  at 45 nm node (std- $V_t$  devices)

Tech. flavor	$V_{dd}$ [V]	$L_g$ [nm]	$L_{eff}$ [nm]	$T_{ox}$ [nm]	$V_t$ [V]	$N_{ch}$ [#/ $cm^3$ ]	$I_{on}$ [mA/ $\mu m$ ]	$I_{off}$ [nA/ $\mu m$ ]	$C_g$ [fF/ $\mu m$ ]	$C_g V_{dd}/I_{on}$ [ps]	$\sigma_{V_t}$ [mV]	$\sigma_{L_g}$ [nm]
HP	1.0	35	17.5	1.1	0.35	$3.8 \cdot 10^{18}$	1100	30	0.41	0.37	44	1.4
LOP	0.9	40	21	1.3	0.43	$3.5 \cdot 10^{18}$	690	3.0	0.42	0.54	43	1.4
LSTP	1.2	45	25	1.7	0.6	$4.1 \cdot 10^{18}$	700	0.05	0.39	0.68	51	1.4



**Fig. 3.5.** Subthreshold swing (left) and DIBL factor (right) for the ITRS technology flavors (model from the voltage-doping transformation [20])

Notice that the 3 flavors should come with triple- $V_t$  devices [9] but we only consider std- $V_t$  devices in this section, in order to ease the flavor comparison.

Fig. 3.5 shows the subthreshold swing  $S$  and DIBL factor  $\eta$  vs.  $L_g$  for the 3 flavors and Table 3.2 summarizes the main MOSFET characteristics in subthreshold regime. At iso- $L_g$ , the HP flavor features the best  $S$  thanks to its thin oxide. At nominal  $L_g$ , the LOP flavor features the best  $S$  because its low nominal  $V_{dd}$  relaxes the constraints on minimum  $T_{ox}$  and  $N_{ch}$  for target gate and subthreshold leakages, which results in a good channel control. For the same reason, it achieves the lowest subthreshold  $I_{on}$  variability.

All flavors exhibit very low  $I_{gate}/I_{sub}$  ratio. In order to explain this observation, let us recall that we consider std- $V_t$  devices in triple- $V_t$  technologies. Within each flavor, low-, std- and high- $V_t$  devices basically exhibit identical  $I_{gate}$  due to their identical  $T_{ox}$ . The typical constraint on  $I_{gate}$  is to keep it lower than  $I_{sub}$  for power/performance trade-off. This constraint gives  $I_{sub,high-V_t}$  as an upper bound on  $I_{gate}$ , which results in  $I_{sub,stdV_t} > I_{sub,highV_t} \geq I_{gate}$ .

Spice-simulation of the benchmark multiplier show that the LOP flavor exhibits the lowest  $E_{min}$ , as shown in Table 3.3, as a consequence from its lowest  $S$ , medium DIBL and lowest variability. The LSTP  $E_{min}$  is very close to LOP thanks to low DIBL.

Table 3.3 also shows the optimum supply  $V_{dd,opt}$  that leads to  $E_{min}$ , as well as the multiplier  $3\sigma$  worst-case delay  $T_{del}$ , delay variability and worst-case SNM at  $V_{dd,opt}$ . HP flavor features the lowest delay at  $V_{dd,opt}$  corresponding to a 7 MHz  $f_{clk,opt}$ . LOP delay is close to the HP one, the corresponding  $f_{clk,opt}$



**Table 3.2.** Subthreshold MOSFET characteristics in 45 nm bulk technology

Tech. flavor	$S$ [mV/dec]	$\eta$ [mV/V]	$I_0$ [pA/ $\mu m$ ]	$I_{on}$ var. <sup>†</sup> [—]	$I_{gate}/I_{sub}$ [—]
HP	92.5	183	340	29.7	0.09
LOP	91.3	151	100	26.3	0.06
LSTP	95.7	128	1.3	41.9	0.04
HP <sup>*</sup> <sub>opt</sub>	80.1	60.3	252	15.6	0.84

\*Low- $V_t$  devices with 50 nm  $L_g$ .†Ratio between mean and  $3\sigma$  WC  $I_{on}$  at 0.2V.**Table 3.3.** Subthreshold circuit performances at minimum-energy point in 45 nm bulk technology

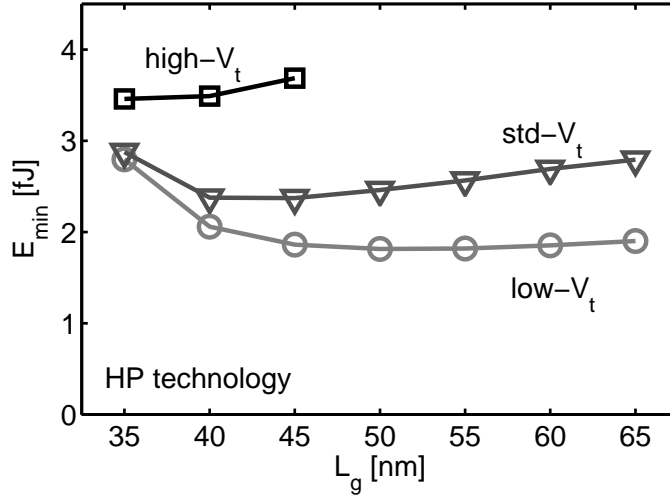
Tech. flavor	$E_{min}$ [fJ]	$V_{dd,opt}$ [V]	$3\sigma$ WC $T_{del}$ [ $\mu s$ ]	$T_{del}$ var. <sup>†</sup> [—]	$P_{stat}$ [nW]	$3\sigma$ WC SNM [mV]
HP	28.8	0.39	0.14	1.76	60.7	61.3
LOP	24.6	0.36	0.75	1.85	9.4	63.3
LSTP	25.2	0.38	62.4	2.54	0.1	68.3
HP <sup>*</sup> <sub>opt</sub>	18.1	0.30	0.59	1.46	9.8	68.5

\*Low- $V_t$  devices with 50 nm  $L_g$ .†Ratio between  $3\sigma$  WC and mean  $T_{del}$  at  $V_{dd,opt}$ .

being 1.4 MHz. LSTP flavor features a prohibitive delay, which limits the clock frequency to the low-kHz range (16 kHz  $f_{clk,opt}$ ). This makes LSTP flavor a bad option for minimum-energy subthreshold circuits, as it considerably restricts the application spectrum. Notice that LOP delay variability is somewhat higher than for HP despite its lower  $I_0$  variability because LOP  $V_{dd,opt}$  is slightly lower, which increases delay variability [25].

Table 3.3 also shows mean static power  $P_{stat}$  and worst-case SNM at  $V_{dd,opt}$ . HP flavor of course exhibits the highest  $P_{stat}$ . All flavors feature very close worst-case SNM. Thanks to its lower  $I_{on}$  variability, LOP SNM remains good even with its lower  $V_{dd,opt}$ .

The 15% energy gain of the LOP flavor is not a breakthrough improvement, which shows that technology flavor selection is not efficient for  $E_{min}$  reduction. In next section, we thus investigate whether device selection within a particular flavor can bring a higher energy gain. We choose the HP flavor as its short delay gives more space for optimum device selection.



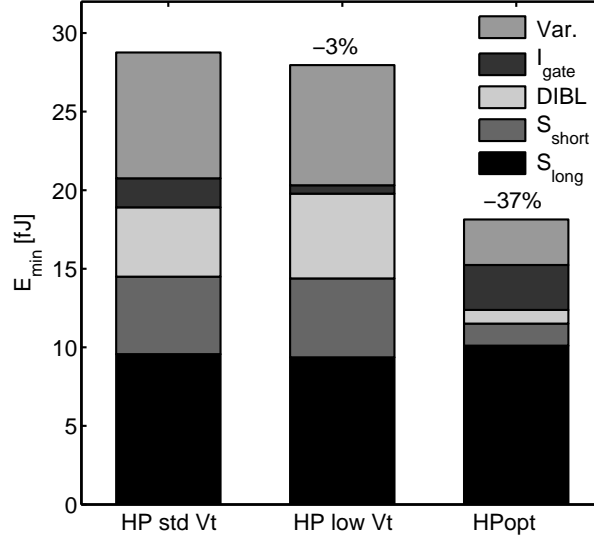
**Fig. 3.6.**  $E_{min}$  vs. gate length in multi- $V_t$  technology (HP technology flavor, considered  $V_t$  values are 0.27, 0.35 and 0.44V).

### 3.6.2 Optimum device selection

Within a given technology flavor, circuit designers have 2 degrees of freedom for device selection:  $V_t$  and  $L_g$ .  $V_t$  can be chosen between 2 or 3 discrete values, while  $L_g$  can take any value higher than the nominal/minimum  $L_g$ . In 45 nm technologies, there are often restrictive design rules that prevent circuits designers from using any  $L_g$ , for regularity issues. We therefore consider  $L_g$  values that are multiples of 5 nm. Fig. 3.6 shows  $E_{min}$  for the devices with 3 considered  $V_t$  vs.  $L_g$ , within the HP technology flavor.

For high- $V_t$  devices,  $E_{min}$  is the highest because of its high  $I_{gate}/I_{sub}$  ratio. Therefore, an  $L_g$  upsize further degrades  $E_{min}$  by worsening this ratio. For low- and std- $V_t$  devices,  $E_{min}$  is comparable at nominal  $L_g$ . When upsizing  $L_g$ ,  $E_{min}$  is first improved thanks to variability and short-channel behavior mitigation but is then degraded by higher  $I_{gate}/I_{sub}$  ratio, as detailed in Section 3.5.1. This ratio is lower for low- $V_t$  devices and  $L_g$  can thus further be upsized to mitigate more efficiently variability and short-channel behavior, while keeping  $I_{gate}/I_{sub}$  low. At 50 nm  $L_g$ , low- $V_t$  devices yields a 37%  $E_{min}$  reduction, as compared to nominal  $L_g$  std- $V_t$  devices, thereby making low- $V_t$  long devices the optimum choice for minimum-energy subthreshold circuits.

Fig. 3.7 shows the breakdown of the contributions from nanometer MOSFET effects to  $E_{min}$  for std- $V_t$ , low- $V_t$  and optimum HP<sub>opt</sub> devices (low- $V_t$  devices with 50 nm  $L_g$ ). It shows that moving from std- $V_t$  to low- $V_t$  devices efficiently reduces gate leakage contribution and thus relaxes the constraint on  $L_g$ . Upsizing  $L_g$  drastically mitigates variability, DIBL and short-channel  $S$  contributions, with



**Fig. 3.7.** Breakdown of  $E_{min}$  contributions in HP technology flavor: ideal devices with long-channel  $S$  (obtained by setting BSIM `cdsc` parameter to zero), short-channel  $S$  degradation, DIBL, gate leakage and variability (capacitance contribution is included in long-channel  $S$ ).  $HP_{opt}$  devices are low- $V_t$  devices with 50 nm  $L_g$ .

a tolerable increase in gate leakage and capacitance (included in long-channel  $S$  contribution).

Subthreshold characteristics of the optimum  $HP_{opt}$  device are given in Table 3.2 and corresponding circuit performances in Table 3.3. It first shows that low- $V_t$  devices with upsized  $L_g$  in HP technology bring lower  $E_{min}$  than standard devices in LOP technology. The  $HP_{opt}$   $I_{gate}/I_{sub}$  ratio is higher than for nominal devices in any flavor but still lower than 1. This indicates that, for energy concern, it is worth tolerating higher  $I_{gate}$  to limit  $S$ , DIBL and variability.  $I_{sub}$  sets an upper bound on this  $I_{gate}$  increase. Secondly, although  $V_{dd,opt}$  is somewhat lower than for standard devices, the delay remains short enough for many ultra-low-power applications and smaller than for LOP flavor with a 1.7 MHz  $f_{clk,opt}$  and similar  $P_{stat}$ . As an extra benefit,  $HP_{opt}$  circuit features reduced delay variability and improved SNM despite its lower  $V_{dd,opt}$ .

We carried out the same optimum device selection in the industrial general-purpose 45 nm technology considered in Fig. 3.1. As this technology is a dual- $V_t$  technology, the lowest- $V_t$  device was already considered as the standard devices in Fig. 3.1. Nevertheless, simulations show that an  $L_g$  upsize from 35 to 60 nm alone leads to 40%  $E_{min}$  improvement with only 10% MOSFET area overhead.

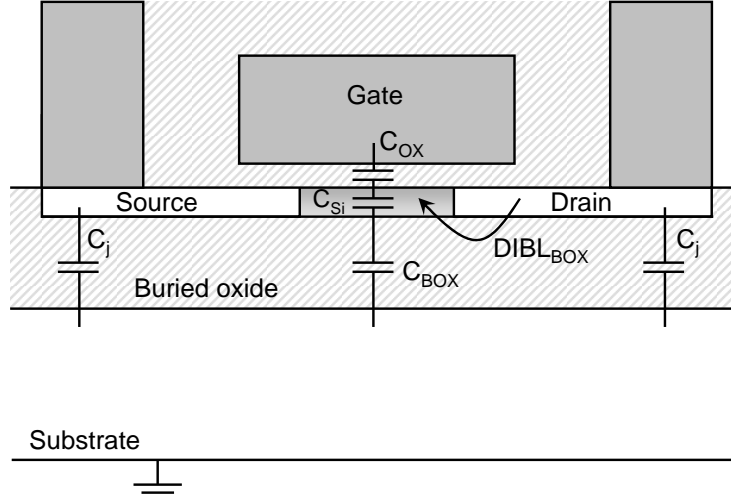


Fig. 3.8. Sketch of an FD SOI MOSFET

This overhead is assumed to weakly impact total die area because interconnection area remain unchanged.

### 3.7 FULLY-DEPLETED SOI TECHNOLOGY

Fully-depleted (FD) ultra-thin-body SOI technology features improved robustness against short-channel effects [26]. Indeed, as shown in Fig. 3.8, the addition of a buried oxide (BOX) enables FD SOI MOSFETs with an active Silicon film much thinner than both junction and depletion depths in bulk MOSFETs. Moreover, as the depletion depth is limited by the BOX, it does not depend on the channel doping  $N_{ch}$ , which can thus be reduced to improve carrier mobility and limit random doping fluctuations [20]. It is thus expected to be indispensable beyond 32 nm node [9]. The BOX also isolates source/drain diffusions from the substrate, which drastically reduces junction capacitances as an extra benefit [26]. In this section, we carry out a prospective study of the FD SOI potential to reduce minimum-energy level in nanometer subthreshold logic circuits. We first present the differences with bulk technology in the Pre-Silicon modeling approach from Section 3.3. We then examine at 45 nm node the impact of SOI technology first on subthreshold MOSFET operation and then on minimum-energy subthreshold circuits.

#### 3.7.1 Pre-Silicon FD SOI MOSFET compact models

We use a modeling approach similar to the one for bulk technology presented in Section 3.3. Although it is basically intended for bulk MOSFETs, we consider

BSIM4 compact model for modeling SOI MOSFETs. This allows us to get a fair comparison between bulk and SOI, without introducing disturbances due to parameter compatibility issues between BSIM and BSIM-SOI models. Moreover, as SOI devices are MOSFET devices, their current-voltage characteristics can basically be modeled by bulk MOSFET equations, provided that accurate parameter values are selected. We start from the 45 nm PTM models in a HP technology flavor by relying on their second-order parameters and we then modify several parameters to emulate FD SOI MOSFETs. In this section, we detail the parameters we change from bulk models: body effect, subthreshold current, variability and capacitances. We consider standard FD SOI wafers with Silicon-film  $T_{Si}$  and buried-oxide  $T_{BOX}$  thicknesses of 10 and 145 nm, respectively [27] and we target HP technology flavor with  $L_g$  and  $T_{ox}$  values equal to bulk ones from Table 3.1.

#### Body effect

FD SOI devices feature body effect much lower than bulk devices. Long-channel calculation [26] of the linearized body-effect coefficient  $\gamma$  for the considered bulk and FD SOI technologies gives 0.30 ( $C_{dep}/C_{ox}$ ) and 0.012 ( $series(C_{Si}, C_{BOX})/C_{ox}$ ) values, respectively. With a negligible loss of accuracy, we consider  $\gamma = 0$  in FD SOI. In order to model this feature in a BSIM4 bulk MOSFET model, we tie the body access of the devices to their source and sets primary body-effect parameters ( $k1$  and  $k2$ ) to zero. Notice that the resulting model is an asymmetric model that can only be used provided that source (resp. drain) voltage remains below drain (resp. source) voltage  $V_s \leq V_d$  for NMOS (resp. PMOS) devices. As this condition is always respected in static CMOS logic style, we can use this model for carrying out subthreshold simulations of the benchmark multiplier.

#### Subthreshold current parameters

We use Skotnicki's equations from the voltage-doping transformation of electrostatic integrity [20] for empirical calculation of the subthreshold swing  $S$  and the DIBL factor  $\eta$ , given the considered  $T_{Si}$  and  $T_{BOX}$  values:

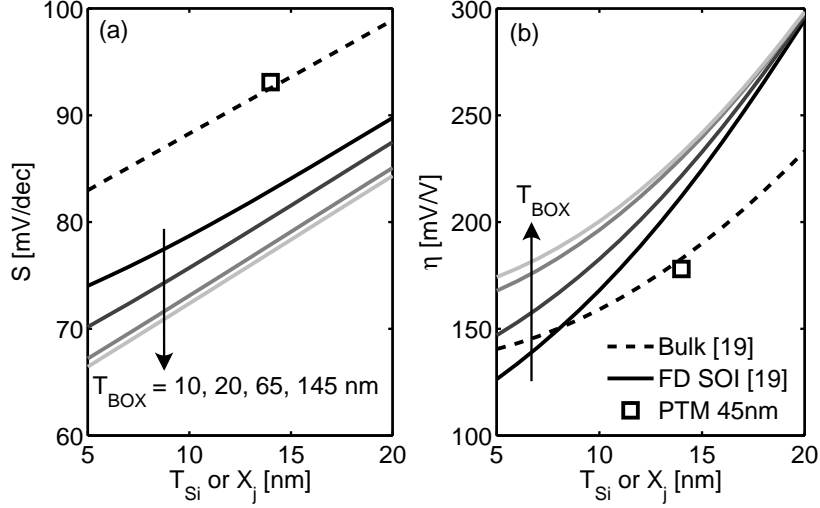
$$S = U_t \times \ln(10) \times \left( 1 + \frac{series(C_{Si}, C_{BOX})}{C_{ox}} + \frac{\epsilon_{Si}}{\epsilon_{ox}} EIS \sqrt{1 + 2 \frac{V_{ds}}{\Phi_d}} \right) \quad (3.10)$$

$$with EIS = \frac{T_{ox,el}}{L_{g,el}} \frac{T_{Si}}{L_{el}} \times \left( 1 + \frac{3}{4} \frac{T_{Si} + \lambda T_{BOX}}{L_{g,el}} \right),$$

$$\eta = 0.8 \times \frac{\epsilon_{Si}}{\epsilon_{ox}} \times EI \quad (3.11)$$

$$with EI = \frac{T_{ox,el}}{L_{g,el}} \frac{T_{Si} + \lambda T_{BOX}}{L_{g,el}} \times \left( 1 + \frac{T_{Si}^2}{L_{g,el}^2} \right),$$

where  $\lambda$  is a fitting parameter to take into account the contribution of BOX fringing field on DIBL effect (see Fig. 3.8), the ideal case being  $\lambda=0$ . An approx-



**Fig. 3.9.** Subthreshold swing (a) and DIBL factor (b) in 45 nm bulk and SOI technologies ( $L_g = 35$  nm,  $T_{ox} = 1.1$  nm at 0.2V)

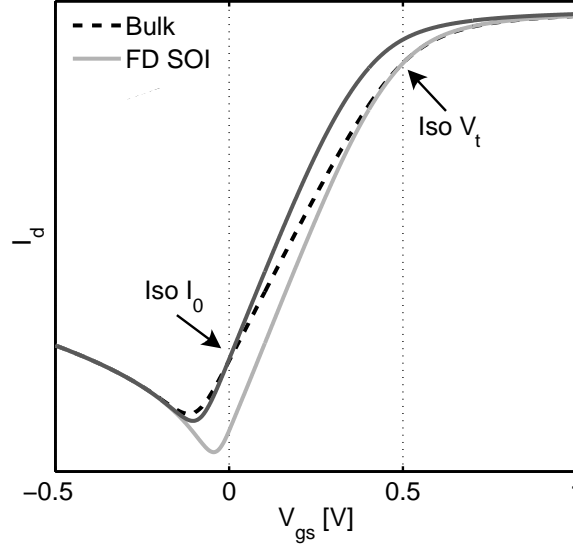
imative expression of  $\lambda$  can be found in MASTAR5 built-in equations [28], which yields 0.1 value for the considered  $T_{Si}$ ,  $T_{BOX}$  and  $L_{g,el}$  parameters.

Fig. 3.9 illustrates resulting  $S$  and  $\eta$  values for FD SOI and bulk technologies, when varying  $T_{Si}$  and  $X_j$ , respectively. The  $S$  and  $\eta$  values for bulk technology are extracted at 14 nm  $X_j$  from PTM model cards and closely fits  $S$  and  $\eta$  values from default PTM model. Both  $S$  and  $\eta$  in FD SOI technology are improved at thin  $T_{Si}$ , while thin  $T_{BOX}$  degrades  $S$  from higher body effect and improves  $\eta$  from reduced BOX fringing field. In the BSIM4 model card for FD SOI MOSFET, we thus tune BSIM `Nfactor` parameter to fit long-channel  $S$ , i.e.  $1 + \frac{series(C_{Si}, C_{BOX})}{C_{ox}}$  terms from Eq. (3.10) and then `cdsc` parameter for short-channel  $S$  adjustment.

For comparison fairness concern, we target a subthreshold  $I_0$  reference current identical to bulk technology. As detailed in Eq. (1.8) and (1.10),  $I_0$  is proportional to  $10^{-V_t/S}$ . The small  $S$  of FD SOI thus implies that  $V_t$  has to be reduced for meeting the  $I_0$  value, as illustrated in Fig. 3.10. We consider two versions of FD SOI MOSFETs for having  $I_0$  identical to bulk [20, 29]:

1. doped channel with same  $N_{ch}$  as bulk and polysilicon gate electrode,
2. undoped channel and work-function engineering with a midgap-metal gate electrode.

In the BSIM4 model card for FD SOI MOSFET, we only tune `vth0` parameter. BSIM `Ndep` parameter for channel doping concentration impacts body effect,  $S$  and  $\eta$  through  $X_{dep}$ . As we tuned other BSIM parameters to reflect FD SOI



**Fig. 3.10.** Sketch of bulk and FD SOI subthreshold  $I/V$  characteristics (low  $V_{ds}$ )

MOSFET behavior, we leave BSIM  $N_{dep}$  parameter unchanged, although we implicitly assume an  $N_{ch}$  of  $10^{16}$  dopant/cm<sup>3</sup> in undoped FD SOI MOSFETs. We neglect the impact of  $N_{ch}$  on carrier mobility. Indeed, its impact on minimum energy through  $I_0$  term is weak as confirmed by the simplification of  $I_0$  terms in  $E_{stat}$  expression from Eq. (3.3). Finally, notice that the use of a midgap-metal gate removes the poly-depletion layer, which results in a 0.3 nm thinner  $T_{ox,el}$  at iso-physical  $T_{ox}$ .

#### Variability

Gate length variability is kept identical in FD SOI model but its impact on subthreshold current is different as  $S$  and  $\eta$  parameters are different. Regarding RDF-induced variability, there is no empirical expression in FD SOI similar to Eq. (3.6) for bulk. We thus use analytical calculation.

Threshold voltage variability induced by RDF can analytically be calculated by partial derivation of  $V_t$  expressions [26] by neglecting  $\Phi_F$  contribution as [20]:

$$\sigma_{V_t, RDF} = \frac{\partial V_t}{\partial n_c} \sigma_{n_c} = \frac{1}{C_{ox}} \frac{\partial Q_c}{\partial n_c} \sigma_{n_c} = \frac{q T_{ox}}{\epsilon_{ox} WL} \sigma_{n_c}, \quad (3.12)$$

where  $Q_c = q n_c / WL$  is the channel depletion charge and  $n_c$  is the number of dopants in the depleted channel region. In bulk technology,  $n_{c,bulk} = N_{ch} X_{dep} WL$  and in FD SOI technology,  $n_{c,SOI} = N_{ch} T_{Si} WL$ . If we assume a Gaussian distribution for the channel dopants, we have  $\sigma_{n_c}^2 = n_c$  [20, 30]. Injecting these relationships in Eq. (3.12) yields the following expressions for bulk and

**Table 3.4.** Comparison of contributions to  $V_t$  standard deviation ( $\sigma_{V_t}$ ) between bulk and FD SOI 45 nm technologies (minimum-sized devices)

$\sigma_{V_t}$ [mV] due to	RDF <sub>emp</sub>	RDF <sub>anal</sub>	RDF <sub>norm</sub>	$T_{Si}$	Other	Total
Bulk	44	62	44	0	15	46
Doped FD SOI	-	46	32	15	15	39
Undoped* FD SOI	-	2.0	1.4	$\sim 0$	15	15.1

\* $V_t$  tuned by work-function engineering with a midgap-metal gate.

FD SOI  $V_t$  standard deviations:

$$\sigma_{V_t, RDF}|_{bulk} = \sqrt[4]{4 q^3 \epsilon_{Si} \Phi_F N_{ch}} \times \frac{T_{ox}}{\epsilon_{ox} \sqrt{WL}}, \quad (3.13)$$

$$\sigma_{V_t, RDF}|_{SOI} = q \sqrt{N_{ch} T_{Si}} \times \frac{T_{ox}}{\epsilon_{ox} \sqrt{WL}}. \quad (3.14)$$

As shown in Table 3.4, the  $\sigma_{V_t}$  due to RDF when analytically calculated (RDF<sub>anal</sub>) from Eq. (3.13) and from Eq. (3.14) at iso- $N_{ch}$  is slightly lower in FD SOI than in bulk. In bulk technology, the  $\sigma_{V_t}$  value analytically calculated (RDF<sub>anal</sub>) from Eq. (3.13) is larger than the value from empirical expression (RDF<sub>emp</sub>) of Eq. (3.6) [21]. As this empirical expression has been shown to offer a reliable approximation, we keep the empirical value for bulk and normalize  $\sigma_{V_t}$  analytical values for FD SOI by the ratio between empirical (RDF<sub>emp</sub>) and analytical (RDF<sub>anal</sub>) bulk values. For FD SOI with undoped channel,  $\sigma_{V_t}$  due to RDF is close to zero.

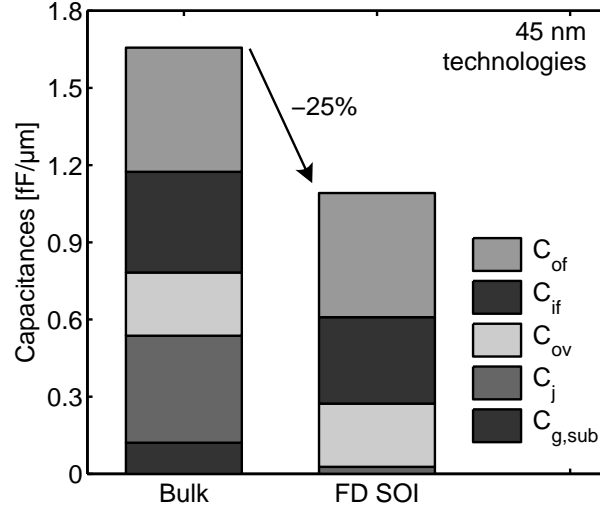
In FD SOI technology, not only RDF contributes to  $V_t$  variations. Indeed, Silicon thickness  $T_{Si}$  variations affect the threshold voltage [31]. Contribution of  $T_{Si}$  variations to  $\sigma_{V_t}$  can be found by partial derivative of FD SOI  $V_t$  expression:

$$\sigma_{V_t, T_{Si}} = \frac{\partial V_t}{\partial T_{Si}} \sigma_{T_{Si}} = \frac{q N_{ch, SOI}}{C_{ox}} \sigma_{T_{Si}}, \quad (3.15)$$

which is  $3\times$  lower than  $\sigma_{V_t, RDF}$  (analytical calculation), as shown in Table 3.4 when considering a  $3\sigma_{T_{Si}}$  of 1.5 nm.

As  $\sigma_{V_t, RDF}$  is low in FD SOI technology, it may not be the dominating source of  $V_t$  variability. In order to avoid biasing bulk vs. FD SOI comparison, we consider an extra 15 mV contribution to total  $\sigma_{V_t}$  to account for  $T_{ox}$  variations, line edge roughness and other variation sources [32]. Similarly to  $\sigma_{V_t, RDF}$ , this  $\sigma_{V_t, other}$  contribution is modeled with an inverse dependence on  $\sqrt{WL}$  according to Pelgrom's model [33]. For the sake of simplicity, we model RDF,  $T_{Si}$  and the other  $V_t$  variability causes as independent sources, meaning that they add in quadrature. The resulting total  $\sigma_{V_t}$  is given in Table 3.4, which shows that doped and undoped FD SOI MOSFETs feature 15% and 65% variability reduction, respectively, despite  $T_{Si}$  variations. Very recently, the low variability of undoped-channel FD SOI MOSFETs has been experimentally demonstrated in [34].





**Fig. 3.11.** Comparison of device capacitance between bulk and FD SOI 45 nm technologies

### Capacitances

As shown in Fig. 3.8, the subthreshold  $C_{g,sub}$  gate capacitance is the series connection of gate-oxide, Silicon-film and buried-oxide capacitances. Fig. 3.11 shows that  $C_{g,sub}$  is negligible in FD SOI technology as confirmed in [13] for double-gate devices. Nevertheless, when using BSIM4 bulk MOSFET compact model, we have no possibility to tune  $C_{g,sub}$  and we leave it at bulk value as a worst-case approach.

The main advantage of SOI technology regarding parasitic capacitances is the reduction of junction capacitances. As shown in Fig. 3.8, the BOX electrically isolates the diffusions from the substrate. The capacitance is thus given by  $C_{BOX}$  that we model by parallel plate approximation. As shown in Fig. 3.11, overlap and outer fringing capacitances are equal in bulk and FD SOI technologies. Inner fringing capacitance is slightly reduced in FD SOI because  $T_{Si}$  replaces  $X_{dep}$  in Eq. (2.9). The predicted total capacitance reduction is 25%.

### 3.7.2 Subthreshold characteristics of FD SOI MOSFETs

Subthreshold MOSFET characteristics in FD SOI technology are given in Table 3.5 and compared to bulk. Doped-channel FD SOI MOSFETs feature better subthreshold swing from improved long-channel contribution and mitigated short-channel behavior, as well as reduced subthreshold  $I_{on}$  variability and total FO4 inverter load capacitance  $C_L$ . However, the DIBL effect is somewhat increased

**Table 3.5.** Subthreshold MOSFET characteristics in 45 nm bulk and SOI technologies (HP flavor)

45 nm technologies	$S$ [mV/dec]	$\eta$ [mV/V]	$I_0$ [pA/ $\mu\text{m}$ ]	$I_{on}$ var. <sup>†</sup> [—]	$C_L$ [fF/ $\mu\text{m}$ ]
Bulk	92.5	183	340	30.8	21.5
Doped FD SOI	72.4	199	340	26.8	17.1
Undoped* FD SOI	70.2	167	340	3.3	18.7

\* $V_t$  tuned by work function engineering with a midgap-metal gate.<sup>†</sup>Ratio between mean and  $3\sigma$  worst-case  $I_{on}$  at 0.2V.**Table 3.6.** Subthreshold circuit performances at minimum-energy point in 45 nm bulk and SOI technologies (HP flavor)

45 nm technologies	$E_{min}$ [fJ]	$V_{dd,opt}$ [V]	WC $T_{del}$ [ $\mu\text{s}$ ]	$T_{del}$ var. <sup>‡</sup> [—]	$P_{stat}$ [nW]	WC SNM [mV]
Bulk	29.6	0.40	0.14	1.89	64.7	62.4
Doped FD SOI	16.5	0.33	0.08	1.76	73.7	44.9
Undoped* FD SOI	11.8	0.27	0.20	1.15	23.0	51.0
Undoped* FD SOI	18.6 <sup>‡</sup>	0.40 <sup>‡</sup>	0.015	1.10	81.8	98.9

\* $V_t$  tuned by work function engineering with a midgap-metal gate.<sup>‡</sup>Ratio between  $3\sigma$  worst-case and mean delay at  $V_{dd,opt}$ .<sup>‡</sup>Iso- $V_{dd}$  comparison.

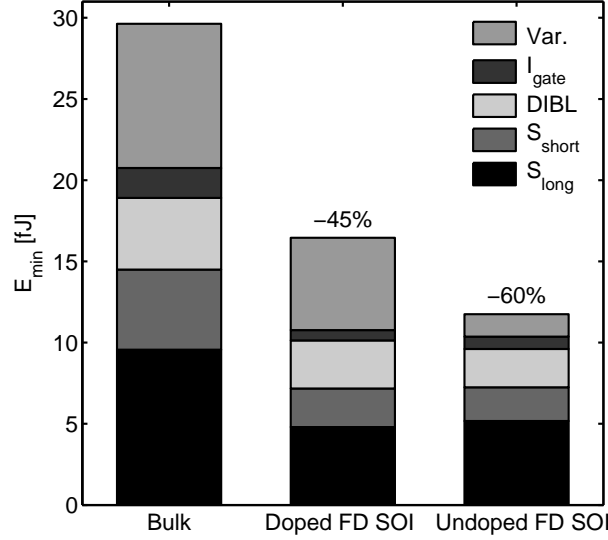
for such a thick BOX (145 nm) because of BOX fringing field. This could be mitigated by the use of thin-BOX technology [20, 29] at the expense of junction capacitance, body effect and thus long-channel  $S$ , or by using a low- $\kappa$  BOX, but this is beyond the scope of this dissertation.

The combination of an undoped channel with midgap-metal gate further improves subthreshold MOSFETs characteristics. Smaller electrical  $T_{ox}$  improves  $S$  and  $\eta$  thanks to poly-depletion removal (0.3 nm), while the low doping level efficiently mitigates  $I_{on}$  variability. Let us examine the impact on minimum-energy subthreshold circuits.

### 3.7.3 Minimum-energy subthreshold circuits in FD SOI technology

Simulations of the benchmark multiplier have been carried out with FD SOI models generated according to Section 3.7.1. Table 3.6 shows the simulated minimum-energy levels  $E_{min}$  in bulk and FD SOI technologies. Notice that bulk  $E_{min}$  is 3% higher than in Table 3.3 because of the addition of  $\sigma_{V_t,other}$  contribution to  $V_t$  variability to account for oxide thickness variations and line edge roughness.

As compared to bulk,  $E_{min}$  is reduced by 45% in doped-channel and 60% in undoped-channel FD SOI technologies. Fig. 3.12 shows a breakdown of the contributions to  $E_{min}$ . First, FD SOI reduces  $E_{min}$  of ideal devices by improving long-



**Fig. 3.12.** Breakdown of  $E_{min}$  contributions: ideal devices with long-channel  $S$ , short-channel  $S$  degradation, DIBL, gate leakage and variability (capacitance contribution is included in long-channel  $S$ ).  $V_t$  of undoped FD SOI technology is tuned by work function engineering with midgap-metal gate.

channel  $S$  and reducing switched capacitances. It then mitigates short-channel degradation of  $S$ . The DIBL contribution to  $E_{min}$  is higher in doped-channel FD SOI (+40%) than in bulk (+30%). The use of a metal gate limits DIBL contribution but increases ideal  $E_{min}$  because of higher switched capacitances. Gate-leakage contribution is equivalent in bulk and FD SOI but variability contribution is drastically reduced in undoped FD SOI. This shows that FD SOI technology has an important potential for minimum-energy subthreshold circuits in nanometer technologies.

Table 3.6 also shows  $V_{dd,opt}$  and corresponding circuit performances.  $V_{dd,opt}$  is lower in FD SOI technologies because of lower  $E_{stat}$ . Despite its lower  $V_{dd,opt}$ , doped-channel FD SOI features an improved delay almost divided by 2 thanks to better  $S$  and low stack effect (no body effect). Its corresponding  $f_{clk,opt}$  is 13 MHz. However,  $P_{stat}$  is somewhat higher than in bulk because of low stack effect and higher DIBL.

Undoped-channel FD SOI features even lower  $V_{dd,opt}$ , which results in an increased delay (5 MHz corresponding  $f_{clk,opt}$ ) and a lower SNM than bulk. Nevertheless, delay variability and  $P_{stat}$  are much smaller. In order to extend this comparison, Table 3.6 shows the performances of undoped-channel FD SOI under 0.4 V, i.e. bulk  $V_{dd,opt}$ . As compared to bulk, corresponding iso- $V_{dd}$  energy

per operation is reduced by 35%, delay is  $10\times$  lower with low variability and SNM is improved by 60%. Finally, at iso-delay<sup>2</sup>, undoped-channel FD SOI still features 59% energy saving and at iso-SNM<sup>3</sup>, undoped-channel FD SOI features 58% energy saving and 50% delay improvement. This shows that FD SOI is of uttermost interest for subthreshold logic as it brings improved performances in all possible trade-offs.

### 3.8 CONCLUSION

In this chapter, we analyzed the minimum-energy point of subthreshold circuits in standard nanometer bulk CMOS technologies. Simulations of a benchmark multiplier show that direct porting to 45 nm technology leads to serious  $E_{min}$  overhead. We confirmed that this overhead partly comes from high variability and subthreshold swing. We reported and demonstrated by circuit simulation and analytical modeling that DIBL and gate leakage contribute to this overhead as much as variability. We then investigated the impact of nanometer MOSFET parameters  $L_g$ ,  $V_t$  and  $T_{ox}$  on  $E_{min}$  and we showed that improving  $E_{min}$  results from a trade-off between variability and DIBL mitigation vs. reduction of the gate/subthreshold  $I_{gate}/I_{sub}$  leakage ratio. This new trade-off in nanometer technologies completely outweighs previously-reported load capacitance vs. subthreshold swing trade-off.

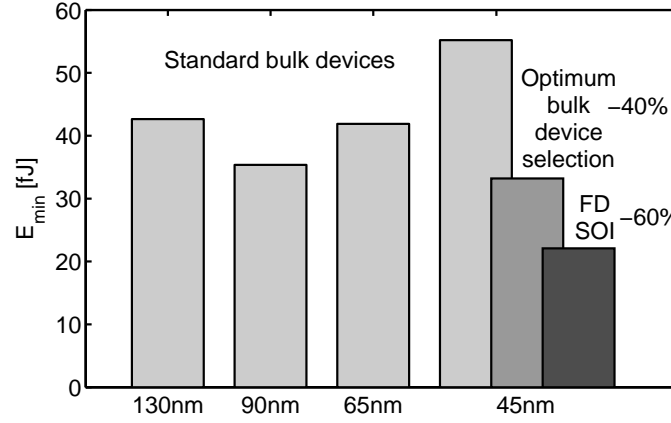
We also studied technology and device selection for  $E_{min}$  reduction. We showed that the energy saving brought by ITRS-recommended Low-Operating Power (LOP) or Low-Standby-Power (LSTP) technology flavor is only 15% as compared to standard High-Performance (HP) flavor at nominal gate length. Moreover, prohibitive delay prevents from using LSTP flavor under minimum-energy subthreshold operation. This makes technology-flavor selection inefficient for optimization of subthreshold circuits. However, we showed that selecting low- $V_t$  15/25 nm-longer devices in a 45 nm HP technology leads to 40% saving in minimum energy, without any process modification. As shown in Fig. 3.13, this optimum device selection allows reaching an  $E_{min}$  level at 45 nm node lower than at 90 nm node.

This study draws a new route for device optimization towards ultimate minimum-energy subthreshold circuits, indicating that efforts should be devoted to minimizing subthreshold swing, DIBL and variability, while tolerating gate leakage increase provided that it remains below the subthreshold leakage level.

We also explored the potential of FD SOI technology for minimum-energy subthreshold circuits, at minimum gate length. We showed that, despite higher DIBL effect due to fringing field in standard thick-BOX SOI technology, the subthreshold swing improvement coupled with the variability mitigation associated to lower channel doping yields up to 60%  $E_{min}$  reduction as compared to bulk. Moreover, as compared to the proposed optimum device selection in bulk

<sup>2</sup>Iso-delay:  $V_{dd} = 0.4 V$  and  $0.28 V$  for bulk and undoped-channel FD SOI, respectively.

<sup>3</sup>Iso-SNM:  $V_{dd} = 0.4 V$  and  $0.30 V$  for bulk and undoped-channel FD SOI, respectively.



**Fig. 3.13.** Evolution of  $E_{min}$  in general-purpose industrial technologies. At 45 nm node, the proposed optimum device selection yields 40% energy saving with slight die area overhead while undoped FD SOI technology reduces energy by 60% (bulk  $E_{min}$  levels come from simulation results in industrial technologies and FD SOI  $E_{min}$  level is calculated from normalization of the comparison results between bulk and FD SOI with pre-Silicon models).

technology, the use of FD SOI with minimum gate length has no delay penalty on superthreshold operation at nominal  $V_{dd}$ , which makes it even more attractive for DFVS circuits. Finally, the extremely-low variability of undoped-channel FD SOI would seriously benefit subthreshold SRAM by improving its stability. This makes FD SOI technology a strong candidate for extending the benefit of technology scaling for minimum-energy subthreshold circuits to nanometer technologies, as illustrated in Fig. 3.13.

## REFERENCES

1. S. Hanson, M. Seok, D. Sylvester and D. Blaauw, "Nanometer device scaling in subthreshold logic and SRAM", in *IEEE Trans. Electron Dev.*, vol. 55, no. 1, pp. 175-185, Jan. 2008.
2. J. T. Kao, M. Masayuki and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture", in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545-1554, Nov. 2002.
3. A. Wang, S. Kosonocky and A. P. Chandrakasan, "Optimal supply and threshold scaling for subthreshold CMOS circuits", in *Proc. IEEE Comp. Soc. Annual Symp. VLSI*, pp. 5-9, 2002.
4. B. Zhai, D. Blaauw, D. Sylvester and K. Flautner, "The limit of dynamic voltage scaling and insomnia dynamic voltage scaling," in *IEEE Trans. VLSI Syst.*, vol. 13, no. 11, pp. 1239-1252, Nov. 2005.
5. B. H. Calhoun and A. P. Chandrakasan, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering," in *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 238-245, Jan. 2006.
6. B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge and D. Sylvester, "Energy efficient near-threshold chip multi-processing", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 32-37, 2007.
7. V. Sze and A. P. Chandrakasan, "A 0.4-V UWB baseband processor", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 262-267, 2007.
8. H. Kaul, M. Anders, S. Mathew, S. Hsu, A. Agarwal, R. Krshnamurthy and S. Borkar, "A 320mV 65 $\mu$ W 411GOPS/Watt ultra-low voltage motion estimator accelerator in 65 nm CMOS", in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 316-317, 2008.
9. Semiconductor Industry Association, "Process, integration, devices and structures", in *The International Technology Roadmap for Semiconductors 2007*, Tech. Rep., Semiconductor Industry Association, Dec. 2007.
10. B. H. Calhoun, A. Wang and A. P. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits", in *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1778-1786, Sep. 2005.
11. T.-H. Kim, J. Keane, H. Eom and C. H. Kim, "Utilizing reverse short-channel effect for optimal subthreshold circuit design", in *IEEE Trans. VLSI Syst.*, vol. 15, no. 7, pp. 821-829, Jul. 2007.
12. S. Hanson, B. Zhai, D. Blaauw and D. Sylvester, "Energy optimality and variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 363-365, 2006.
13. J.-J. Kim and K. Roy, "Double-gate MOSFET subthreshold circuit for ultralow power applications", in *IEEE Trans. Electron Dev.*, vol. 51, no. 9, pp. 1468-1474, Sep. 2004.
14. B. C. Paul, A. Bansal and K. Roy, "Underlap DGMOS for digital subthreshold operation", in *IEEE Trans. Electron Dev.*, vol. 53, no. 4, pp. 910-913, Apr. 2006.

15. A. Raychowdhury, X. Fong, Q. Chen and K. Roy, "Analysis of super cut-off transistors for ultra-low power digital circuits", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 2-7, 2006.
16. B. C. Paul, A. Raychowdhury and K. Roy, "Device optimization for digital subthreshold logic operation", in *IEEE Trans. Electron Dev.*, vol. 52, no. 2, pp. 237-247, Feb. 2005.
17. B. C. Paul and K. Roy, "Oxide thickness optimization for digital subthreshold operation", in *IEEE Trans. Electron Dev.*, vol. 55, no. 2, pp. 685-688, Feb. 2008.
18. M. V. Dunga *et al.*, "BSIM4.6.1 MOSFET model", available on-line at [www-device.eecs.berkeley.edu/bsim3/bsim4.html](http://www-device.eecs.berkeley.edu/bsim3/bsim4.html).
19. W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration", in *IEEE Trans. Electron Dev.*, vol. 53, no. 11, pp. 2816-2823, Nov. 2006.
20. T. Skotnicki *et al.*, "Innovative materials, devices and CMOS technologies for low-power mobile multimedia", in *IEEE Trans. Electron Dev.*, vol. 55, no. 1, pp. 96-130, Jan. 2008.
21. A. Asenov, A. R. Brown, J. H. Davies, S. Kaya and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decanometer and nanometer-scale MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.
22. Y. Cao and L. T. Seok, "Mapping statistical process variations toward circuit performance variability: an analytical modeling approach", in *Proc. ACM/IEEE Des. Autom. Conf.*, pp. 658-663, 2005.
23. H. Fuketa, M. Hashimoto, Y. Mitsuyama and T. Onoye, "Correlation verification between transistor variability model with body biasing and ring oscillation frequency in 90nm subthreshold circuits", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 3-8, 2008.
24. N. R. Mohapatra, M. P. Desai, S. G. Narendra and V. Ramgopal Rao, "Modeling of parasitic capacitances in deep submicrometer conventional and high- $\kappa$  dielectric MOS transistors", in *IEEE Trans. Electron Dev.*, vol. 50, no. 4, pp. 959-966, Apr. 2003.
25. B. Zhai, S. Hanson, D. Blauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 20-25, 2005.
26. J. P. Colinge, "The SOI MOSFET", in *Silicon-on-Insulator Technology: Materials to VLSI*, 3<sup>rd</sup> Ed., Springer, pp. 151-246, 2004.
27. C. Fenouillet-Beranger, "Fully-depleted SOI technology using high-K and single-metal gate for 32nm node LSTP applications featuring  $0.179\mu\text{m}^2$  6T-SRAM bit-cell", in *Proc. IEEE Int. Electron Dev. Meeting*, pp. 267-270, 2007.
28. T. Skotnicki *et al.*, "A user's guide to MASTAR 4", available at [www.itrs.net/Links/2007ITRS/LinkedFiles/PIDS/MASTAR5/Mastar5-ITRS\\_2007/instructions.pdf](http://www.itrs.net/Links/2007ITRS/LinkedFiles/PIDS/MASTAR5/Mastar5-ITRS_2007/instructions.pdf), 59 p., 2005.
29. T. Numata and S. Takagi, "Device design for subthreshold slope and threshold voltage control in sub-100-nm fully-depleted SOI MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 51, no. 12, pp. 2161-2167, Dec. 2004.

30. T. Mizuno, J. Okamura and A. Toriumi, "Experimental study of threshold voltage fluctuations using an 8k MOSFETS array", in *VLSI Symp. Tech. Dig.*, pp. 41-42, 1993.
31. G. Tsutsui, M. Saitoh, T. Nagumo and T. Hiramoto, "Impact of SOI thickness fluctuation on threshold voltage variation in ultra-thin body SOI MOSFETs", in *IEEE Trans. Nanotechnology*, vol. 40, no. 3, pp. 369-373, May 2005.
32. G. Roy, A. R. Brown, F. Adamu-Lema, S. Roy and A. Asenov, "Simulation study of individual and combined sources of intrinsic parameter fluctuations in conventional nano-MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 53, no. 12, pp. 3063-3070, Dec. 2006.
33. M. Pelgrom, A. Duinmaijer and A. Welbers, "Matching properties of MOS transistors", in *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433-1440, Oct. 1989.
34. O. Weber *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding", in *Dig. IEEE Int. Electron Dev. Meeting*, in press, 4 p., 2008.

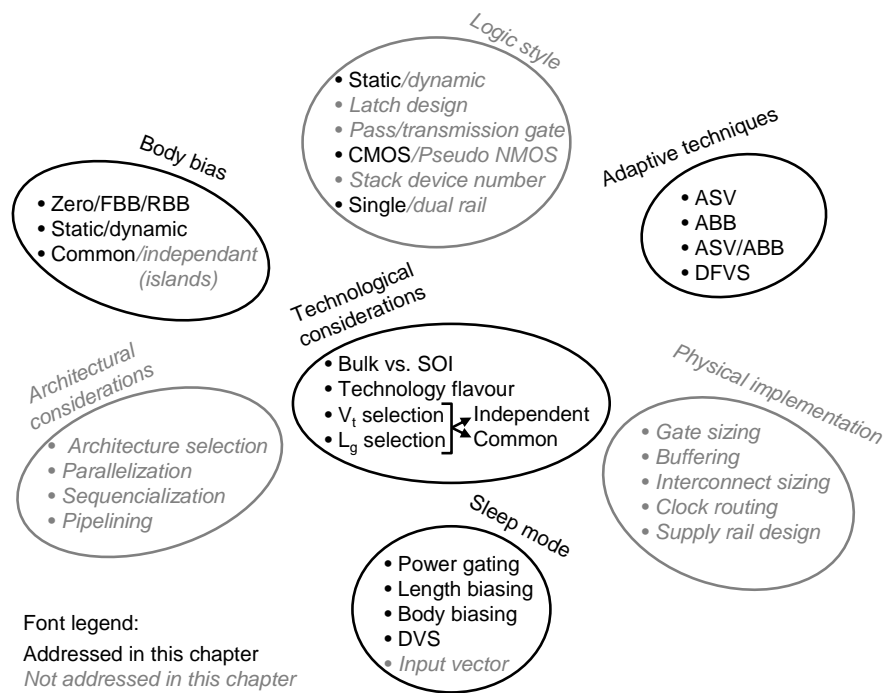


## CHAPTER 4

---

# CIRCUIT DESIGN CHOICES FOR PRACTICAL ENERGY MINIMIZATION IN NANOMETER SUBTHRESHOLD CIRCUITS

---



**Fig. 4.1.** Typical degrees of freedom in digital circuit design. The impact of these design choices on energy consumption may significantly change when porting them to subthreshold circuits in nanometer CMOS technologies, leading to different circuit design paradigms. *Let us examine it.*

## Abstract

---

As shown in Chapter 2, practical energy per operation under robustness and throughput constraints can be far higher than minimum energy. In this chapter, we revisit classical circuit design choices in the light of nanometer subthreshold digital circuits for ULP applications, the design target being to make practical energy reach the minimum energy level [CP6]. We show that fully-depleted SOI brings important practical-energy savings for the whole throughput range of ULP applications. We also demonstrate that the versatility of nanometer technologies is a powerful option to minimize practical energy, as it allows to shift minimum-energy point to different application throughputs. Nevertheless, we demonstrate that independent dual- $V_t$  assignment is inefficient in nanometer subthreshold circuits because of the large delay difference between std- and high- $V_t$  logic gates and the high variability of short paths.

We then show that adaptive reverse body biasing with negative voltage is an efficient technique to compensate for modeling errors or global process/temperature variations. It allows to limit design margins while keeping minimum-energy point at the target application throughput, under various operating conditions. On the contrary, forward body biasing suffer from increased minimum-energy level and bad behavior with discrete bias voltage values. Moreover at 45 nm node, we point out that reverse body biasing is only efficient in low-power technology flavor and we suggest that at next nodes it may no longer be practical because of decreasing body-bias coefficient and increasing band-to-band tunneling leakage.

Finally, we investigate the efficiency of sleep-mode techniques - dynamic reverse body biasing and power gating - for reducing active and stand-by leakage. For active-leakage reduction, sleep-mode techniques are less efficient than technology selection and static reverse body biasing, as they suffer from the energy overhead associated to mode transition. However, for reducing stand-by leakage, power gating is a very efficient technique in nanometer subthreshold circuits. Nevertheless, we showed that circuit robustness can be under risk when using badly-sized power switches and that engineering the power switch can bring significant energy reduction with lower robustness degradation.

## Contents

---

<b>4.1 Introduction</b>	95
<b>4.2 Technology selection</b>	96
<b>4.3 Body biasing for circuit adaptation</b>	105
<b>4.4 Sleep-mode techniques</b>	114
<b>4.5 Conclusion</b>	125

---

## 4.1 INTRODUCTION

In Chapter 2, we showed that dynamic and static energies per operation follow opposite evolutions with technology scaling into the nanometer era: dynamic energy is reduced thanks to capacitance reduction, whereas static energy increases from leakage currents, short-channel effects and variability. This leads to an optimum node for minimizing practical energy per operation of subthreshold circuits under robustness and throughput constraints. Practical energy for low-throughput applications is minimized at 0.18/0.13  $\mu\text{m}$  nodes, whereas practical energy for medium-throughput applications is minimized at 90/65 nm nodes and minimum-energy level increases significantly when reaching 45 nm node. Nevertheless, in practice, the technology choice is not only dictated by energy concern. Indeed, the die area and the underlying manufacturing cost are important criterion for choosing a technology node. Therefore, designing a subthreshold circuit in a 45 nm technology is a meaningful target.

In Chapter 3, we showed that optimum MOSFET selection within a versatile yet standard technology menu reduces the minimum-energy level by 40% and that fully-depleted SOI technology further improves this figure. Nevertheless, operating at minimum-energy point is not straightforward as we showed in Chapter 1 that it results from a perfect match between the application target throughput  $f_{op}$  and the optimum clock frequency  $f_{clk,opt}$  of minimum-energy point. In 45 nm technology, we showed in Chapter 2 (Fig. 2.16) that for the throughput range of ULP applications ( $\approx 10\text{ k-}10\text{ MOp/s}$ ), circuits mainly lie in robustness-limited energy-inefficient  $R3$  throughput region, i.e.  $f_{clk,opt} \gg f_{op}$ , with orders-of-magnitude static energy overhead. Seok *et al.* show in [1] that a static energy overhead is also added when the application features a low duty cycle, i.e. a low time ratio between active and stand-by periods of the application. To solve this energy issue, a subthreshold processor from Kwong *et al.* in [2] uses MTCMOS power gating, whereas another subthreshold processor from Hanson *et al.* in [3] uses reverse body biasing. As these circuits have different parameters and use different technologies, they are difficult to compare and it is thus not clear which technique is the most efficient one.

Fig. 4.1 lists some of the typical degrees of freedom in digital circuit design. All of them hold for ULP subthreshold circuits in nanometer technologies but their impact on energy consumption may significantly differ from the classical understanding designers have of them based on their experience in nominal- $V_{dd}$  high-performance/low-power design. In this chapter, we thus use the analysis framework of practical energy under robustness and throughput constraints proposed in Chapter 1, to revisit the relevant circuit design choices from Fig. 4.1 in the case of a subthreshold circuit in nanometer CMOS technologies. The target is to make practical energy meet minimum-energy level. Using an industrial 45 nm bulk technology, we carry out a systematic in-depth study of the efficiency of these design choices to help designers make the right decisions, at early design stages within numerous possibilities.

The organization of this chapter is based on a classification of the design degrees of freedom as presented in Fig. 4.1. In Section 4.2, we first consider the impact of technology selection at 45 nm node on energy consumption, as it is the first design choice. We then address body biasing for circuit adaptation in Section 4.3. Sleep-mode techniques are finally investigated in Section 4.4.

Finally, notice that throughout this chapter, we again address intrinsic variability in all simulations by statistical Monte-Carlo extraction of  $3\sigma$  worst-case SNM,  $3\sigma$  worst-case delay and mean leakage current.

## 4.2 TECHNOLOGY SELECTION

In Chapter 3, we showed that optimum MOSFET selection and fully-depleted (FD) SOI technology significantly improves minimum energy per operation, whereas technology flavor selection only brings minor improvement. In this section, we extend this investigation to practical energy under robustness and throughput constraints for ULP applications.

### 4.2.1 Bulk vs. FD SOI

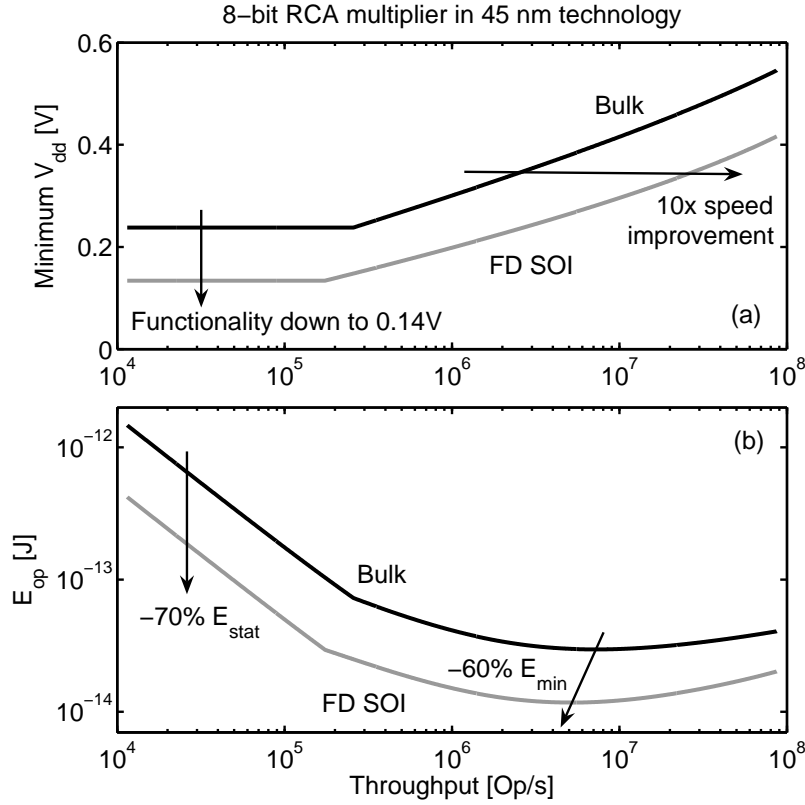
Let us first consider high-performance technologies as modeled in Chapter 3: standard bulk technology and undoped-channel FD SOI technology, both at 45 nm node. Simulated minimum  $V_{dd}$  for meeting 99.9% functional yield and delay constraint of the benchmark 8-bit RCA multiplier is plotted in Fig. 4.2 (a). Thanks to reduced subthreshold swing, DIBL and variability, FD SOI lowers the minimum functional  $V_{dd}$  down to 0.14V. Moreover, at iso- $V_{dd}$ , the reduced subthreshold swing, junction capacitances, variability and stack effect yields a  $10\times$  reduction in  $3\sigma$  worst-case delay and thus a  $10\times$  improvement in speed performances. It means that minimum  $V_{dd}$  for meeting the throughput constraint is lower in FD SOI.

These facts together result in an important reduction of practical energy per operation for the whole throughput range of ULP applications up to 70% in  $R3$  throughput region and 60% close to minimum-energy point. For the sake of generality, we consider an industrial 45 nm standard bulk technology in the rest of this chapter.

### 4.2.2 Technology flavor selection

The industrial 45 nm technology we consider features 2 technology flavors: a thin-oxide mid- $V_t$  short-channel flavor denominated as general-purpose (GP) and a low-power (LP) flavor with thick-oxide high- $V_t$  mid-channel devices. Both flavors are dual- $V_t$  technologies, i.e. with std- $V_t$  and high- $V_t$  devices. Main MOSFET parameters are given in Table 4.1.

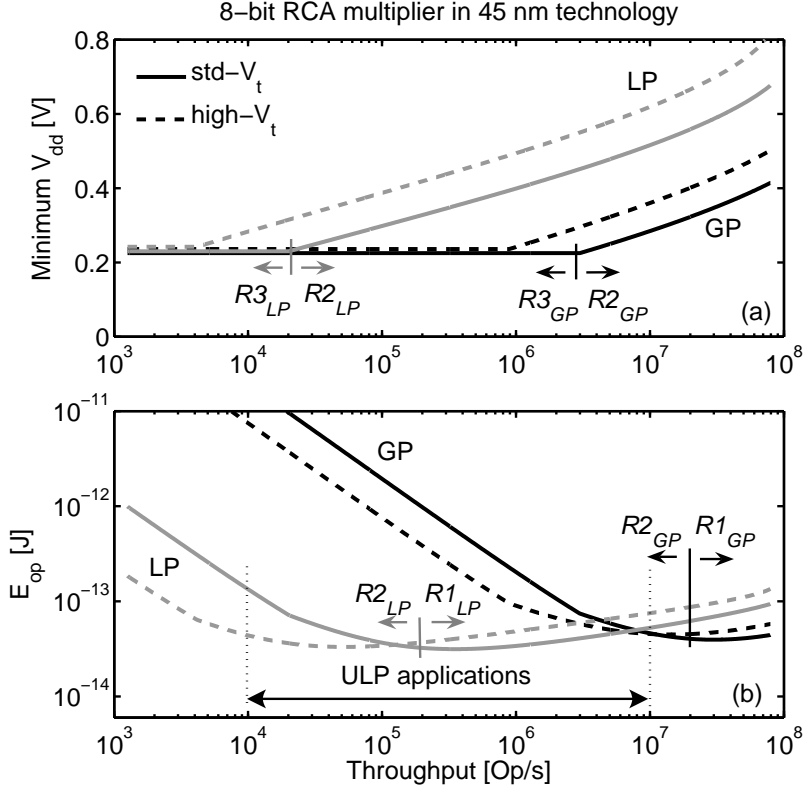
Simulated minimum  $V_{dd}$  for meeting 99.9% functional yield and delay constraint of the benchmark 8-bit RCA multiplier is plotted in Fig. 4.3 (a) for both



**Fig. 4.2.** Comparison of bulk and undoped-channel FD SOI technologies: (a) minimum  $V_{dd}$  under robustness and throughput constraints with (b) corresponding practical energy per operation (variability-aware Spice simulation of the 8-bit RCA benchmark multiplier in high-performance 45 nm technologies). FD SOI enables operation down to 0.14V and a 10 $\times$  speed improvement. In addition from the 60% saving in minimum energy reported in Chapter 3, FD SOI provides 70% static energy reduction in  $R3$  throughput region.

GP and LP flavors. As a consequence of its higher  $V_t$  and thus lower subthreshold reference current  $I_0$ , LP flavor features an increased delay at a given  $V_{dd}$  and thus requires a higher  $V_{dd}$  for meeting the throughput constraint. Minimum  $V_{dd}$  for meeting robustness constraint is comparable in GP and LP technologies. As a result, the boundary between speed-limited  $R2$  and robustness-limited  $R3$  throughput regions is shifted to lower throughputs for LP flavor.

Notice that in the considered technology, compact models of GP MOSFETs are BSIM4 models [4], whereas PSP models [5] are used for LP MOSFETs. Moreover, the variability model is somewhat different in GP and LP models. Therefore, special attention should be paid when comparing GP and LP flavors



**Fig. 4.3.** Comparison of GP and LP flavors: (a) minimum  $V_{dd}$  under robustness and throughput constraints with (b) corresponding practical energy per operation (variability-aware Spice simulation of the 8-bit RCA benchmark multiplier in industrial 45 nm bulk technology). The LP flavor shifts the practical energy curve to lower target application throughputs. Consequently, for the throughput range of ULP applications ( $\approx 10$  k–10 MOp/s), LP flavor features lower practical energy per operation.

in order to ensure that comparison results are significant. In this case, the difference in throughput-limited minimum  $V_{dd}$  is high and can thus be regarded as significant whereas the difference in robustness-limited minimum  $V_{dd}$  is too small to be regarded as significant.

Corresponding practical energy per operation  $E_{op}$  is plotted in Fig. 4.3 (b). The lower  $I_0$  of LP flavor implies a shift of the  $E_{op}$  curve to lower throughputs. Thanks to its lower minimum  $V_{dd}$  in R1 region, GP flavor features lower practical  $E_{op}$  for fixed throughputs higher than 8 MOp/s, where dynamic energy component dominates, as the switched capacitance is roughly equivalent between GP and LP flavors. Nevertheless,  $E_{op}$  in GP flavor dramatically increases for lower

**Table 4.1.** MOSFET parameters in the considered industrial 45 nm technology

Tech. flavor	Drawn $L_g$ [nm]	Printed $L_g^\dagger$ [nm]	$T_{ox}$ [nm]	$V_{dd,nom}$ [V]	$V_{t,nom}$ [V]
GP	40	35	1.2	0.9	0.41/0.55
LP	40	42	1.7	1.1	0.55/0.68

<sup>†</sup> Different printed  $L_g$  with identical drawn  $L_g$  are achieved by adaptation of the Poly mask during mask generation.

throughputs as the circuit deeply enters  $R2$  and  $R3$  regions dominated by static energy component. For throughputs below 8 MOp/s, LP flavor thus features the lowest  $E_{op}$  thanks to low  $I_0$ . This shows that there is an optimum technology flavor depending on the target throughput of the considered application. In the case of ULP applications, LP flavor features lower  $E_{op}$  for nearly the whole throughput range ( $\approx 10$  k-10 MOp/s), as a consequence, we will focus on LP flavor in next sections. The main effect of changing the technology flavor is thus a shift of the  $E_{op}$  curve vs. the application throughput. As the compact models are BSIM4 for GP MOSFETs and PSP for LP MOSFETs, the slight difference in minimum-energy levels between GP and LP flavors can come from the models so that it is not considered as significant.

#### 4.2.3 MOSFET selection

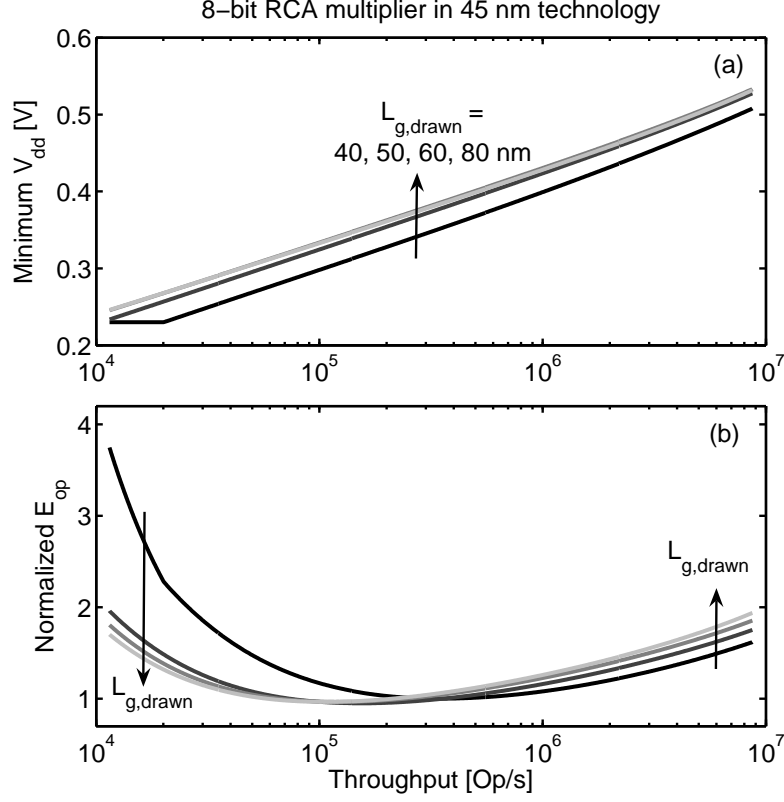
As both GP and LP flavors are dual- $V_t$  technologies, minimum  $V_{dd}$  and corresponding practical  $E_{op}$  have been simulated for high- $V_t$  devices and are plotted in Fig. 4.3 (dashed lines). It shows that  $V_t$  selection, common to all logic gates, also implies a translation, a shift of the  $E_{op}$  curve vs. throughput. This can be explained as follows. In throughput-limited  $R2$ - $R1$  regions, energy per operation at a given  $V_{dd}$  does not primarily depend on  $V_t$ :

$$E_{op} = \frac{1}{2} N_{sw} C_L V_{dd}^2 + L_D C_L 10^{\frac{-V_{dd}}{S}} V_{dd}^2, \quad (4.1)$$

from Eq. (3.1) and (3.3). However, the delay depends on  $V_t$  through  $I_0$  and consequently subthreshold  $I_{on}$ . Minimum  $V_{dd}$  for meeting a throughput constraint thus depends on  $V_t$  and changing  $V_t$  modifies minimum  $V_{dd}$  vs. throughput, which in turn shifts the  $E_{op}$  curve vs. throughput.

Fig. 4.3 shows that  $V_t$  selection offers a finer granularity in the shift of the  $E_{op}$  curve than technology flavor selection, which results in finer circuit tuning to make minimum-energy point correspond to the target throughput of the application, i.e. making  $f_{clk,opt}$  meet  $f_{op}$ . In next sections, we use std- $V_t$  devices in LP flavor unless otherwise specified, for the sake of generality.

As proposed in Chapter 3, an upsize of the device gate length  $L_g$  results in a reduction of minimum-energy level  $E_{min}$ . Fig. 4.4 shows the corresponding impact on practical energy under robustness and throughput constraints. Gate



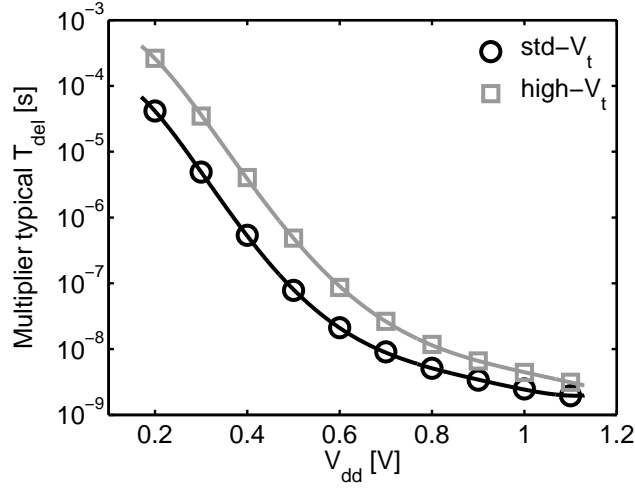
**Fig. 4.4.** Impact of gate length upsize on (a) minimum  $V_{dd}$  and (b) practical energy per operation in LP flavor (printed  $L_g$  is 2 nm longer than drawn  $L_g$ ). In LP flavor, the benefit of  $L_g$  upsize for  $E_{min}$  reduction is small. However, an upsized  $L_g$  benefits to practical energy at low application throughputs.

length upsize increases gate delay in the considered technology, which makes minimum  $V_{dd}$  for throughput constraint slightly rise. In LP flavor, gate length upsize does not improve minimum-energy level  $E_{min}$ , as short channel effects are less important than in GP flavor (e.g. DIBL  $\eta$  coefficient is 100 mV/V). However, the resulting moderate mitigation of subthreshold swing and DIBL is sufficient to save energy at low throughputs.

#### 4.2.4 Independent dual- $V_t$ assignment

In previous section, we showed that a global  $V_t$  selection allows to match  $f_{clk,opt}$  with the target throughput. It does not bring significant improvement in minimum-energy level because the leakage current  $I_{leak}$  you save by using



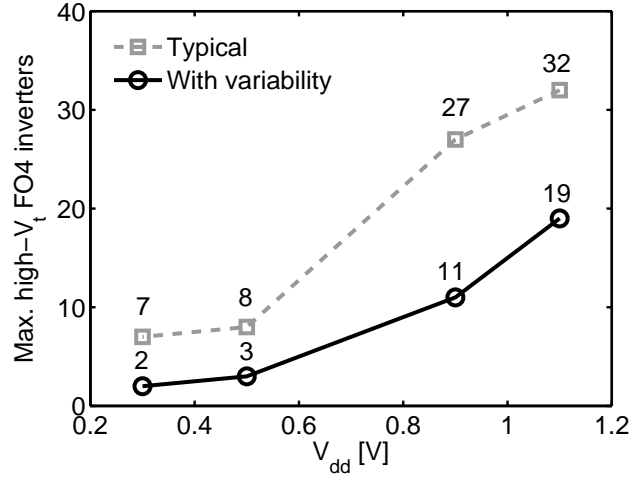


**Fig. 4.5.** Comparison of delay evolution vs  $V_{dd}$  in dual- $V_t$  technology (8-bit benchmark multiplier in industrial LP 45 nm technology, typical delay without variability)

high- $V_t$  devices is lost in terms of delay and thus execution time of the operation, thereby yielding minor modifications in static energy at a given  $V_{dd}$ . However, a dual- $V_t$  technology also allows to assign the threshold voltage of the devices to each logic gate independently. It has thus been proposed for high-speed low-power operation to use high- $V_t$  devices in non-critical paths to save leakage power, while keeping std- $V_t$  devices in the critical path to preserve speed performances [6]. Regarding FVS subthreshold circuits, this could reduce  $I_{leak}$  without raising minimum  $V_{dd}$  for throughput constraint and thereby reducing static energy at a given  $V_{dd}$ . Let us investigate the efficiency of this technique in nanometer subthreshold circuits.

Fig. 4.5 shows the evolution with  $V_{dd}$  of the typical delay without variability of the benchmark multiplier with std- $V_t$  and high- $V_t$  devices in the considered 45 nm LP technology. The first observation is that the delay difference becomes increasingly important when reaching the subthreshold region because of the exponential dependence of subthreshold  $I_{on}$  on  $V_t$  through  $I_0$  parameter. At nominal 1.1V  $V_{dd}$ , the high- $V_t$  multiplier is 60% slower whereas at subthreshold 0.3V  $V_{dd}$ , it is  $7\times$  slower. This will clearly limit the use of independent dual- $V_t$  assignment technique as high- $V_t$  devices can only be assigned to paths with a logic depth far smaller than the critical path.

In order to quantify this observation, we consider the following case: a subthreshold circuit with the 8-bit RCA benchmark multiplier as a component. Std- $V_t$  devices are assigned to all logic gates, by default. As the multiplier features a large logic depth (23 complex gates), it is likely that it will contain the critical path of the circuit. Let us see how short a non-critical path should be to

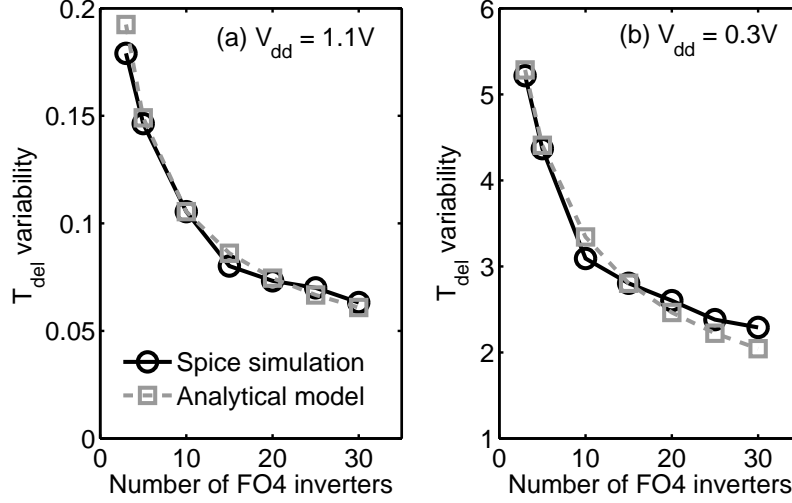


**Fig. 4.6.** Maximum number of high- $V_t$  FO4 inverters in a non-critical path for keeping its delay lower than the critical path from a std- $V_t$  8-bit multiplier (45 nm LP technology). Both variability and  $V_{dd}$  scaling to subthreshold regime reduces this number, thereby canceling the potential of independent dual- $V_t$  assignment for nanometer subthreshold circuits.

assign it high- $V_t$  devices without making it become critical, i.e. without raising its delay above the multiplier delay. Therefore, we extract the delay of a high- $V_t$  FO4 inverter from simulation of a 10-stage inverter chain and compare this delay with the std- $V_t$  multiplier delay under various supply voltages from 1.1V down to 0.3V. As the delay of an  $N$ -stage inverter chain is directly proportional to  $N$  (verified by simulation), we can easily predict how many high- $V_t$  FO4 inverters will result in a delay higher than the multiplier critical path. The result is shown in Fig. 4.6, first without variability, i.e. when considering typical delay for both the multiplier and the FO4 inverter (dashed line). At nominal 1.1V  $V_{dd}$ , a chain of 32 FO4 inverters has a delay just below the multiplier delay. However, when scaling  $V_{dd}$  down to 0.3V, the maximum number of inverters in the path for high- $V_t$  assignment is reduced to 7. This shows that high- $V_t$  devices can only be assigned to very short paths in subthreshold circuits.

Moreover, short paths feature higher variability due to random intrinsic device variations such as random dopant fluctuations (RDF) because variability is less averaged between logic gates in short paths. As these purely random variability components cannot be compensated, it may further decrease the maximum number of logic gates in high- $V_t$  paths.

In order to quantify that, we need a model of variability vs. the logic depth i.e. the number of stages  $N$  in the inverter chain. The delay of the path is the sum of the delay of all its logic gates, which are proportional to  $C_L V_{dd} / I_{on}$ . At nominal  $V_{dd}$ ,  $I_{on}$  and thus  $T_{del}$  can be modeled by a normal distribution



**Fig. 4.7.** Impact of logic depth on delay variability: (a) at 1.1V normalized difference between  $3\sigma$  worst-case and typical delays ( $3\sigma/\mu$  of normal distribution) and (b) at 0.3V ratio between  $3\sigma$  worst-case and typical delays ( $\sigma^3$  of lognormal distribution) in 45 nm LP technology (high- $V_t$  devices). This shows a good match between analytical models ( $1/\sqrt{N}$  at 1.1V and Eq. (4.2) at 0.3V) and simulation results.

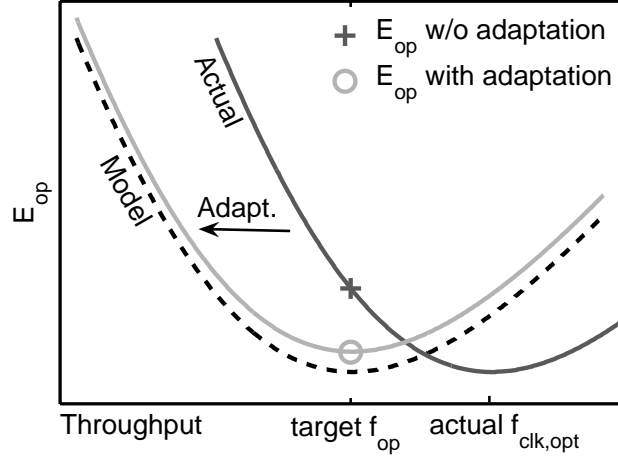
because of their alpha-power dependence on normally-distributed  $V_t$ . The sum of normally-distributed  $T_{del}$  is another normal distribution with the standard deviation averaged in  $1/\sqrt{N}$  [7]. As shown in Fig. 4.7 (a), there is a very good agreement between this  $1/\sqrt{N}$  law and Spice simulations of  $N$ -stage FO4 inverter chain.

In subthreshold regime,  $I_{on}$  is a subthreshold current and its distribution thus follows a lognormal law because of its exponential dependence on normally-distributed  $V_t$ . In [8], the delay variability of a path of subthreshold logic gates is shown to follow this law:

$$\sigma(\ln T_{del}) = \sqrt{\ln \left( 1 + \frac{1}{N} (10^{t^2} - 1) \right)}, \quad (4.2)$$

where  $t = \sigma_{V_t}/S$ . As shown in Fig. 4.7 (b), Spice-simulated variability of the delay closely matches this law.

From these variability models, we predict the  $3\sigma$  worst-case delay of an  $N$ -stage FO4 high- $V_t$  inverter chain and compares it to the  $3\sigma$  worst-case delay of the multiplier at various supply voltages. Fig. 4.6 shows the calculated maximum number of FO4 inverters in a high  $V_t$  path to keep its worst-case delay below the critical path worst-case delay. The impact of variability on nominal  $V_{dd}$  operation is important: it reduces the maximum logic depth of high  $V_t$  paths to 19 FO4



**Fig. 4.8.** Impact of modeling errors: a mismatch between the target  $f_{op}$  throughput and the actual  $f_{clk,opt}$  of minimum-energy point systematically results in energy overhead. This needs to be compensated by circuit adaptation even if the adaptive technique implies a small  $E_{min}$  overhead.

inverters and the impact on subthreshold operation is disastrous: high- $V_t$  devices can only be assigned to paths with 2 FO4 inverters. This clearly demonstrates the inefficiency of individual dual- $V_t$  assignment in nanometer subthreshold circuits.

Finally, notice that similar techniques rely on independent gate length up-size to mitigate subthreshold circuits in non-critical path [9], also called length biasing, and independent dual- $T_{ox}$  assignment for gate leakage mitigation. Nevertheless, a look at Table 4.1 shows that a different  $T_{ox}$  assignment in industrial technology comes with a strong  $V_t$  difference and thus a larger subthreshold delay difference between thin-oxide GP and thick oxide LP logic gates, as shown by the large shift in minimum  $V_{dd}$  for throughput constraint between GP and LP flavors from Fig. 4.3. Moreover, manufacturability issues in nanometer technologies favour regular layout, which often prevents from assigning different gate lengths to adjacent gates. For these reasons, independent dual- $T_{ox}$  and  $L_g$  assignment are likely to be inefficient in nanometer subthreshold circuits, similarly to independent dual- $V_t$  assignment.

#### 4.2.5 Discussion

As shown in Fig. 4.3 (b) by taking the minimum of the four  $E_{op}$  curves, the versatility of the technology provides a high potential for minimizing  $E_{op}$  for a wide throughput range by appropriate flavor and  $V_t$  selection. Correct technology selection thus enables making practical energy reach the minimum-energy level at the target throughput by matching  $f_{op}$  and  $f_{clk,opt}$ .

Nevertheless, this beautiful picture is quite theoretical. In practice, simulating the whole system to determine which technology option enables minimum-energy operation at the target throughput is a complex task. Indeed, an error in circuit modeling or a global MOSFET parameter deviation from extrinsic process variability may result in a shift of the actual  $E_{op}$  curve vs. throughput. As shown in Fig. 4.8, this systematically results in energy overhead as the target throughput dictated by the application cannot be adapted. There is thus a strong need for adaptive techniques to shift back minimum-energy point to the target throughput, even at the cost of a small  $E_{min}$  penalty. This is addressed in next section.

### 4.3 BODY BIASING FOR CIRCUIT ADAPTATION

Traditionally, digital circuits are designed with the target to meet the timing constraints at the worst-case process, voltage and temperature (PVT) corner. This requires the introduction of design margins, which implies energy overhead. Throughout this dissertation, we have considered intrinsic random variability by taking a design margin on  $V_{dd}$  to ensure 99.9% timing yield. It means that we have considered the Monte-Carlo statistically-extracted  $3\sigma$  worst-case delay to derive the minimum  $V_{dd}$  to meet the throughput constraint. As intrinsic variability is not spatially-correlated, it can hardly be compensated, so that design margins are the only way to deal with it.

Up to this point, we did not consider global  $V_t$  variations due to extrinsic variability nor global temperature variations. As these variations are global, they can be compensated by adaptive techniques in order to avoid taking design margins to handle worst-case conditions. Main techniques for adaptation of subthreshold circuits are adaptive supply voltage (ASV) and adaptive body biasing (ABB) [10, 11].

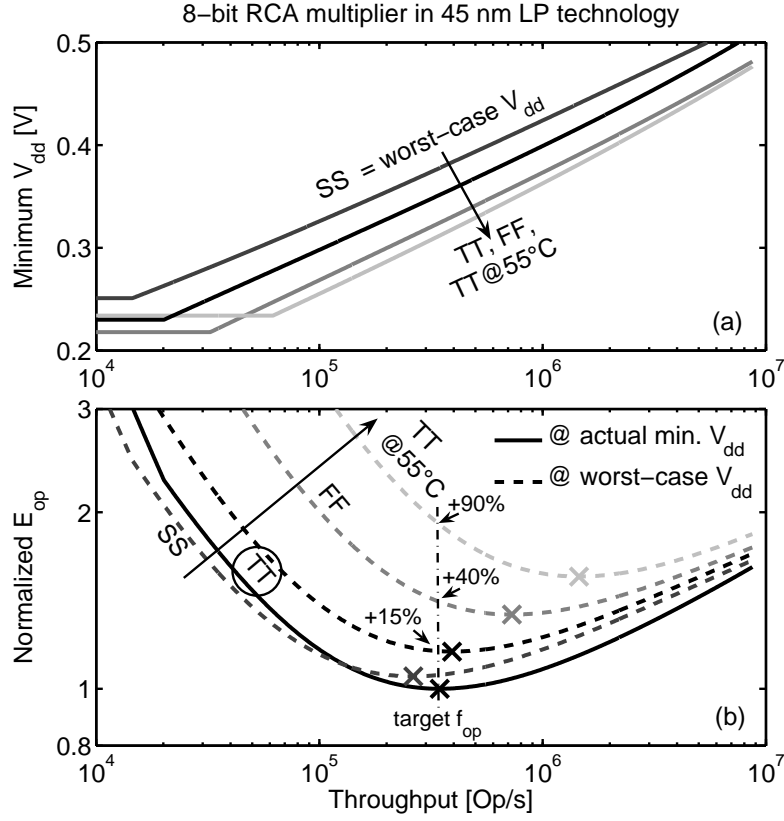
In this section, we first quantify the impact of global process/temperature variations on practical energy. We then analyze the effect of body bias on subthreshold MOSFET operation and the impact on practical energy in nanometer subthreshold circuits. We finally compare the use of ASV and ABB for circuit adaptation.

#### 4.3.1 Impact of global process/temperature corners

There are several phenomenons that may require circuit adaptation to avoid design margins:

- extrinsic global process variations;
- global temperature variations;
- variations of circuit performance with lifetime due to device aging phenomenons (hot-carrier effects, negative-bias temperature instability, etc.).

Notice that, for the sake of simplicity, we do not consider  $V_{dd}$  variations in this dissertation.



**Fig. 4.9.** Impact of global temperature and extrinsic process variations on (a) minimum  $V_{dd}$  and (b) practical energy per operation in 45 nm LP technology. The operating temperature is 25°C, unless otherwise specified. Solid lines represent the operation at the actual minimum  $V_{dd}$  i.e. without design margin while solid lines represent operation at the worst-case  $V_{dd}$  of SS corner.

As subthreshold drain current exponentially depends on  $V_t$ , the highest dependence of circuit performance on these three phenomena comes from their effect on  $V_t$  and consequently on subthreshold reference current  $I_0$ . The phenomena that increase  $V_t$ , such as an SS (slow NMOS, slow PMOS) process corner, a temperature lowering or negative-bias instability in PMOS devices, increase circuit delay and reduce leakages. Circuit adaptation is needed to compensate for the delay increase in order to avoid timing violations. The phenomena that reduce  $V_t$ , such as FF (fast NMOS, fast PMOS) process corner or a temperature rise, reduce delay and increases leakages. Adaptation is needed to mitigate the corresponding static energy overhead by reducing the positive time slack due to in-advance completion of the operation. This is illustrated in Fig. 4.9 (a) where

**Table 4.2.** Effects of reverse body biasing on MOSFET subthreshold operation

Tech. node	Tech. flavor	$\gamma$ value [mV/V]	$I_0$ reduction [ $\times/V$ ]	$S$ [–]	$\eta$ [–]	$C_{g,sub}$ [–]
130	GP	150	125	reduction	-	reduction
45	GP	85	14	-	increase	reduction
45	LP	120	50	reduction	increase	reduction

minimum  $V_{dd}$  for meeting both the robustness and throughput constraints is plotted for SS, TT (typical NMOS, typical PMOS) and FF process corners at 25°C and for TT corner at 55°C. Minimum  $V_{dd}$  is the highest for SS corner at 25°, which is thus the worst-case corner.

Fig. 4.9 (b) shows the corresponding energy consumption for the same process/temperature corners. Practical  $E_{op}$  is plotted in dashed lines when considering design margins, i.e. when operating at the worst-case  $V_{dd}$  (minimum  $V_{dd}$  of SS corner). For comparison purpose, practical  $E_{op}$  for TT corner at 25° is also plotted in solid line without taking design margins i.e. when considering the actual minimum  $V_{dd}$  of this corner. This shows that process/temperature variations shift the minimum-energy point to different applications throughputs. Moreover, design margins introduce large energy overheads. Let us assume that we were able to select a technology flavor and a  $V_t$  value that makes minimum-energy point meet the target throughput i.e.  $f_{op} = f_{clk,opt}$ . Under this condition, the design margin introduces 15 and 40% energy overheads due to increased leakage for TT and FF corners at 25°C, respectively. Furthermore, a temperature rise to 55°C implies an energy overhead up to 90% at TT corner. This clearly shows the need for circuit adaptation in order to keep minimum energy per operation.

#### 4.3.2 Effects of body bias on subthreshold MOSFET operation

From Eq. (1.9), BB modifies the threshold voltage through  $-\gamma V_{bs}$  term, where  $\gamma$  is the linearized body-effect coefficient. This is a deterministic phenomenon even in nanometer technologies, as experimentally verified in [12]. Table 4.2 shows the  $\gamma$  values (at  $V_{ds} = 0.3V$ ) in 45 nm GP and LP technologies, compared with 0.13  $\mu m$  technology. These values are computed from Spice extraction of  $V_t$  values between  $V_{BB}=0$  and  $-0.6V$ . Notice that these are  $\gamma$  values of minimum- $L_g$  devices, which are significantly lower than the theoretical long-channel  $\gamma$  expression of  $C_{dep}/C_{ox}$ .

Table 4.2 shows that the effect of BB on  $V_t$  is reduced with technology scaling, as reported in [13]. A 1V  $V_{BB}$  shifts  $V_t$  by 150 mV in 0.13  $\mu m$  technology and only by 85 mV in 45 nm GP technology. LP flavors features intermediate  $\gamma$  value due to its thicker  $T_{ox}$  and longer  $L_g$ .

The resulting reduction of subthreshold reference current  $I_0$  is also given in Table 4.2, computed from  $I_0$  Spice extraction between  $V_{BB}=0$  and  $-0.6V$ . It exponentially follows the evolution of  $\gamma$  and it is thus weak in 45 nm GP technol-

ogy. Qualitatively speaking, on one hand, reverse body biasing (RBB) improves  $S$  and reduces  $C_{g,sub}$  thanks to an increase of the depletion depth  $X_{dep}$  and thus a reduction of  $C_{dep}$ . On the other hand, RBB increases the DIBL effect ( $\eta$  coefficient) for the same reason. Forward body biasing (FBB) have the opposite effect, refer to Chapter 2 for models of  $S$ ,  $C_{g,sub}$  and  $\eta$  parameters. The weak effect of BB on DIBL in  $0.13\ \mu\text{m}$  technology comes from the low original  $\eta$  value in this technology. Notice that in  $45\text{ nm}$  GP technology,  $S$  weakly depends on BB because of two concurrent effects: an increase of  $X_{dep}$  from RBB worsens improves the long-channel  $S$ , whereas it worsens short-channel degradation of  $S$  (see Eq. (3.6) for instance). As short-channel effects are higher in GP than in LP flavor because of shorter  $L_g$ , the second effect is more important and counteracts the first one. Furthermore, notice that when considering pure high-performance flavor with very-short  $L_g$ , we observed in [CP6] that  $S$  can even be degraded by RBB.

### 4.3.3 Impact of body bias on practical energy

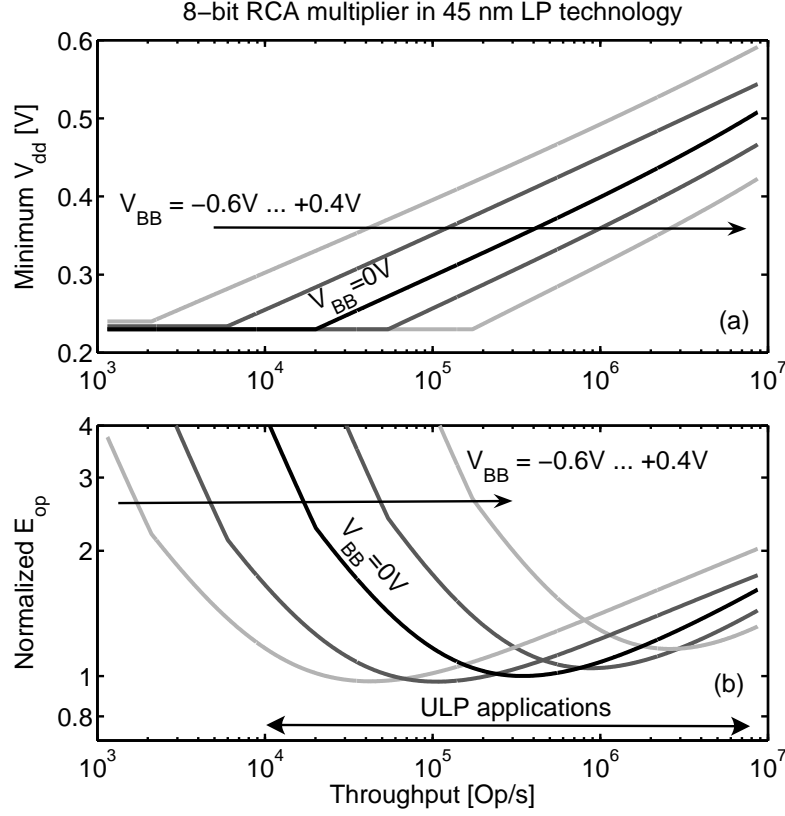
These effects of BB on MOSFET subthreshold operation in turn impact practical energy of subthreshold circuits. The most direct impact of RBB for instance comes from the  $I_0$  reduction, which results in longer delay and lower subthreshold  $I_{leak}$ . As shown in Fig. 4.10, the delay modification results in a shift of the minimum  $V_{dd}$  for meeting the throughput constraint to higher (resp. lower)  $V_{dd}$  values for reverse (resp. forward) body bias, corresponding to the underlying  $V_t$  modification, similarly to  $V_t$  selection as explained in Section 4.2.3. With a  $1\text{ V}$  BB range ( $-0.6$  to  $0.4\text{ V}$ ), practical energy for the whole throughput range of ULP applications ( $10\text{ k}$  to  $10\text{ MOp/s}$ ) can be kept below  $1.4 \times E_{min}$ , with an underlying  $V_{dd}$  adaptation between  $0.3$  and  $0.42\text{ V}$ .

This shows that BB is a powerful technique for circuit adaptation to shift minimum-energy point by fine tuning of  $f_{clk,opt}$ , as represented in Fig. 4.11 (a), with an  $f_{clk,opt}$  shift of  $60 \times /V$ . Interestingly, notice in Fig. 4.11 (b) that the optimum  $V_{dd,opt}$  (dashed line) of minimum-energy point is quite insensitive to the body voltage  $V_{BB}$ , provided that the drain-to-substrate junction is fully turned off ( $V_{BB} < 0.4\text{ V}$ ). It suggests that adaptation can be performed through ABB only i.e. with fixed  $V_{dd}$ .

We also observe in Fig. 4.10 (b) a modification of  $E_{min}$  with  $V_{BB}$ . Fig. 4.11 (b) emphasizes this fact (solid line). Recall from Chapter 3 that an increase in  $S$ , DIBL,  $C_L$  and  $I_{gate}/I_{sub}$  ratio directly affects  $E_{min}$ . The application of an RBB first reduces  $E_{min}$  thanks to reduction of  $S$  and  $C_L$  through  $C_{g,sub}$ . However, above a certain negative  $V_{BB}$ ,  $E_{min}$  is deteriorated because DIBL increases and  $I_{sub}$  becomes lower than  $I_{gate}$  level. FBB increases  $E_{min}$  first by degradation of  $S$  and increase of  $C_L$  and then by dramatic increase of the junction leakage  $I_{junc}$ , which dominates  $I_{leak}$  similarly to  $I_{gate}$  with high RBB. Adaptation through FBB thus comes at the cost of  $E_{min}$  penalty and should be avoided.

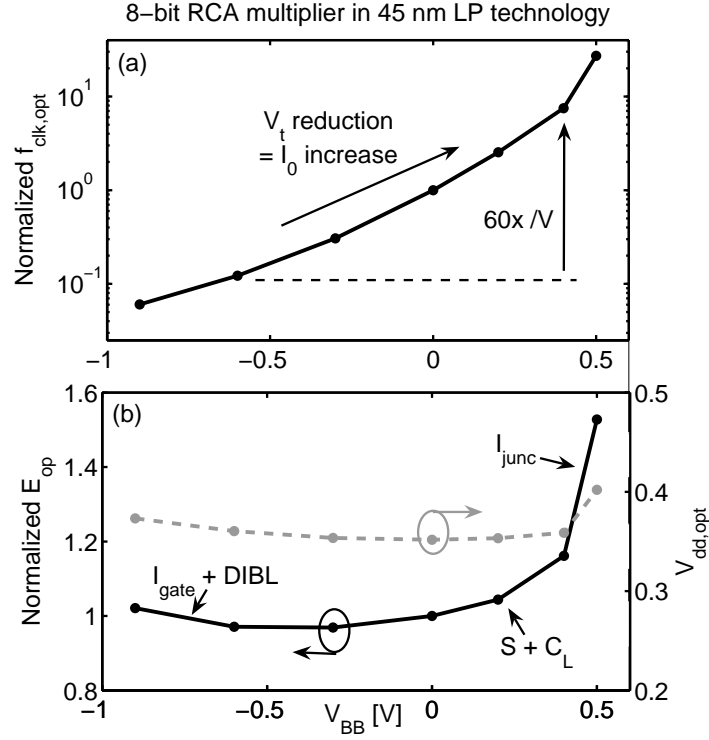
Let us compare the RBB impact in the different technologies. Table 4.3 shows the efficiency of RBB to reduce the multiplier  $I_{leak}$  under  $0.3\text{ V}$   $V_{dd}$ , which is





**Fig. 4.10.** Impact of body bias on (a) minimum  $V_{dd}$  and (b) practical energy per operation in 45 nm LP technology ( $V_{BB} = -0.6, -0.3, 0, 0.2, 0.4V$ ). Body biasing shifts the  $E_{op}$  curve vs. throughput and also modifies the minimum-energy level.

dramatically deteriorated in 45 nm technology, even in LP flavor, as compared to 0.13  $\mu m$  technology. The first reason of this deterioration is the reduction of body-effect coefficient  $\gamma$  in nanometer technologies and the second is the increase of DIBL effect with RBB. The underlying detrimental RBB impact on multiplier delay under 0.3V  $V_{dd}$  is also given in Table 4.3. It shows that this effect is also highly attenuated in 45 nm GP flavor, whereas it is increased in LP flavor. This comes from the very high  $V_t$  values of the LP flavor. Indeed, the circuit in LP flavor under 0.3V  $V_{dd}$  is deep in the subthreshold region with a full exponential dependence on  $I_{on}$  and thus  $T_{del}$  on  $V_t$ , whereas in GP flavor the circuit is closer to the subthreshold region boundary at 0.3V, with a weaker dependence of  $I_{on}$  and  $T_{del}$  on  $V_t$ . Consequently, although Fig. 4.10 shows that RBB is an efficient adaptive technique to make the actual  $f_{clk,opt}$  meet the target  $f_{op}$  in 45 nm LP technology, it is less efficient for GP technology at 45 nm node. It means that a



**Fig. 4.11.** Impact of body-bias on minimum-energy point in 45 nm LP technology: (a) optimum clock frequency of minimum-energy point and (b) minimum-energy level with corresponding optimum  $V_{dd}$ .

**Table 4.3.** Impact of reverse body biasing on subthreshold circuit

Tech. node	Tech. flavor	$I_{leak}$ reduction [ $\times/V$ ]	$T_{del}$ increase [ $\times/V$ ]
130	GP	140	39
45	GP	6.7	6.1
45	LP	45	52

larger  $V_{BB}$  voltage has to be on-chip generated or externally supplied in order to compensate identical variations. Similarly, it may no longer be the case for LP technology at next nodes, as the RBB efficiency degrades with technology scaling because of the reduction of the body-effect coefficient  $\gamma$ .

#### 4.3.4 Circuit adaptation

As shown in Fig. 4.3, once the technology flavor and the MOSFET devices are selected, minimum-energy operation can only be reached for one particular throughput. Adaptation is required when the target throughput  $f_{op}$  does not match the optimum clock frequency  $f_{clk,opt}$  of minimum-energy point. This can occur in two cases.

**Case 1 - compensation of deviations from model:** the actual  $E_{op}$  curve vs. throughput may be different from the simulated one because of modeling errors, global temperature variations, device aging or extrinsic process variations. In the case of a modeling error, post-Silicon compensation can be achieved at test time by application of static compensation parameters ( $V_{dd}$  for adaptive supply voltage scheme or  $V_{BB}$  for adaptive body biasing scheme) extracted from measurement of a few manufactured chips. In the case of process, temperature and aging variations, the compensation should preferably be done at run time by using a critical path replica with an adaptive supply voltage (ASV) scheme in a delay-locked loop [10, 14], which becomes increasingly frequent in both low-power and high-performance nanometer circuits.

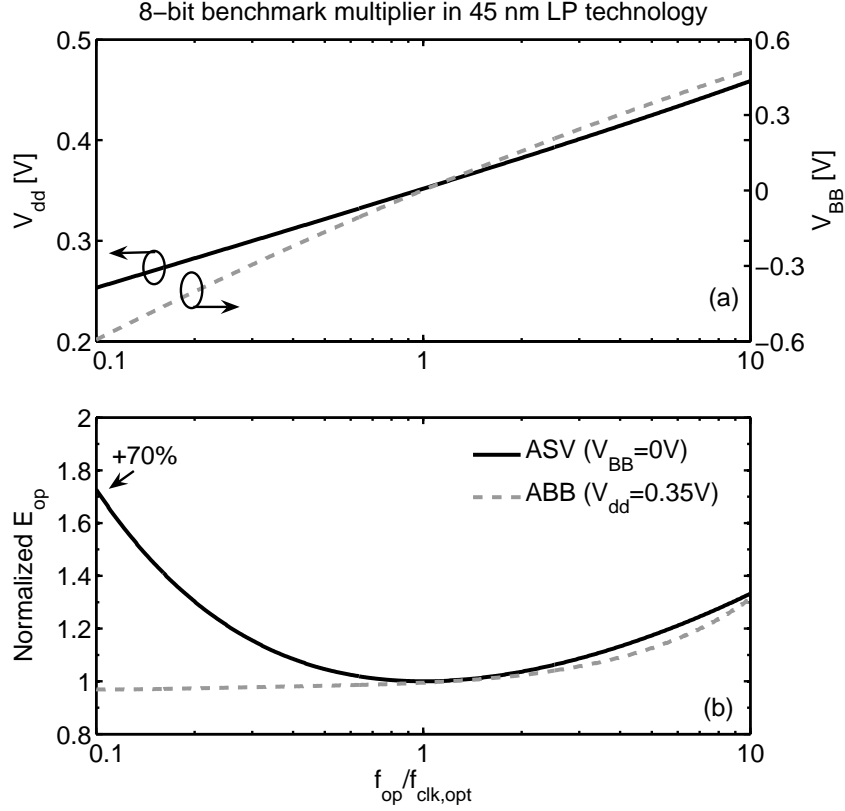
**Case 2 - adaptation to dynamic workload:** the target  $f_{op}$  may vary during run time if the circuit workload is dynamic. For an ULP circuit, this could correspond to a low-performance mode with 100 kOp/s and a mid-performance mode with 10 MOp/s for instance. In this case, run-time adaptation can be achieved with reconfiguration parameters ( $f_{clk} = f_{op}$  and  $V_{dd}/V_{BB}$ ) statically encoded in the power-state look-up table.

Next experiment addresses both cases by computation of the practical  $E_{op}$  for a wide throughput range centered on  $f_{clk,opt}$  with ASV and ABB schemes, which is similar to an actual  $f_{clk,opt}$  that differs from  $f_{op}$ .

First, Fig. 4.12 (a) shows the minimum  $V_{dd}$  for the 8-bit benchmark multiplier to support target application throughputs different from  $f_{clk,opt}$ . A 200mV  $V_{dd}$  range between 0.25 and 0.45V can accommodate two decades of  $f_{op}$  variations. Corresponding practical  $E_{op}$  in Fig. 4.12 (b) follows the usual evolution, being dominated by dynamic component for  $f_{op} > f_{clk,opt}$  and by static component for  $f_{op} < f_{clk,opt}$ . At  $f_{op} = 0.1 f_{clk,opt}$ , there is a 70% energy overhead with ASV technique.

Secondly, when considering a fixed 0.35V  $V_{dd}$  equal to the  $V_{dd}$  of minimum-energy point, minimum  $V_{BB}$  to support  $f_{op}$  is represented in Fig. 4.12 (a) and shows that two decades  $f_{op}$  variations can be accommodated by a 1.1V  $V_{BB}$  range between -0.6 and 0.5V. The impact of ABB on  $E_{op}$  shown in Fig. 4.12 (b) is comparable to ASV at high throughputs. However, at low throughputs, ABB is much better as it keeps  $E_{op}$  at the  $E_{min}$  level.

Regarding practical implementation, the compensation circuit to deliver the adaptive  $V_{dd}$  or  $V_{BB}$  can be based on digitally-controlled DC/DC converters and charge pumps or multiple supply sources [14]. In this case, only discrete voltage levels can be assigned to  $V_{dd}$  or  $V_{BB}$  [15]. We analyze the effect of voltage quantization in Fig. 4.13 by assuming 50 and 200mV steps in  $V_{dd}$  and  $V_{bb}$ , respectively.

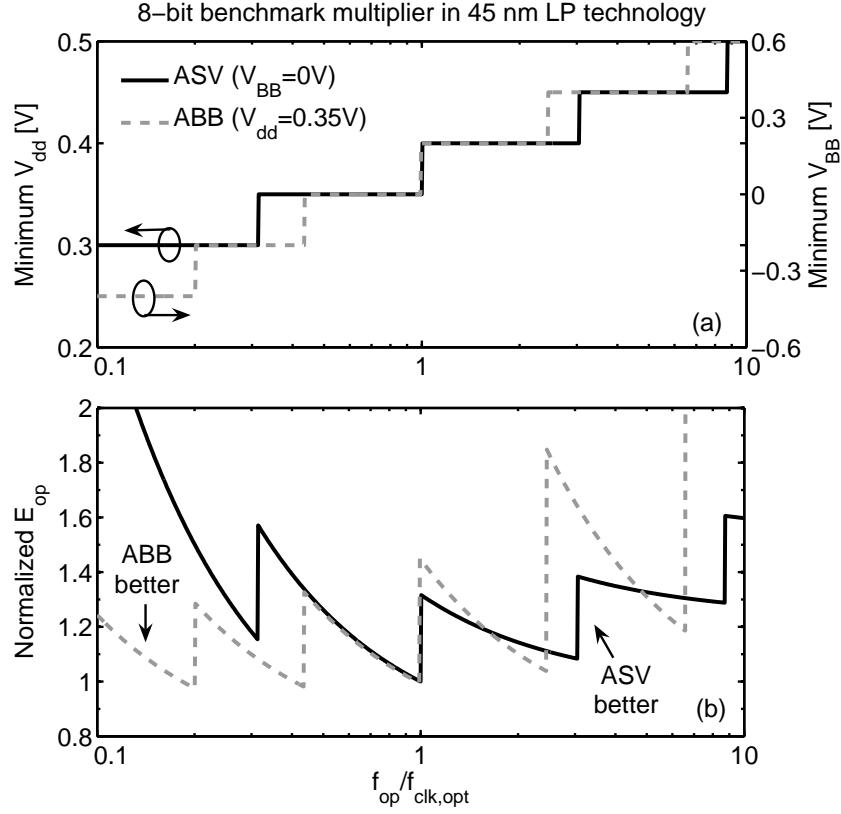


**Fig. 4.12.** Comparison of ASV and ABB techniques to compensate for a mismatch between target throughput and optimum clock frequency of minimum-energy point: (a) minimum  $V_{dd}$  and  $V_{BB}$  and (b) corresponding practical energy normalized to  $E_{min}$ . ABB brings more than 70% energy saving as compared to ASV at low throughputs.

The superiority of ABB over ASV stays at low throughputs. However, at high throughputs, the overhead of voltage quantization is much more pronounced with ABB scheme because of high sensitivity of junction leakage on  $V_{BB}$ . While ASV keeps the overhead below 40% for  $f_{op}$  up to  $8 \times f_{clk,opt}$ , ABB leads to energy overhead higher than 80% for a more narrow  $f_{op}$  range, only  $6 \times f_{clk,opt}$ .

#### 4.3.5 Discussion

In this section, we showed that circuit adaptation in nanometer subthreshold circuits is very useful to keep minimum-energy operation i.e. keep  $f_{clk,opt} = f_{op}$ , against global process/temperature variations, modeling errors or dynamic workload variations. The use of an ABB scheme is shown to be more efficient than



**Fig. 4.13.** Effect of discrete voltage levels on adaptive techniques: (a) minimum  $V_{dd}$  and  $V_{BB}$  and (b) corresponding practical energy normalized to  $E_{min}$ . ASV features lower dramatical quantization overhead at high throughputs.

ASV as confirmed by Hanson *et al.* in  $0.13\mu m$  in [3]. ABB is particularly effective for negative bias voltages (RBB). Indeed, forward body biasing suffers from two main drawbacks. First, it increases the minimum-energy level in 45 nm technology because of subthreshold swing degradation, subthreshold gate capacitance increase and, ultimately, high junction leakage. Secondly, when the bias voltage can only take discrete values, the energy overhead due to quantization can be higher than 100% with forward body biasing. This indicates that designers should better rely on ABB with only negative bias voltages. To do so, they should select the technology flavor and device type (std- or high- $V_t$ ) that brings  $f_{clk,opt}$  as close as possible to  $f_{op}$ , but by ensuring that  $f_{clk,opt}$  remains above  $f_{op}$  in any case. They should thus design at the worst-case process/temperature corner for speed concern (SS process at low temperature) and then use adaptive RBB to benefit from a potential positive time slack to reduce static energy. In

the case of a dynamically-varying workload, this means that designers should select the technology to enable operation at the minimum-energy point for the highest target throughput and rely on adaptive RBB when in low-throughput modes rather than the opposite solution.

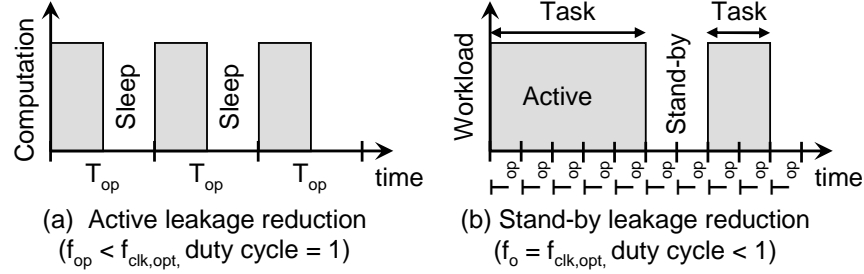
We also showed that the efficiency of body biasing exhibits a worrying evolution with technology scaling. The consequence is that in nanometer technologies, stronger reverse body-bias voltages should be used to achieve a constant leakage reduction. This observation coupled with the increase of global process variations in nanometer technologies can have detrimental implications in the design of ULP circuits. Indeed, even if in 45 nm LP technology flavor, a -0.6V body bias seems sufficient to mitigate high leakage-current deviations (here represented by a  $10\times$  shift in  $f_{clk,opt}$ ), it is not the case in 45 nm GP technology flavor and probably neither in LP flavor of next technology nodes. As a matter of fact, supplying +1V and -1V or higher bias voltages for RBB to a chip operating with 0.3V  $V_{dd}$  is a fanciful target, especially for low-cost volume-constrained applications such as RFID tags or biomedical devices, which cannot rely on multiple off-chip voltage sources. Moreover, although the low  $V_{dd}$  of minimum-energy circuits relaxes the electrical field across the junction and consequently band-to-band tunneling (BTBT) current, this leakage component becomes increasingly important in nanometer technologies with very high doping levels and abrupt doping profiles [16]. Associated with the increased negative  $V_{BB}$  to achieve subthreshold-leakage reduction in nanometer technologies, BTBT may dramatically increase and completely ruin the benefit brought by RBB.

Adaptive RBB could thus no longer be efficient and the interest of technologies with low sensitivity against process and temperature variations such as fully-depleted SOI will thus be even more pronounced in nanometer subthreshold circuits than in high-performance nominal- $V_{dd}$  circuits.

#### 4.4 SLEEP-MODE TECHNIQUES

Sleep-mode techniques to cut off leakage are massively adopted in high-performance/low-power circuits when migrating to leaky nanometer technologies. The target is to place the circuit in sleep mode with low leakage current when no operation is required, i.e. when the application is in stand-by state [17]. Moreover, in DFVS systems for minimum-energy operation it has recently been proposed to use sleep-mode techniques to reduce leakage in active mode if the operation is completed ahead of timing deadline [18]. Let us formerly express these two situations within the framework of practical energy in FVS subthreshold circuits we developed in this dissertation. Illustration is given in Fig. 4.14.

**Case 1 - active leakage reduction:** when technology selection is inefficient to bring  $f_{clk,opt}$  down to  $f_{op}$ , the target throughput lies in  $R2/R3$ . Rather than operating at minimum  $V_{dd}$  to meet robustness ( $R3$ ) and throughput ( $R2$ ) constraints, the circuit can be operated at  $V_{dd,opt}$ . The delay is thus shorter than the throughput period  $T_{op}$  and as shown in Fig. 4.14 (a), the resulting positive



**Fig. 4.14.** Application cases of sleep-mode techniques: (a) active leakage reduction when the circuit is in  $R2/R3$  throughput regions and (b) stand-by leakage reduction when the application features stand-by periods without workload.

time slack can be used to enter sleep mode and save static energy. The energy overhead for entering sleep mode and waking up the circuit is quite high in this case as it is repeated at each operation.

**Case 2 - stand-by leakage reduction:** when the application of a minimum-energy circuit ( $f_{clk,opt} = f_{op}$ ) features long periods of inactivity in between tasks, i.e. stand-by periods, a sleep-mode is added to save static energy, as shown in Fig. 4.14 (b). The time ratio between active and stand-by periods is referred to as duty cycle. In this case, the energy overhead for entering sleep-mode and waking up the circuit are low because it is divided by the number of operations in the tasks.

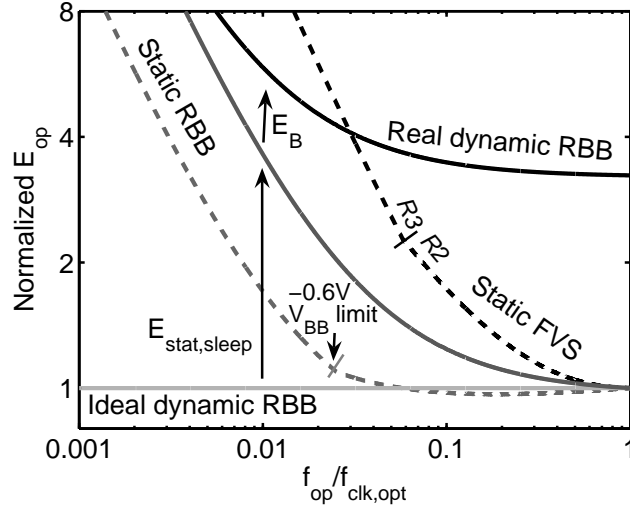
In this section, we compare the efficiency of the two most widespread sleep-mode techniques for high-performance circuits [19]: dynamic reverse body biasing or power gating. The goal is to determine the most effective one in nanometer subthreshold circuits to address active/stand-by leakage reduction. We first independently examine the impact of these techniques on practical energy before comparing them and discussing the results.

#### 4.4.1 Impact of dynamic reverse body biasing on practical energy

The first technique consists in applying a reverse body bias when in sleep mode, while leaving a zero body bias when in active mode. This is dynamic RBB also called Virtual-Threshold CMOS (VTCMOS). Practical energy with dynamic RBB can be expressed as:

$$\begin{aligned}
 E_{op,DRBB} &= E_{dyn} + E_{stat,act} + E_{stat,sleep} + E_B \\
 &= E_{min} + E_{stat,sleep} + E_B \\
 &= E_{min} + V_{dd,opt} I_{leak,sleep} \times (T_{op} - T_{del,opt}) + \frac{1}{2} C_B V_{BB}^2, \quad (4.3)
 \end{aligned}$$

where  $E_{stat,sleep}$  is the static energy in sleep mode due to non-zero  $I_{leak,sleep}$  leakage currents and  $T_{del,opt}$  is the delay at minimum-energy point  $T_{del,opt} =$

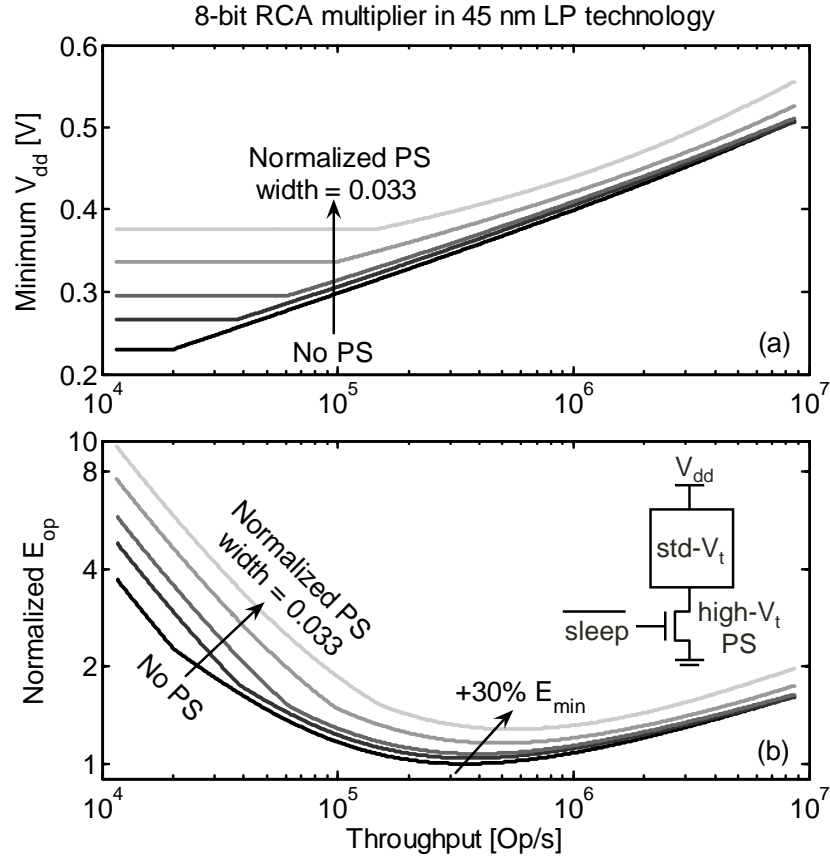


**Fig. 4.15.** Impact of dynamic reverse body biasing (solid lines) on practical energy for active leakage reduction (case 1:  $f_{op} < f_{clk,opt}$ , duty cycle = 1). We consider a  $-0.6V$   $V_{BB}$  as the maximum affordable voltage for ULP circuits. Practical energy with static FVS and RBB schemes (dashed lines) is given as a comparison baseline. This shows that the energy overhead  $E_B$  for generating the body bias completely ruins the static energy saving brought by dynamic RBB. Moreover, even without considering  $E_B$ , dynamic RBB is less efficient than static RBB coupled with FVS.

$1/f_{clk,opt}$ .  $E_B$  is the energy required to charge the capacitances  $C_B$  associated to the body nodes, i.e. N-well and isolated P-substrate in a triple-well process. Spice simulation of the benchmark multiplier gives 70 fJ, that has to be compared to the 32 fJ  $E_{min}$  at  $V_{BB} = 0$ . This suggests that the overhead of charging  $C_B$  will dramatically impact practical  $E_{op}$  in case 1, when sleep-mode is entered at each operation period. Notice that we make the basic assumption of ideal clock gating in sleep mode, which fully eliminates dynamic energy component in sleep mode.

Fig. 4.15 shows the simulated practical  $E_{op}$  in case 1 for active leakage reduction with dynamic RBB. Notice that we consider a maximum reverse  $V_{BB}$  of  $-0.6V$  for area/volume-constrained ULP applications. Detailed in this figure are the contributions of ideal dynamic RBB energy, which corresponds to  $E_{min}$ , and of  $E_{stat,sleep}$  and  $E_B$ . The overhead of  $E_B$  is mostly important when  $f_{op}$  is close to  $f_{clk,opt}$ . An improvement over static FVS (baseline, minimum  $V_{dd}$  for meeting robustness and throughput constraints) is yielded when  $f_{op}$  is  $30\times$  lower than  $f_{clk,opt}$ . However, static RBB operating at minimum  $V_{dd}$  (similar to Section 4.3) features lower  $E_{op}$  than dynamic RBB even when neglecting  $E_B$ , thanks to reduced active-mode leakage with static RBB. This shows that dynamic RBB is not an interesting solution for active leakage reduction. Stand-by leakage reduction by dynamic RBB is addressed in Section 4.4.3.

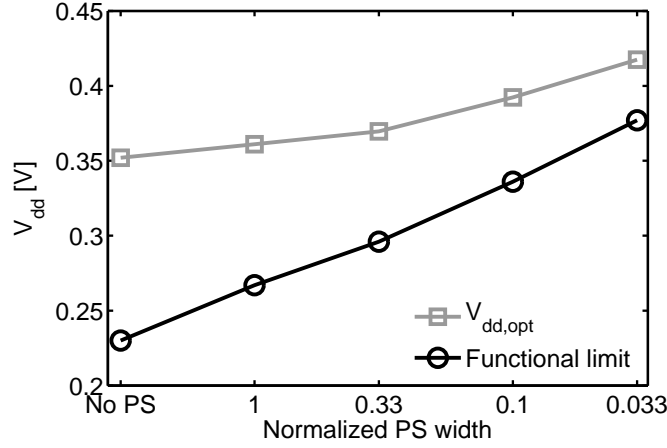




**Fig. 4.16.** Impact of the power switch (PS) on active-mode operation: (a) minimum  $V_{dd}$  under robustness and throughput constraints with (b) corresponding practical energy per operation normalized to  $E_{min}$ . The power switch is a high- $V_t$  NMOS device with minimum  $L_g$  and its width is normalized to the total width of NMOS stacks in the circuit ( $36 \mu\text{m}$ ). Considered normalized widths are 1, 0.33, 0.1, 0.033.

#### 4.4.2 Impact of power gating on practical energy

The second usual technique is to disconnect the circuit from the power supply when in sleep mode. This is achieved through the insertion of a sleep transistor or power switch (PS) between one or both the supply rails of the circuit and the power supply sources [17]. In order to achieve high leakage reduction, a high- $V_t$  device is often used as power switch and this technique is thus also called multi-threshold CMOS (MTCMOS) power gating (PG). A study on power gating for subthreshold circuits was recently carried out in [1] to address stand-by leakage reduction. Seok *et al.* show that sleep-mode energy due to the remaining



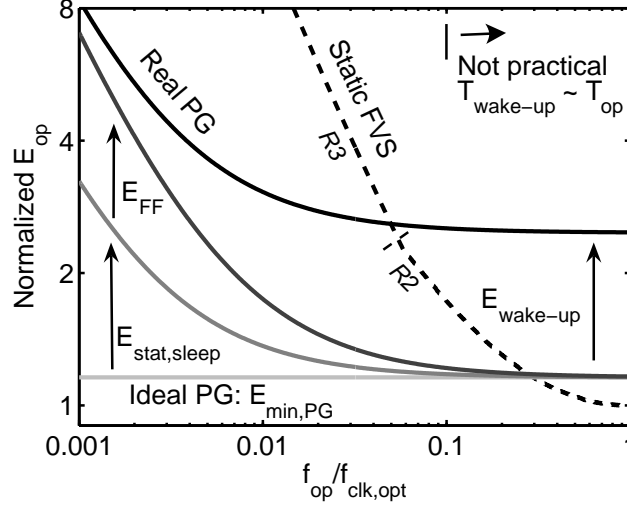
**Fig. 4.17.** Impact of the power switch on minimum functional  $V_{dd}$  and optimum  $V_{dd}$  of minimum-energy point. Both are raised and the margin in between is reduced due to the insertion of the power switch.

leakage in sleep mode cannot be overlooked for very low duty cycle values and that in subthreshold regime, the impact of the power switch on delay in active mode is more important than at nominal  $V_{dd}$ . Indeed, the insertion of the power switch affects active-mode operation as it generates a resistive voltage drop, which reduces the effective  $V_{dd}$ . Let us first examine that.

#### Active mode

Fig. 4.16 (a) shows that the insertion of an NMOS high- $V_t$  power switch modifies minimum  $V_{dd}$  of the benchmark multiplier. It degrades the static noise margins and the delay, thereby raising minimum  $V_{dd}$  in all throughput regions. The impact on minimum  $V_{dd}$  is more severe when the width of the power switch is decreased because it generates higher voltage drop. Another consequence is the increase of  $E_{min}$  level up to +30% because of the delay degradation, as shown in Fig. 4.16 (b). The evolution of minimum  $V_{dd}$  for robustness constraints (functional limit) and  $V_{dd,opt}$  of minimum-energy point is shown in Fig. 4.17 vs. the power switch width: the smaller the power switch, the smaller the margin between  $V_{dd,opt}$  and the functional limit. This shows that not only delay degradation should be investigated when engineering the power switch as in [1]. Indeed, special care should also be taken to avoid ruining static noise margins, which are already very low without a power switch in nanometer subthreshold circuits.

Notice that as usual minimum  $V_{dd}$  is statistically extracted from Monte-Carlo simulations with intrinsic device variability. However, we make the assumption of coarse-grain power gating and we thus do not consider variability for the power switch as its large width strongly mitigates random doping fluctuations and other



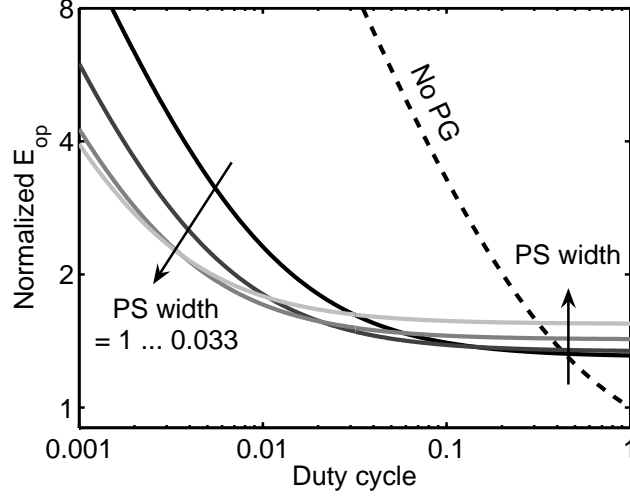
**Fig. 4.18.** Impact of power gating (solid lines) on practical energy for active leakage reduction (case 1:  $f_{op} < f_{clk,opt}$ , duty cycle = 1) with the energy contributions from sleep-mode leakage, non-gated high- $V_t$  flip-flops and wake-up (NMOS high- $V_t$  power switch, minimum  $L_g$ , normalized width = 0.1). Practical energy with static FVS scheme (dashed line) is given as a comparison baseline. When  $f_{op}$  is close to  $f_{clk,opt}$ , active and wake-up energies dominate, while flip-flop energy dominates at very low  $f_{op}$ .

intrinsic variability sources. The power switch variability would make things worse and fine-grain power gating is thus not recommended for subthreshold logic.

#### Sleep mode

The leakage in sleep mode is strongly mitigated by the power gating but it is not equal to zero. Sleep-mode static energy  $E_{stat,sleep}$  has thus to be considered carefully [1]. Fig. 4.18 shows practical energy of the benchmark multiplier with power gating in case 1 to address active leakage reduction. Ideal energy is equal to minimum-energy level of power-gated (PG) circuit, which is higher than non-PG circuit because of the delay penalty of the power switch.  $E_{stat,sleep}$  becomes important when  $f_{op}$  is  $100\times$  lower than  $f_{clk,opt}$ .

In a PG circuit, sequential elements such as flip-flops cannot be disconnected from the power supply, to preserve circuit state. MTCMOS state-retention flip-flops have to be used with a high- $V_t$  shadow latch (balloon circuit), which is not power-gated [20]. Other architectures of state-retention MTCMOS flip-flops can be found in [21, 22, 23]. These flip-flops do have leakage currents and their contribution to  $E_{op}$  cannot be overlooked. To include their associated energy contribution in these results, we simulate a latch based on cross-coupled inverters with high- $V_t$  devices with  $L_g = 60$  nm to reduce their leakage. The corresponding



**Fig. 4.19.** Impact of power switch sizing (solid lines) on practical energy for stand-by leakage reduction ( $f_{op} = f_{clk,opt}$ , duty cycle  $< 1$ ). Practical energy without power gating (dashed line) is given as a comparison baseline. Downsizing the power switch increases  $E_{min}$  and thus  $E_{op}$  when the duty cycle is close to 1. For low duty cycle values,  $E_{op}$  is reduced with smaller power switches down to the energy limit of the non-gated low-leakage memory.

leakage under 0.3V  $V_{dd}$  is 2.5 pA per flip-flop. We consider that each output of the multiplier features an associated flip-flop. The corresponding energy contribution  $E_{FF}$  is plotted in Fig. 4.18.  $E_{FF}$  level is comparable to  $E_{stat,sleep}$  level.

Finally, wake-up energy  $E_{wake-up}$  to drive the gate of the power switch and recharge internal nodes of the PG circuit when leaving sleep mode has to be taken into account as well. Spice simulations gives an  $E_{wake-up}$  of 42 fJ to which mainly internal-capacitance recharge contributes. The total practical energy of a PG circuit in case 1 can thus be expressed as:

$$E_{op,PG} = E_{dyn} + E_{stat,act} + E_{stat,sleep} + E_{wake-up} . \quad (4.4)$$

Total  $E_{op,PG}$  is represented in Fig. 4.18, which shows that PG is efficient only if  $f_{op}$  is  $20\times$  lower than  $f_{clk,opt}$ . Moreover, notice that power gating cannot practically be achieved when  $f_{op}$  is too close from  $f_{clk,opt}$  because the simulated wake-up latency  $T_{wake-up}$  of  $3\mu s$  is in this case longer than the operation period  $T_{op}$ .

When addressing stand-by leakage reduction (case 2),  $E_{wake-up}$  is divided by the number of operations in the task and the associated overhead is thus much smaller. Moreover, in this case, flip-flops can be power-gated. Indeed, with long stand-by periods, we have time to push circuit state (critical registers) into a low-leakage SRAM and the whole logic circuit can be power-gated. This is the

option used in the subthreshold processor in [2]. We make the assumption of  $5\times$  leakage current reduction between the core-process flip-flops and the memory-process SRAM (0.5 pA per output bit). Fig. 4.19 shows the associated  $E_{op}$ , when neglecting  $E_{wake-up}$  as well as the energy required to push/pull circuit state into/from the memory. Nevertheless, we still consider the  $5\times$  lower leakage in the memory thanks to special low-leakage memory devices. The resulting  $E_{op}$  remains below  $2\times$  non-PG  $E_{min}$  for duty cycles down to 0.01. The effect of downsizing the power switch is an increase of PG  $E_{min}$  thus  $E_{op}$  when the duty cycle is close to 1, whereas it reduces  $E_{op}$  at very low duty cycles down to the memory  $E_{stat,sleep}$  limit.

This shows that optimum power switch sizing results from a trade-off between  $E_{stat,sleep}$  reduction thanks to sleep-mode  $I_{leak}$  reduction and  $E_{min}$  increase because of delay penalty. Let us go deeper in power switch engineering.

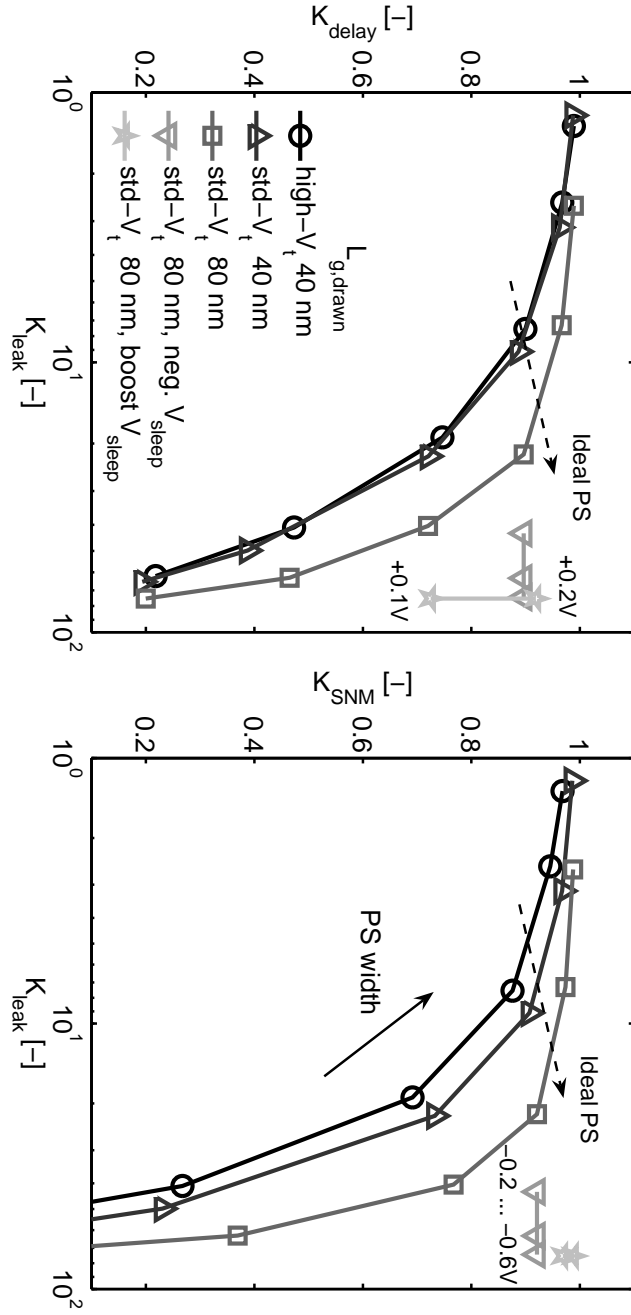
#### Power-switch engineering

Sizing the sleep device is typically a complex task, as it depends on the circuit discharge pattern, which is hard to predict accurately. It results from a trade-off between high leakage reduction in standby mode and low delay penalty in active mode, with also die area to take into consideration. In [1], Seok *et al.* propose to analyze the power switch with two properties: the delay increase factor  $K_{delay}$  and the sleep-mode leakage reduction factor  $K_{leak}$ <sup>1</sup>. However, they do not take circuit robustness into account. Therefore, we introduce in the analysis a robustness property: the SNM reduction factor  $K_{SNM}$  of the Monte-Carlo-extracted  $3\sigma$  worst-case SNM.

Fig. 4.20 plots the properties of the basic high- $V_t$  power switch for different widths. Optimum PS width selection results from a trade-off between low  $K_{delay}$  and  $K_{SNM}$  vs. high  $K_{leak}$ . A narrow power switch improves the leakage reduction but impacts more the delay and the robustness and as shown in Fig. 4.20 (b) can even lead to functional failure (negative  $3\sigma$  worst-case SNM) if the normalized width of the power switch is below 0.1.

In order to improve one of these factors without deteriorating the other ones, we look at different types of power switches. First, for area concern, we consider a std- $V_t$  power switch that may be downsized as compared to high- $V_t$  for constant  $K_{delay}$  and  $K_{SNM}$ . The  $K_{delay}$  vs.  $K_{leak}$  trade-off is identical to high- $V_t$  power switch but the  $K_{SNM}$  vs.  $K_{leak}$  trade-off is somewhat improved. Next, we upsize gate length of the std- $V_t$  power switch to 80 nm, which more significantly improves both trade-offs thanks to subthreshold swing and DIBL reduction. Indeed both these effects have bad impact on the  $I_{on}/I_{off}$  ratio and should thus be mitigated. Finally, the sleep signal can also be engineered by using a charge pump: negative voltage in sleep mode or voltage above  $V_{dd}$  in active mode. Fig. 4.20 shows that a 200mV boost on sleep in active mode yields a comparable  $K_{delay}$  vs.  $K_{leak}$  trade-off with a better  $K_{SNM}$  vs.  $K_{leak}$  trade-off, as a 600mV negative bias on sleep (with adapted widths of the power switch). Boosting sleep

<sup>1</sup>The  $K_{leak}$  definition we use here is the invert of the definition from [1].



**Fig. 4.20.** Effect of sizing and biasing on the performances of the power switch. An std- $V_t$  switch reduces the impact on SNM ( $K_{\text{SNM}}$ ) at iso-leakage reduction ( $K_{\text{leak}}$ ). A gate length upsize improves both impacts on delay ( $K_{\text{del}}$ ) and SNM at iso-leakage reduction. Biasing further improves the trade-offs.

**Table 4.4.** Comparison of leakage reduction techniques to address active and stand-by leakage components

	Active $V_{dd}$ 10 kOp/s	Active Norm. $E_{op}$ 10 kOp/s	Stand-by $V_{dd}$ duty=0.001	Stand-by Norm. $E_{op}$ duty=0.001
FVS std- $V_t$	0.23	4.2	0.35	260
FVS high- $V_t$	0.24	1.4	0.36	40
FVS $L_g=80\text{nm}$	0.24	1.8	0.34	69
Static RBB	0.30	1.07	0.45	51
Dynamic RBB	0.35	4.1	0.35	28
Standard PG	0.37	2.7	0.37	5.0
Optimum PG	0.36	2.5	0.36	2.1

signal in conjunction with a narrow  $L_g$ -upsized std- $V_t$  power switch is thus a good optimization, that we use in the techniques comparison in next section.

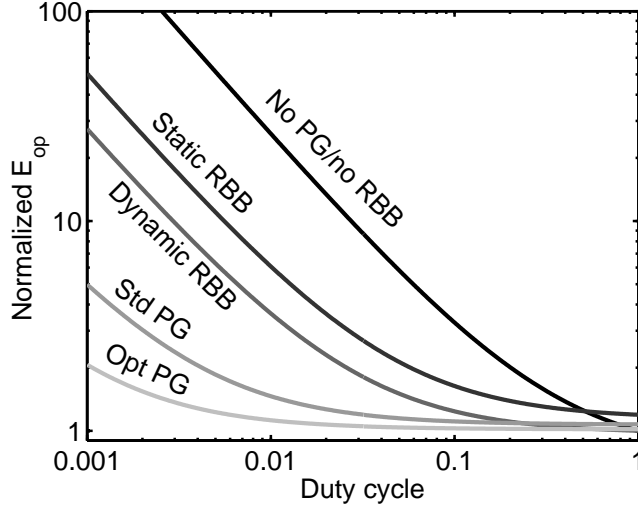
#### 4.4.3 Discussion

Let us finally use previous observations to compare the considered sleep-mode techniques (dynamic RBB and power gating) with static leakage-reduction techniques to address both active and stand-by leakage components.

##### Case 1 - active leakage reduction

Table 4.4 summarizes the operating  $V_{dd}$  and practical  $E_{op}$ , in active-leakage- and standby-leakage-dominated cases.  $E_{op}$  is normalized to  $E_{min}$  of std- $V_t$  FVS circuit without leakage reduction technique (baseline).

In the active-leakage-dominated case, we choose 10 kOp/s for the low  $f_{op}$  as the lower bound of the throughput range for ULP applications. In this case, target  $f_{op}$  is  $35\times$  lower than  $f_{clk,opt}$  of minimum-energy point. At this throughput, the circuit only relying on FVS lies in  $R3$  region with an  $E_{op}$  equal to  $4.2\times E_{min}$ . The use of high- $V_t$  device or an upsized gate length considerably improves  $E_{op}$  while static RBB ( $V_{BB} = -0.6\text{V}$ ) reduces  $E_{op}$  overhead to only 7%. As explained in Section 4.4.1, the overhead of charging the body accesses completely ruins the efficiency of dynamic RBB ( $V_{BB} = -0.6\text{V}$ ) for active leakage reduction. Although wake-up energy overhead is somewhat lower than body-access charging, power gating with either a basic power switch (high- $V_t$  and minimum  $L_g$ ) or an optimized one (std- $V_t$  with  $L_g = 80\text{nm}$  and 200mV-boosted  $\overline{sleep}$  in active-mode) also features higher  $E_{op}$  than a simple circuit with high- $V_t$  or upsized- $L_g$  devices (without sleep-mode nor BB technique). This confirms that the versatility of the technology is extremely powerful for bringing practical  $E_{op}$  to  $E_{min}$  level. With moderate extra design cost, static RBB gives even more interesting results. On the opposite, sleep-mode techniques have high extra design cost and poor efficiency.



**Fig. 4.21.** Comparison of the different techniques for stand-by leakage reduction ( $f_{op} = f_{clk,opt}$ ,  $duty\ cycle < 1$ ). Power-gating techniques yield the lower practical energy per operation.

#### Case 2 - stand-by leakage reduction

When it comes to the reduction of stand-by leakage, things are different because the impact of body-access-charging and wake-up energies on  $E_{op}$  associated to dynamic RBB and power-gating techniques, respectively, is divided by the number of operations in the task. In Fig. 4.21,  $E_{op}$  is plotted from simulation without these energy components and with a low-leakage state-retention SRAM for the power gating scheme. In this case, dynamic RBB features lower  $E_{op}$  than static RBB. Recall that in this case we assume that the circuit operates at minimum-energy point of simple FVS circuit ( $f_{op} = f_{clk,opt}$ ). Consequently, static RBB requires higher  $V_{dd}$  to support the throughput constraint, leading to higher leakage and thus higher sleep-mode energy. Power gating is more efficient than dynamic RBB and optimization of the power switch further reduces  $E_{op}$ .

Table 4.4 summarizes the results with all techniques when considering a duty cycle of 0.001. In this case, static techniques are less efficient than sleep-mode techniques. In particular, power gating with an optimized power switch brings practical  $E_{op}$  down to only  $2.1 \times$  higher than  $E_{min}$ .



## 4.5 CONCLUSION

In Chapter 3, we showed how optimum MOSFET selection and the use of a fully-depleted SOI technology can significantly improve minimum-energy level of nanometer subthreshold circuits. In this chapter, we extended this work by investigating ways to make practical energy under robustness and throughput constraints meet the minimum-energy level. We therefore revisited typical circuit design choices in this light.

We first showed that changing technology flavor and the device  $V_t$  shifts the minimum-energy point to different throughputs. The versatility of nanometer technologies is thus a powerful tool to minimize practical energy, as technology/device selection allows the circuit to operate close to minimum-energy point for a wide throughput range from tens of kOp/s to tens of MOp/s. For most of the throughput range of ULP applications, low-power technology flavor brings the lowest practical energy. We also demonstrated that independent high- $V_t$  assignment to save leakage in non-critical paths is not feasible in nanometer subthreshold circuits because of the large delay difference between std- and high- $V_t$  subthreshold logic gates and the high variability of short paths.

Modeling errors, global process or temperature variations and device aging may imply a wrong estimation of the throughput of minimum-energy point. It may result in a bad technology choice for making minimum-energy point meet the target application throughput, leading to practical energy overhead up to 90%. We then showed that adaptive body biasing is an efficient technique, more efficient than adaptive supply voltage, to compensate for this throughput mismatch. However, this is only true for reverse body biasing as forward body biasing increases minimum-energy level and badly behaves with discrete bias voltage values. Relying on adaptive reverse-only body biasing, the constraints for designers when making the technology selection is thus to ensure that the throughput of minimum-energy point will not be higher than the target throughput. This means that the technology selection should be achieved by considering the worst-case process/temperature corner for speed (SS process at low temperature). Adaptive reverse body biasing is then used to remove the design margins by increasing  $V_t$  and thus limiting subthreshold current in case of a positive timing slack. At 45 nm node, we point out that reverse body biasing is only efficient in low-power technology flavor and we suggest that at next nodes it may no longer be practical because of decreasing body-bias coefficient and increasing band-to-band tunneling leakage. This reemphasizes the need for a technology with less sensitivity against global process and temperature variations such as fully-depleted SOI.

Finally, we investigated the efficiency of sleep-mode techniques - dynamic reverse body biasing and power gating - for reducing:

- active leakage when the throughput of minimum-energy point falls well above the target throughput,
- stand-by leakage when the circuit has long inactivity periods.

For active-leakage reduction, sleep-mode techniques have high energy overheads to achieve the transition between active and sleep modes. Technology selection (high- $V_t$  devices in low-power flavor) and static reverse body biasing are consequently more efficient than sleep-mode techniques in this case, with the additional benefit of lower extra design cost. However, for reducing stand-by leakage in nanometer subthreshold circuits, power gating is a very efficient technique but we showed that the circuit robustness can be under risk when using badly-sized power switches. Engineering the power switch with longer- $L_g$  std- $V_t$  devices and boosted *sleep* signal is shown to bring significant energy reduction with lower robustness degradation.

## REFERENCES

1. M. Seok, S. Hanson, D. Sylvester and D. Blaauw, "Analysis and optimization of sleep modes in subthreshold circuit design", in *Proc. ACM/IEEE Des. Autom. Conf.*, pp. 694-699, 2007.
2. J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann and A. Chandrakasan, "A 65 nm sub- $V_t$  microcontroller with integrated SRAM and switched-capacitor DC-DC converter", in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 318-319, 2008.
3. S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhla, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester and D. Blaauw, "Exploring variability and performance in a sub-200-mV processor", in *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 881-891, Apr. 2008.
4. M. V. Dunga *et al.*, "BSIM4.6.1 MOSFET model", available on-line at [www-device.eecs.berkeley.edu/bsim3/bsim4.html](http://www-device.eecs.berkeley.edu/bsim3/bsim4.html).
5. X. Li *et al.*, "PSP 102.3", available on-line at [pspmodel.asu.edu/downloads/psp102p3-summary.pdf](http://pspmodel.asu.edu/downloads/psp102p3-summary.pdf).
6. L. Wei, Z. Chen, K. Roy, M. C. Johnson, Y. Ye and V. K. De, "Design and optimization of dual-threshold circuits for low-voltage low-power applications", in *IEEE Trans. VLSI Syst.*, vol. 7, no. 1, pp. 16-24, Mar. 1999.
7. D. Blaauw, K. Chopra, A. Srivastava and L. Scheffer, "Statistical timing analysis: from basic principles to state of the art", in *IEEE Trans. Comp.-Aided Des. Integrated Circuits Syst.*, vol. 27, no. 4, pp. 589-607, Apr. 2008.
8. B. Zhai, S. Hanson, D. Blaauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 20-25, 2005.
9. N. Sirisantan, L. Wei and K. Roy, "High-performance low-power CMOS circuits using multiple channel length and multiple oxide thickness", in *Proc. IEEE Int. Conf. Comp. Des.*, pp. 227-232, 2000.
10. J. T. Kao, M. Masayuki and A. P. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture", in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545-1554, Nov. 2002.
11. J. T. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan and V. De, "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage", in *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396-1402, Nov. 2002.
12. H. Fuketa, M. Hashimoto, Y. Mitsuyama and T. Onoye, "Correlation verification between transistor variability model with body biasing and ring oscillation frequency in 90nm subthreshold circuits", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 3-8, 2008.
13. A. Keshavarzi, S. Ma, S. Narendra, B. Bloechel, K. Mistry, T. Ghani, S. Borkar and V. De, "Effectiveness of reverse body bias for leakage control in scaled dual  $V_t$  CMOS ICs", in *Proc. IEEE/ACM Int. Symp. Low-Power Electron. Des.*, pp. 207-212, 2001.

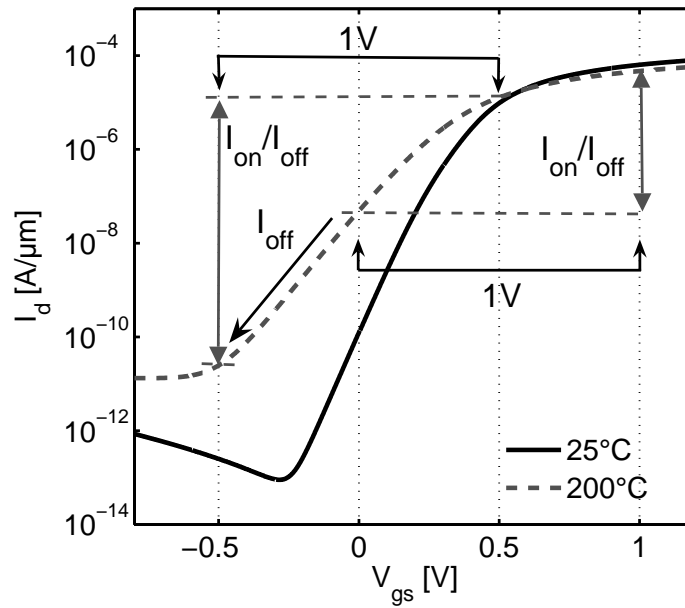
14. B. H. Calhoun and A. P. Chandrakasan, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering", in *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 238-245, Jan. 2006.
15. H. Suzuki, M. Kurimoto, T. Yamanaka, H. Takata, H. Makino and H. Shinohara, "Post-Silicon programmed body-biasing platform suppressing device variability in 45 nm CMOS technology", in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 15-20, 2008.
16. A. Agarwal, S. Mukhopadhyay, A. Raychowdhury, K. Roy and C. H. Kim, "Leakage power analysis and reduction for nanoscale circuits", in *IEEE Micro*, vol. 26, no. 2, pp. 68-80, Mar.-Apr., 2006.
17. S. Mutoh, T. Douseki, Y. Matsuya, S. Shigematsu and J. Yamada, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS", in *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847-854, Aug. 1995.
18. Y. Ikenaga, M. Nomura, Y. Nakazawa and Y. Hagihara, "A circuit for determining the optimal supply voltage to minimize energy consumption in LSI circuit operations", in *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 911-918, Apr. 2008.
19. H. Xu, R. Vermuri and W.-B. Jone, "Run-time active leakage reduction by power gating and reverse body biasing: an energy view", in *Proc. IEEE Int. Conf. Comp. Des.*, pp. 618-625, 2008.
20. S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe and J. Yamada, "A 1-V high-speed MTCMOS circuit scheme for power-down application circuits", in *IEEE J. Solid-State Circuits*, vol. 32, no. 6, pp. 861-869, Jun. 1997.
21. J. T. Kao and A. P. Chandrakasan: "MTCMOS sequential circuits", in *Proc. European Solid-State Circuits Conf.*, pp. 317-320, 2001.
22. B. H. Calhoun and A. P. Chandrakasan, "Standby power reduction using dynamic voltage scaling and canary flip-flop structures", in *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1504-1511, Sep. 2004.
23. D. Levacq, V. Dessard and D. Flandre, "Ultra-low power flip-flops for MTCMOS circuits", in *Proc. IEEE Int. Symp. Circuits Syst.*, pp. 4681-4684, 2005.

## CHAPTER 5

---

### BUILDING ULTRA-LOW-POWER HIGH-TEMPERATURE DIGITAL CIRCUITS IN STANDARD SOI CMOS TECHNOLOGY

---



**Fig. 5.1.** Simulated NMOS  $I_d/V_{gs}$  characteristics in 0.13  $\mu\text{m}$  partially-depleted SOI technology at 25°C and 200°C (floating-body device with  $W/L = 1/0.13$  [ $\mu\text{m}$ ] under  $V_{ds} = 1$  V). High-temperature operation leads to higher  $I_{off}$  and degraded  $I_{on}/I_{off}$  ratio. This can be dealt with by operating with  $V_{gs}$  between -0.5 V and +0.5 V instead of between 0 V and +1 V. *Let us see how.*

### Abstract

---

In high-temperature environments ( $> 150^{\circ}\text{C}$ ), static power/energy consumption completely dominates, even at  $0.13\text{ }\mu\text{m}$  node. As no technology option in scaled technology nodes solves this issue, we propose a new logic style, named Ultra-Low-Power (ULP), which achieves negative  $V_{gs}$  self-biasing, to benefit from the small area and low dynamic power of scaled technologies while keeping ultra-low leakage, even at high temperature [CP1][PA1]. In  $0.13\text{ }\mu\text{m}$  partially-depleted SOI CMOS technology, ULP logic style reduces static power consumption at  $200^{\circ}\text{C}$  by 3 orders of magnitude at the expense of increased delay and area, with good robustness against process variations [CP2][JP1]. Moreover, ULP logic gates feature excellent noise robustness thanks to SNM higher than  $V_{dd}/2$ , which is never achieved in standard CMOS logic style. Functionality of ULP logic style is demonstrated by measurement results of ULP-inverter ring oscillators in  $0.13\text{ }\mu\text{m}$  technology.

### Contents

---

<b>5.1</b>	<b>Introduction</b>	131
<b>5.2</b>	<b>High-temperature MOSFET behavior</b>	131
<b>5.3</b>	<b>ULP transistor</b>	132
<b>5.4</b>	<b>ULP logic style</b>	137
<b>5.5</b>	<b>Impact of PVT variations on performances</b>	144
<b>5.6</b>	<b>Validation of ULP logic style</b>	146
<b>5.7</b>	<b>Conclusion</b>	150

---

## 5.1 INTRODUCTION

In Chapter 4, we analyzed solutions for limiting static energy of FVS subthreshold circuits for ULP applications mainly targeting consumer or biomedical markets. In the industrial sector, ULP circuits are also required for distributed process monitoring and control. In comparison to consumer and biomedical devices, industrial applications have very different environment conditions. In applications such as oil drilling, downhole monitoring, combustion engine or harsh-environment industrial process control, the operating temperature can be very high, up to  $300^{\circ}\text{C}$ . At high temperature, the behavior of MOSFETs is degraded and leads to robustness issues at low voltage and orders-of-magnitude higher leakage currents [1, 2], which can hardly be dealt with by traditional leakage-reduction techniques. As suggested in Chapter 1, FVS circuits thus lie in  $R2/R3$  regions dominated by static power/energy component for the whole throughput range of ULP applications.

In this chapter, we propose a new leakage-mitigation technique based on the Ultra-Low-Power (ULP) transistor concept to benefit from the low die area and low dynamic power of standard CMOS technologies in scaled nodes, while keeping ultra-low leakage, even at high temperature. The proposed ULP transistor achieves negative  $V_{gs}$  self-biasing and allows to build an ULP logic style with orders-of-magnitude reduction of leakage current at the expense of increased delay. We show that ULP logic style is a robust and straightforward technique to build ultra-low-power high-temperature digital circuits in scaled standard SOI technology.

This chapter is organized as follows. In Section 5.2, we briefly review the MOSFET behavior at high temperature, illustrated with an industrial  $0.13\,\mu\text{m}$  partially-depleted (PD) SOI technology. We present the concept of ULP transistor and investigates its properties in the considered technology in Section 5.3. The building of ULP logic gates based on the ULP transistor is described in Section 5.4 and gate performances are evaluated. Section 5.5 deals with the impact of process, voltage and temperature (PVT) variations on performances of ULP logic style. Finally, the building of ULP logic circuits is validated in Section 5.6.

## 5.2 HIGH-TEMPERATURE MOSFET BEHAVIOR

Temperature increase has several key effects on MOSFET operation [1]:

- carrier mobility reduction in the channel,
- junction leakage increase,
- subthreshold swing degradation,
- $V_t$  lowering.

The degradation of carrier mobility leads to lower drain current, which in turn implies a delay penalty in digital circuits. This penalty is not an issue in ultra-low-power applications as long delays are tolerated thanks to low computational load and reduced clock frequencies. However, the other effects are detrimental because they all increase static power through leakage currents. SOI technology is used most of the time, as it considerably mitigates these effects [1, 3]. Moreover, micron-scale processes with 0.8-1  $\mu\text{m}$  channel length and very high  $V_t$  are often used in order to keep subthreshold leakage under control [4],[CO2]. Such technologies operate under high  $V_{dd}$  (2.5-5 V), which combined with the higher capacitance of the micron-scale devices leads to high dynamic power consumption.

As detailed in Chapter 2, scaled deep-submicron and nanometer CMOS technologies feature lower  $V_t$  with higher subthreshold leakage current as a result. Fig. 5.1 depicts the NMOS  $I_d/V_{gs}$  characteristics at 25°C and 200°C, in an industrial 0.13  $\mu\text{m}$  partially-depleted SOI CMOS technology (oxide thickness  $T_{ox}$ =2.0 nm, silicon film thickness  $T_{Si}$ =150 nm,  $V_t$ =0.34 V in saturation at 25°C). Under 1 V  $V_{ds}$ , the operation at 200°C increases  $I_{off}$  leakage current by a factor 400. Moreover, temperature increase implies a reduction of  $I_{on}/I_{off}$  ratio because of subthreshold slope degradation and  $V_t$  lowering, which degrades the static power/delay trade-off of digital circuits and could lead to robustness issues, i.e. reduced SNM and degraded output logic levels [1]. As shown in Fig. 5.1, a way to restore acceptable  $I_{on}/I_{off}$  ratio is to operate in subthreshold regime with negative  $V_{gs}$ . With a fixed 1 V  $V_{gs}$  swing, operation between -0.5 V and +0.5 V  $V_{gs}$ , instead of between 0V and +1V, yields an important  $I_{on}/I_{off}$  ratio improvement. However, signals in digital circuits have logic levels: 0 V for low logic level and  $V_{dd}$  for high logic level. It is thus not possible in standard CMOS logic style to operate with negative  $V_{gs}$  as the gate of NMOS (resp. PMOS) devices are connected to logic signals and their source to ground (resp.  $V_{dd}$ ). The concept of ULP transistor, presented in next section allows to get negative  $V_{gs}$ , by adaptive source self-biasing.

### 5.3 ULP TRANSISTOR

The general ULP concept was first introduced in [5], with proposal of ULP basic blocks such as ULP voltage reference [6] and ULP diode [7], as well as analog circuits thereof. In [8, 9], the building of a 7-transistor SRAM cell based on the ULP diode in 0.13  $\mu\text{m}$  PD SOI technology was proposed, with correct operation and ultra-low leakage demonstrated up to 250°C.

ULP transistor is a new concept from this ULP family [PA1]. The goal is to extend the ultra-low-leakage property of ULP-diode-based SRAM cell to computing circuits by designing an ULP logic style. In this section, the concept of the ULP transistor is presented: its structure, the leakage-reduction mechanism and its current-voltage characteristics.



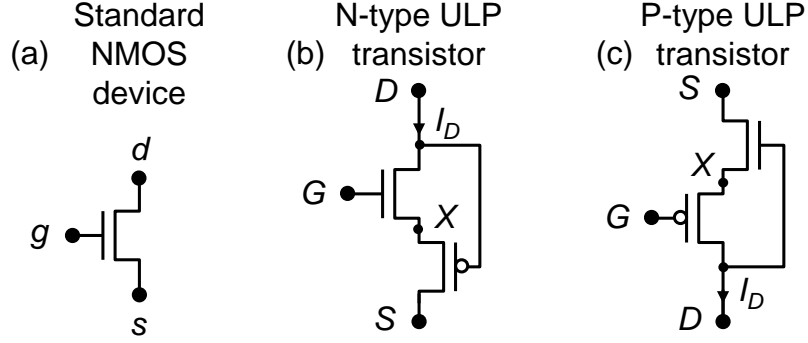


Fig. 5.2. Transistor structures

### 5.3.1 Principle

The structures of N-type and P-type ULP transistors are depicted in Fig. 5.2 and compared to standard NMOS device. As the principle of P-type ULP transistor is similar to N-type one, we only discuss here the N-type ULP transistor. It is composed by an NMOS device stacked upon a PMOS device. PMOS gate is connected to NMOS drain. In partially-depleted SOI technology, floating-body devices are considered for layout compactness issues. The structure behaves externally as an NMOS transistor with 3 accesses: gate ( $G$ ), source ( $S$ ) and drain ( $D$ ). Throughout this chapter, we use upper case letters to reference the global accesses of the ULP transistor ( $G$ ,  $S$ ,  $D$ ) whereas lower case letters are used to reference the accesses of CMOS devices ( $g$ ,  $s$ ,  $d$ ) either isolated or within the ULP transistor.

### 5.3.2 Leakage reduction mechanism

Let us first consider the standard NMOS device from Fig. 5.2 with its source tied to ground. In  $0.13\ \mu\text{m}$  SOI CMOS technology, as shown in Fig. 5.3, its off-state current  $I_{off,NMOS}$  is dominated by subthreshold leakage  $I_{sub}$ :

$$I_{sub} = I_0 \times 10^{\frac{V_{gs} + \eta V_{ds}}{S}} \times \left( 1 - e^{\frac{-V_{ds}}{U_{th}}} \right) \quad (5.1)$$

$$I_{off,NMOS} = I_0 \times 10^{\frac{\eta V_{ds}}{S}} \times \left( 1 - e^{\frac{-V_{ds}}{U_{th}}} \right) \quad (5.2)$$

where the subthreshold swing  $S$  is equal to  $\ln(10) n U_{th}$  and  $I_0$  is proportional to  $10^{-V_t/S}$ . Temperature affects  $U_{th}$ , which increases  $S$ , as well as  $V_t$ , which increases  $I_0$  and in turn  $I_{off}$ . Notice that in partially-depleted SOI floating-body devices, the impact of back-gate bias can be neglected.

Measured  $I_{off,NMOS}$  vs.  $V_{ds}$  in the considered  $0.13\mu\text{m}$  SOI technology is plotted in Fig. 5.3 at  $25^\circ\text{C}$  and  $200^\circ\text{C}$  (solid and dashed lines, respectively). In this figure, the model from Eq. (5.2), with fitted parameters, is plotted with circle markers, showing very good match ( $n=1.55$ ,  $\eta=120\text{ mV/V}$ ,  $I_0=0.48$  and  $81\text{ nA}/\mu\text{m}$  at  $25^\circ\text{C}$  and  $200^\circ\text{C}$ , respectively). The discrepancy above  $0.8\text{V}$  is due to the floating-body effect of MOSFETs in partially-depleted SOI technology. In this figure, the Spice-simulated current is also plotted at  $200^\circ\text{C}$  with cross markers, in order to validate the accuracy of the considered MOSFET compact models<sup>1</sup> we use in this Chapter.

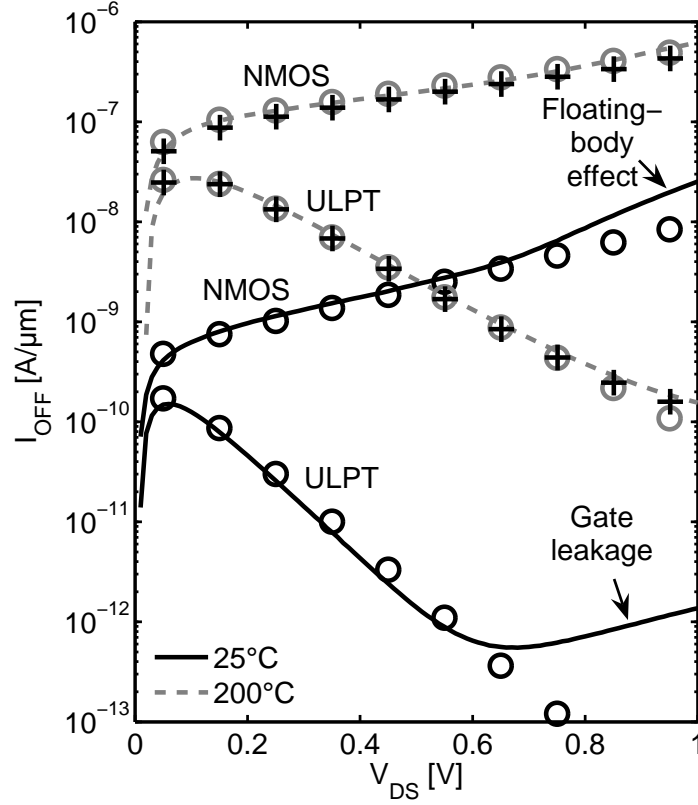
Let us now focus on the ULP transistor. The leakage reduction mechanism is based on the self-biased negative  $V_{gs}$  of the NMOS and PMOS devices inside the ULP transistor. For the N-type ULP transistor from Fig. 5.2,  $V_{gs}$  of the internal NMOS and PMOS devices depend on the voltage of the internal node  $X$ .  $V_{gs}$  of the NMOS is  $V_{GX}$ , whereas  $V_{gs}$  of the PMOS is  $V_{DX}$ . If NMOS and PMOS devices have symmetrical  $V_t$ , by symmetry  $V_{XS}$  is equal to  $V_{DS}/2$ , when  $V_{GS}=0$ . Devices thus operate with  $V_{gs}$  equal to  $-V_{DS}/2$ , leading to ultra-low leakage. The NMOS device of the ULP transistor thus has  $V_{gs}=-V_{DS}/2$ ,  $V_{ds}=V_{DS}/2$ . Replacing these terms in Eq. (5.1) yields the drain subthreshold leakage current of an N-type ULP transistor with  $V_{GS}=0$  as:

$$I_{off,ULP} = I_0 \times 10^{\frac{(\eta-1)V_{DS}}{2S}} \times \left(1 - e^{\frac{-V_{DS}}{2V_{th}}}\right). \quad (5.3)$$

Fig. 5.3 shows the measured subthreshold current of N-type ULP transistor at  $25^\circ\text{C}$  and  $200^\circ\text{C}$  (dashed and solid lines, respectively). When  $V_{DS}$  increases, the subthreshold current first increases because  $V_{ds}$  of NMOS and PMOS devices increase too. Then, it strongly decreases as  $V_{gs}$  of the devices become more and more negative. Model from Eq. (5.3) is plotted too with circle markers. The agreement with measured  $I_{off,ULP}$  is very good, except above  $0.7\text{V}$  at  $25^\circ\text{C}$  because in these conditions, leakage current is no longer dominated by subthreshold leakage but by gate-tunneling leakage current [9],[CP1]. At  $200^\circ\text{C}$ , Spice simulations also show almost perfect match with measurements.

At  $1\text{ V}$ , the use of ULP transistor reduces  $I_{off}$  by more than 4 orders of magnitude at  $25^\circ\text{C}$  and more than 3 orders of magnitude at  $200^\circ\text{C}$ . With a measured subthreshold swing of  $140\text{ mV/dec}$  at  $200^\circ\text{C}$ , NMOS device should have an increase of  $V_t$  by  $500\text{ mV}$  in order to get an  $I_{off}$  current as low as the ULP inverter. The resulting  $V_t$  of NMOS devices should thus be higher than  $0.8\text{ V}$ , which is hardly proposed by chip manufacturers in scaled technologies. ULP transistor is thus an efficient and straightforward way to achieve ultra-low  $I_{off}$  in scaled standard technologies with neither extra mask nor process cost.

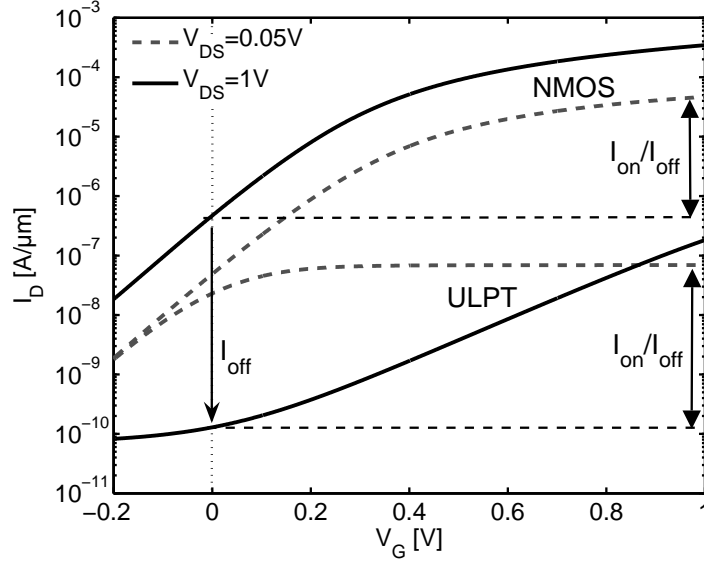
<sup>1</sup>Considered MOSFET compact models are industrial BSIM3SOI models, which are basically valid up to  $125^\circ\text{C}$ . We thus recalibrate the parameters to fit transistor measurements at  $200^\circ\text{C}$ , in order to get representative results in next sections. Excellent agreement between simulations and measurements is achieved by only modifying four BSIM parameters: `Nfactor`, `eta0`, `vth0` and `Isbjt`.



**Fig. 5.3.** Measured  $I_{off}$  (lines) of NMOS device and N-type ULP transistor (ULPT) at 25°C and 200°C ( $W/L = 25/0.13 \mu\text{m}$ ). Circle markers are model from Eq. (5.2) and (5.3). Cross markers depict Spice simulation results with MOSFET compact models recalibrated at 200°C.

### 5.3.3 $I_D/V_{GS}$ characteristics

The ULP transistor can be analyzed in terms of the typical DC characteristic:  $I_D/V_{GS}$  curves. Fig. 5.4 shows the simulated  $I_D/V_{GS}$  curves of the N-type ULP transistor with both 0.05 and 1 V  $V_{DS}$ . The  $I_d/V_{gs}$  of standard NMOS device is shown too for comparison purpose. At low  $V_{DS}$  and negative  $V_{GS}$ , the ULP transistor is the series connection of a PMOS device with  $V_{gs}$  close to 0 V and an NMOS device with negative  $V_{gs}$ , which limits the current. Under these conditions, the current of ULP transistor with negative  $V_{GS}$  is the same as the NMOS current with negative  $V_{gs}$ . When  $V_{GS}$  increases, the NMOS device inside the ULP transistor turns on and the PMOS with  $V_{gs}$  close to 0 V limits the current.



**Fig. 5.4.** Simulated  $I_D/V_{GS}$  curves of standard NMOS device and N-type ULP transistor (ULPT) at  $200^\circ\text{C}$  ( $W/L = 1/0.13\ [\mu\text{m}]$ )

The ULP transistor current thus saturates at the level of PMOS  $I_{off}$  current with low  $V_{ds}$ .

At high  $V_{DS}$  and  $V_{GS}$  equal to 0 V, the  $I_{off}$  current of ULP transistor is the current of an NMOS device with  $V_{gs} = -V_{DS}/2$  as explained in previous section. When  $V_{GS}$  increases, the NMOS device becomes less reversely biased and  $I_D$  increases towards  $I_{off}$  of a PMOS device with high  $V_{ds}$ . The current slope of the ULP transistor is thus close to half the subthreshold slope ( $1/2S$ ) of the NMOS device, as when  $V_{GS}$  sweeps from 0 to  $V_{dd}$ ,  $V_{gs}$  of the PMOS device inside the ULP transistor sweeps from  $-V_{ds}/2$  to 0 V.

Let us summarize these observations, considering the figures of merit for digital circuits.

- First observation is that  $I_D$  of N-type ULP transistor is always limited by the internal PMOS device, which never leaves the subthreshold regime. It yields low  $I_{off}$  and  $I_{on}$  currents.
- $I_{off}$  of the ULP transistor is the current of devices with  $V_{gs} = -V_{dd}/2$  and  $V_{ds} = V_{dd}/2$ . It leads to ultra-low leakage as devices are reversely biased and DIBL effect is mitigated.
- $I_{on}$  of the ULP transistor is the current of a device with  $V_{gs} = 0$ . This will lead to long delay for ULP logic gates.

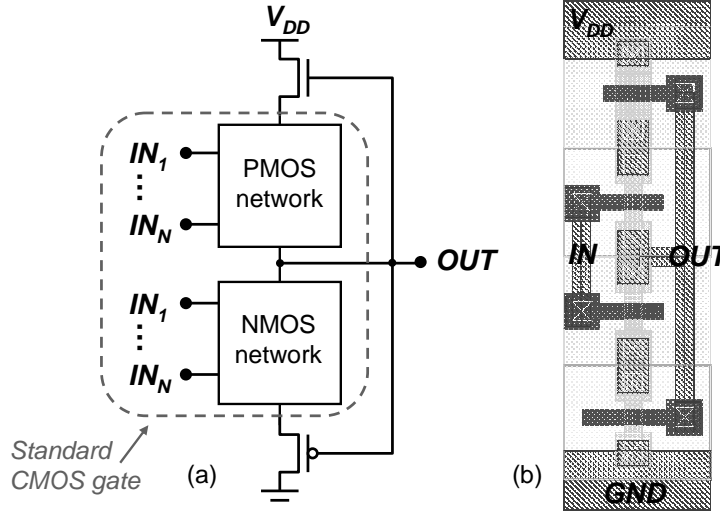


Fig. 5.5. ULP logic style: structure (a) and inverter layout(b)

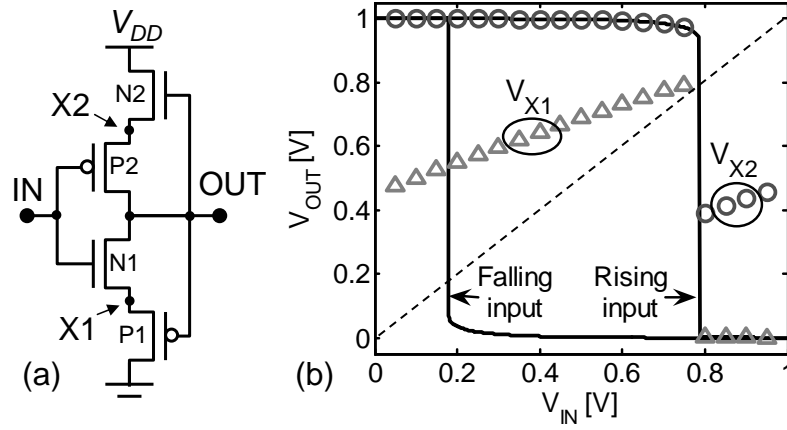
- Regarding logic-gate robustness, it is important to have a good  $I_{on}/I_{off}$  ratio, with  $I_{on}$  taken for a transistor with low  $V_{DS}$  and  $I_{off}$  for a transistor with high  $V_{DS}$ . As shown in Fig. 5.4, this is clearly the case as  $I_D$  of ULP transistor is always a subthreshold current (no saturation) and DIBL is mitigated, even if the current slope vs.  $V_{GS}$  is half the MOSFET subthreshold slope.

## 5.4 ULP LOGIC STYLE

The concept of ULP transistor can be used to build logic gates in what we call the ULP logic style [PA1]. In this section, we first present the ULP logic style architecture and its DC characteristics with the impact of intrinsic device variability. We then evaluate its performances for ULP applications by comparing it to other low-leakage approaches.

### 5.4.1 Architecture and layout

The structure of ULP logic gates is presented in Fig. 5.5(a) [PA1]. It is based on the equivalent standard CMOS gate with addition of 2 extra devices, whose gates are connected to the output node to cut off subthreshold leakage. The header is an NMOS device and the footer is a PMOS device in order to place the NMOS pull-down and PMOS pull-up networks in the configuration of N- and P-type ULP transistors, respectively. The connection of the output node to the gate of the header and footer devices acts as a feedback loop: when the



**Fig. 5.6.** ULP inverter: (a) schematic and (b) simulated voltage transfer curve ( $T=200^{\circ}\text{C}$ ,  $V_{dd}=1\text{ V}$ ,  $W/L = 0.15/0.13\text{ }[\mu\text{m}]$ , except for  $N2$ :  $W/L = 0.3/0.13\text{ }[\mu\text{m}]$ )

inputs are low and the output is high, it cuts off the PMOS footer, thereby biasing the source of the NMOS pull-down network to a positive value and thus achieving negative  $V_{GS}$  self-biasing. As shown in Fig. 5.5(b) for the ULP inverter, SOI implementation leads to compact layout as the footer (resp. header) can be abutted to the pull-down NMOS (resp. pull-up PMOS) and floating-body devices are used to avoid area-consuming body-node connections.

#### 5.4.2 DC behaviour

In order to assess the correct operation of the ULP logic style, let us examine the DC characteristic of the ULP inverter depicted in Fig. 5.6(a). The simulation of the voltage transfer curve at a temperature of  $200^{\circ}\text{C}$  is also plotted in Fig. 5.6(b). We directly notice first that the switching voltage is different when the output node is initially high and when it is initially low, which means that the curve features hysteresis. Furthermore, the ULP inverter shows very good logic levels even with an NMOS as header and a PMOS as footer. To explain these features, let us consider a rising edge of the input voltage. The internal voltages  $V_{X1}$  and  $V_{X2}$  are also plotted in Fig. 5.6 to ease the explanation.

- When  $V_{IN}=0\text{ V}$ ,  $V_{OUT}$  is equal to  $V_{X2}$  because  $P2$  presents a very low impedance thanks to its high  $V_{gs}$ . Moreover,  $N2$  has a low equivalent impedance with its zero  $V_{gs}$  as compared to  $N1$ - $P1$ . Indeed, by symmetry in  $N1$ - $P1$  ULP transistor,  $V_{X1}$  is close to  $V_{dd}/2$  and thus  $N1$  and  $P1$  devices inside N-type ULP transistor have  $V_{gs}$  close to  $-V_{dd}/2$ , as explained in Section 5.3.2. Therefore,  $V_{X2}$  and thus  $V_{OUT}$  are both equal to  $V_{dd}$ . From Fig. 5.4, this is confirmed by the  $I_{on}/I_{off}$  ratio of nearly 1000 between an ULP transistor with  $V_{GS} = V_{dd}$  and  $V_{GS} = 0$ .

- When  $V_{IN}$  increases, the symmetry in  $N1$ - $P1$  ULP transistor implies that  $V_{X1}$  is close to  $(V_{IN} + V_{OUT})/2$ .  $N1$  and  $P1$  still have negative  $V_{gs}$  so that  $V_{OUT}$  remains close to  $V_{X2}$  and thus to  $V_{dd}$ .
- When  $V_{IN}$  gets closer to  $V_{X1}$ ,  $V_{gs}$  of  $N1$  and  $P1$  is nearly 0. The equivalent impedance of  $N2$  is no longer negligible as compared to  $N1$  and  $P1$ .  $V_{OUT}$  is still equal to  $V_{X2}$  but  $V_{X2}$  decreases somewhat. As  $V_{OUT}$  decreases,  $V_{gs}$  of  $P1$  becomes less reversely biased.
- Finally, the rise of  $V_{IN}$  combined with the fall of  $V_{OUT}$  makes the gate switch. Indeed  $N1$  leaves the subthreshold regime and discharges  $V_{OUT}$  to  $V_{X1}$ . Moreover this fall of  $V_{OUT}$  increases  $|V_{gs}|$  of  $P1$  and decreases  $V_{gs}$  of  $N2$ . The  $I_{on}$  current is now the subthreshold current of  $P1$ , which increases exponentially as  $|V_{gs}|$  is rising. The output connection thus works as a positive feedback loop and leads to a sharp transition of  $V_{OUT}$ .

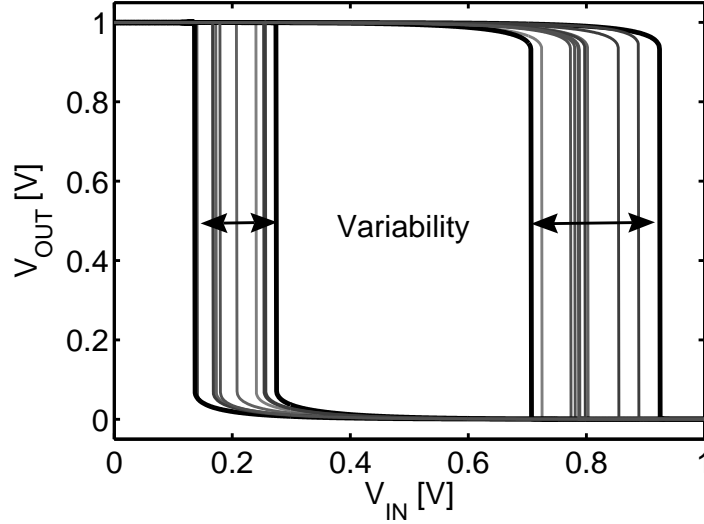
Similar reasoning can be applied to the case of a falling input. The switching voltages for rising ( $V_{IH}$ ) and falling ( $V_{IL}$ ) input edges are not exactly symmetrical regarding  $V_{dd}/2$  because at 200°C subthreshold currents of NMOS and PMOS devices are not identical in the considered technology. Under 1 V  $V_{dd}$  with all minimum-sized devices ( $W/L = 0.15/0.13 [\mu m]$ ), the switching voltages are 0.75 and 0.15 V. The width of NMOS  $N2$  header, which has a somewhat lower subthreshold current than PMOS  $P1$  footer due to process imbalance, can be upsized to  $0.3 \mu m$  to balance switching voltages without area overhead, as shown in Fig. 5.5(b). This leads to 0.79 and 0.22 V switching voltages.

Thanks to the good output logic levels ( $V_{OL}$  and  $V_{OH}$ ) and the hysteresis, ULP logic style features very high SNM, higher than  $V_{dd}/2$ , which is never achieved with standard CMOS logic styles:  $SNM_H = V_{IH} - V_{OL} = 0.79$  V and  $SNM_L = V_{OH} - V_{IL} = 0.78$  V, at  $V_{dd} = 1$  V. This leads to excellent noise robustness.

All logic functions from standard CMOS logic gates can be implemented in ULP logic style. In standard CMOS logic style, the width of 2 or 3 stacked devices in the pull-up or the pull-down network is multiplied by 2 or 3 in order to keep roughly equivalent  $I_{on}$  for pull-up and pull-down networks. The same principle can be applied to pull-up and pull-down networks inside ULP logic gates. The header and footer devices limit  $I_{off}$  and their width are thus kept constant for static power concern:  $0.3 \mu m$  and  $0.15 \mu m$ , respectively. For an ULP NAND3 gate, this leads to very symmetric SNM:  $SNM_L = 0.78$  V and  $SNM_H = 0.79$  V. For an ULP NOR3 gate, SNM are less symmetric because of the inherent process imbalance between NMOS and PMOS devices in the considered technology:  $SNM_L = 0.72$  V and  $SNM_H = 0.88$  V.

### 5.4.3 Impact of intrinsic variability on robustness

When scaling CMOS devices down, the mismatch on device features increases due to intrinsic variability [10]. For sensible structures such as SRAM cells, this mismatch could degrade the SNM and leads to functional breakdown and we



**Fig. 5.7.** Monte-Carlo simulated ULP inverter voltage transfer curve ( $T=200^{\circ}C$ ,  $V_{dd} = 1$  V,  $W/L = 0.15/0.13$  [ $\mu m$ ], except for  $N2$ :  $W/L = 0.3/0.13$  [ $\mu m$ ])

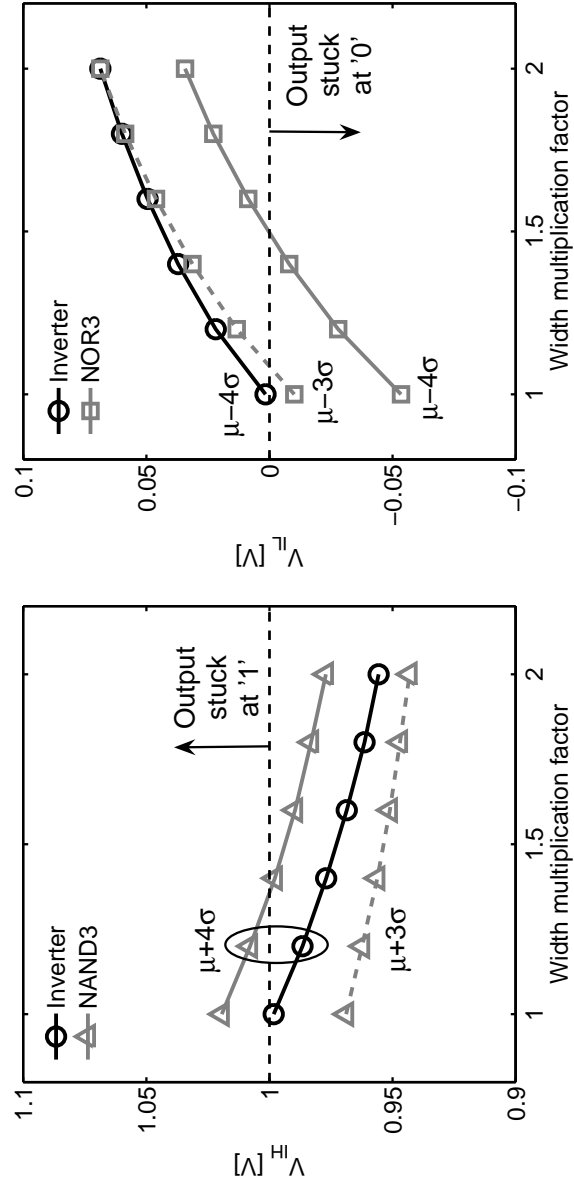
showed in Chapter 2 that logic gates operating in subthreshold regime may also suffer from bad SNM because of low  $I_{on}/I_{off}$  ratio and magnified sensitivity against device variability.

Despite their standard  $V_{dd}$ , the devices inside ULP logic gates also operate in the subthreshold regime because of the negative  $V_{gs}$  self-biasing. As high-temperature operation degrades  $I_{on}/I_{off}$  ratio, it is thus necessary to study the robustness of ULP logic gates under intrinsic device variability.

Monte-Carlo simulation of the ULP inverter voltage transfer curve is shown in Fig. 5.7. Mismatch parameters provided by the foundry are considered ( $\sigma_{V_t}=34$  mV for minimum-sized devices,  $\sigma_{L_{eff}}=4$  nm,  $\sigma_W=10$  nm and  $\sigma_{T_{ox}}=0.033$  nm). Fig. 5.7 shows that variability does not degrade output logic levels thanks to the good  $I_{on}/I_{off}$  ratio of ULP transistor (Section 5.3.3). However, the switching voltages are influenced by intrinsic variability. This could be an issue if switching voltages get too close to 0 or  $V_{dd}$  levels. Indeed, a  $V_{IH}$  above  $V_{dd}$  for example means that the logic gate has its output stuck at low logic level as the voltage seen at its input is never recognized as a high logic level.

Amongst all variability sources,  $V_t$  variability has the strongest impact on subthreshold current [11], as it exponentially depends on  $V_t$ . As  $V_t$  variability due to random doping fluctuation is inversely proportional to  $\sqrt{WL}$  [10], it can be efficiently mitigated by increasing the width of the devices. Notice that length upsize is not practical for ULP logic style as it decreases  $I_{on}$ , which is already very low. Monte-Carlo simulations were carried out to extract the switching voltages





**Fig. 5.8.** Simulated worst-case switching voltages of ULP logic gates vs. device width ( $T=200^\circ C$ ,  $V_{dd}=1 V$ , logic gates are sized as described in Section 5.4.2 and the device widths are increased altogether by the common multiplication factor)

of ULP inverter, ULP NAND3 and ULP NOR3 logic gates, when increasing the width of all devices by a common factor. Worst-case switching voltages are shown in Fig. 5.8. In this figure, we first observe that ULP NAND3 has the worst  $V_{IH}$  and that ULP NOR3 has the worst  $V_{IL}$ . For minimum-sized NAND3 gate,  $4\sigma$  worst-case  $V_{IH}$  is higher than  $V_{dd}$ , which leads to several gates with output stuck at high logic level. Increasing the width of all devices by a factor 1.4 efficiently improves worst-case  $V_{IH}$  variability, which drops below  $V_{dd}$ , thereby ensuring correct operation. Same reasoning can be applied to NOR3 gate, which requires an increase of all devices by a factor 1.5 in order to raise worst-case  $V_{IL}$  above 0V. As NAND3 and NOR3 gates are worst-case choice for analysis of  $V_{IH}$  and  $V_{IL}$  respectively, this demonstrates that safe operation of ULP logic style can be achieved by small device width upsize. As minimum-sized devices are hardly used in high-temperature circuits, this upsize is fairly affordable.

#### 5.4.4 Performance evaluation

In order to assess the efficiency of ULP logic style, performances are compared with standard CMOS logic style at gate level through simulation of an inverter with fan-out of 4 (FO4). The considered technology is a dual- $V_t$  technology. Low- $V_t$  devices ( $V_t=0.34\text{V}$ ) are used for ULP logic style to increase drivability and high- $V_t$  devices ( $V_t=0.43\text{V}$ ) are used for standard CMOS logic style to lower static current. Simulation results are summarized in Table 5.1.

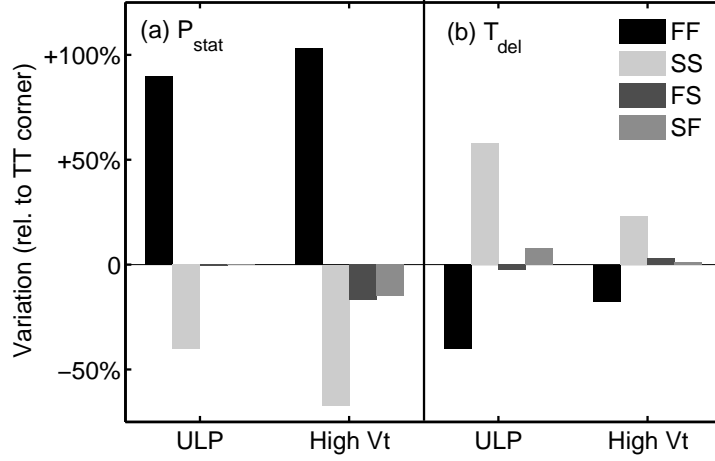
Under 1V  $V_{dd}$ , ULP inverter provides static power reduction by nearly 3 orders of magnitude as compared to standard high- $V_t$  CMOS inverter. As a result, the instantaneous total power consumption  $P_{inst}$  is also very low lying in the pW range at the expense of longer delay  $T_{del}$ . Notice that the 34ns delay for FO4 inverter is sufficient to support low operation throughputs for ULP applications. The power consumption of high- $V_t$  CMOS inverter at 10k, 100k and 1MHz lies in the nW-range, being dominated by static power. Notice also that the difference between static and total power of high- $V_t$  standard CMOS inverter is not only due to the switching of FO4 load capacitance but also to the short-circuit current during a transition and the transient behavior of subthreshold leakage in floating-body devices.

As shown in Table 5.1, lowering the supply voltage of high- $V_t$  CMOS inverter to 0.5V yields significant  $P_{inst}$  reduction but it remains in the nW range, still dominated by static power. In order to provide a wider comparison, standard CMOS inverter has also been simulated in a  $1\mu\text{m}$  partially-depleted SOI technology dedicated to high-temperature operation (body-tied devices with  $T_{ox}=25\text{nm}$ ,  $T_{Si}=250\text{nm}$  and  $V_t=1.6\text{V}$ ). With this technology, static power is efficiently reduced but remains  $30\times$  higher than for ULP inverter in  $0.13\mu\text{m}$ . Moreover, there are large area, delay and dynamic-power overheads. This clearly shows the benefit of using ULP logic style in scaled standard technologies.

**Table 5.1.** Comparison of FO4 inverter performances in partially-depleted SOI technology (T=200°C)

Inverter type	Process [ $\mu m$ ]	$V_{dd}$ [V]	Area [ $\mu m^2$ ]	$T_{del}$ [ns]	$P_{stat}$ [nW]	$P_{inst}$ @ 10 kHz [nW]	$P_{inst}$ @ 100 kHz [nW]	$P_{inst}$ @ 1 MHz [nW]
ULP low- $V_t$	0.13	1	5.4	34.6	0.031	0.041	0.17	0.81
CMOS high- $V_t$	0.13	1	2.6	0.072	26.8	27.7	30.1	36.4
CMOS high- $V_t$	0.13	0.5	2.6	0.31	6.8	7.1	7.7	8.4
CMOS high- $V_t$	1	1	320	510	0.98	1.3	3.5	52.6 <sup>†</sup>

<sup>†</sup> Simulated at 1.23V in order to support 1 MHz (delay  $\approx$  100 ns).



**Fig. 5.9.** Performance variation of ULP and high- $V_t$  standard CMOS inverters against process variations: (a) static power and (b) delay ( $T=200^\circ C, V_{dd}=1V$ , corner denomination: T=typical, F=fast, S=slow, first letter is for NMOS and second for PMOS)

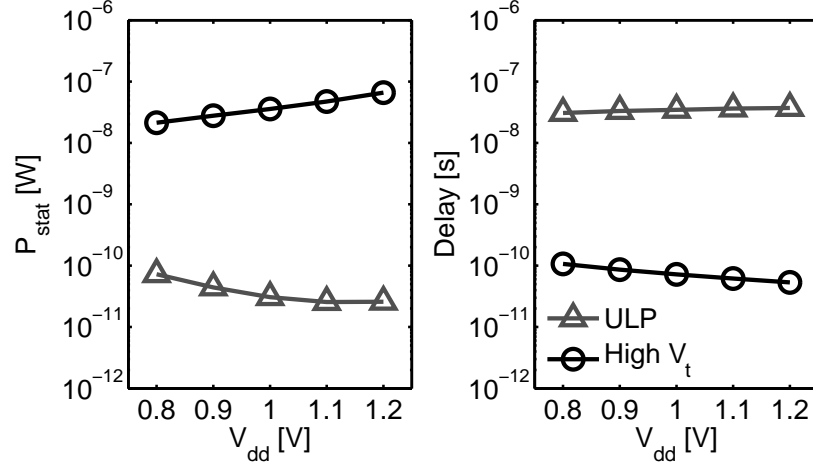
## 5.5 IMPACT OF PVT VARIATIONS ON PERFORMANCES

In this section, performances of the ULP inverter are simulated against process, voltage and temperature variations in order to validate the reliability of ULP logic style.

### 5.5.1 Process variations

In scaled CMOS technologies, global process variations can strongly affect performances of digital circuits. In SOI technology, the use of adaptive body-biasing (ABB) technique has a large die-area overhead as the body of each device needs an independent bias connection, which implies difficult routing in addition to device-area overhead. Moreover, the high-temperature operation increases junction leakage, which makes body biasing undesirable for static power concern. As circuits thus cannot rely on ABB, it is necessary to investigate performance stability under global process variations.

Fig. 5.9 shows the simulated variations in static power and delay of ULP and high- $V_t$  standard CMOS inverter, with global process corners. At FF (Fast NMOS, Fast PMOS) and SS (Slow NMOS, Slow PMOS) corners, static power is more stable for ULP than for high- $V_t$  inverter. At FS and SF crossed corners, static power of ULP inverter remains roughly constant because pull-up and pull-down networks of ULP inverter are composed of an ULP transistor, i.e. with both NMOS and PMOS devices. Variations of ULP transistor  $I_{off}$  are thus

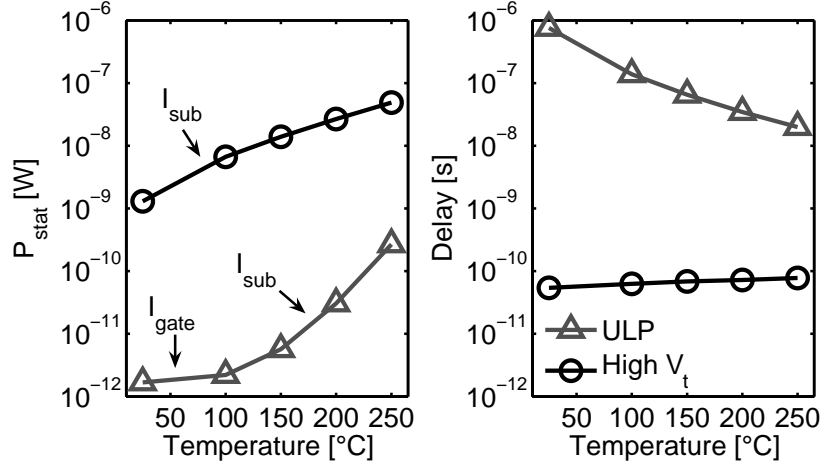


**Fig. 5.10.** Performances of ULP and high- $V_t$  standard CMOS inverters against supply voltage variations ( $T=200^\circ C$ )

mitigated at crossed corners. Moreover, this also leads to weak sensitivity of SNM against crossed corners: simulations show 5%  $SNM_H$  (resp. 3%  $SNM_L$ ) maximum deviations at worst-case FS (resp. SF) corner, with minor impact on robustness results from Section 5.4.3. Delay of the ULP inverter varies more with global process corners than the delay of high- $V_t$  inverter because ULP transistor  $I_{on}$  is a subthreshold current, as reported in Section 5.3.3, which is thus more sensitive to  $V_t$  variations.

### 5.5.2 Voltage variations

Performance dependence on voltage variations is also important for robustness and scalability. Simulations of the ULP inverter show less than 5% deviation in the SNM normalized to  $V_{dd}$  for voltage from 0.8V to 1.2V. The ULP inverter is thus scalable and robust. Performances of ULP inverter are plotted vs.  $V_{dd}$  in Fig. 5.10 and compared to high- $V_t$  standard CMOS inverter. As shown in Fig. 5.3, lowering  $V_{dd}$  leads to higher ULP transistor  $I_{off}$  and thus small static power increase of ULP inverter. As explained in Section 5.3.3, ULP transistor  $I_{on}$  is the current of a device with  $V_{gs}=0V$ , which has a small dependence on  $V_{dd}$ . Moreover, when lowering  $V_{dd}$ , the output swing of an inverter is reduced. This leads to a reduction of the ULP inverter delay, proportional to  $C_L V_{dd} / I_{on}$  when lowering  $V_{dd}$ . Both static power and delay of ULP inverters are very stable against supply voltage variations.



**Fig. 5.11.** Performance of ULP and high- $V_t$  standard CMOS inverters against temperature variations ( $V_{dd}=1V$ )

### 5.5.3 Temperature variations

Simulation of the ULP inverter for a wide temperature range from room temperature to  $250^\circ C$  show that its static noise margins are modified by less than 6%. Static power and delay for various temperatures are shown in Fig. 5.11 and compared to high- $V_t$  standard CMOS inverter. Delay of the ULP inverter decreases with temperature because ULP transistor  $I_{on}$  is a subthreshold current, which increases with temperature. At room temperature, ULP logic style is significantly slower. The maximum operating frequency at room temperature is in the range of 40 to 60 kHz. Static power of ULP inverter remains roughly constant for temperatures lower than  $100^\circ C$ . As shown in Fig. 5.3, ULP transistor  $I_{off}$  is dominated at these temperatures by gate leakage, which features low temperature dependence. At  $150^\circ C$ , subthreshold leakage becomes dominant and static power starts to increase. The increase is more important than for high- $V_t$  inverter because of the factor 2 in front of  $S$  ( $\sim U_{th}$ ) term in Eq. (5.3). For the whole temperature range, static power of ULP inverter remains lower by at least two orders of magnitude.

## 5.6 VALIDATION OF ULP LOGIC STYLE

In order to further validate the efficiency of ULP logic style, we present measurement results of a ring-oscillator test vehicle and simulations of a benchmark multiplier. We then qualitatively discuss the efficiency of ULP logic style as compared to other leakage-reduction techniques in the context of high-temperature applications.

**Table 5.2.** Comparison of ULP transistor currents in 0.13  $\mu\text{m}$  bulk technology ( $W/L = 1/0.13 [\mu\text{m}]$ ,  $V_{dd}=0.5 \text{ V}$ ,  $T=25^\circ\text{C}$ )

Source	$I_{on} [\text{nA}]$	$I_{off} [\text{pA}]$	$I_{on}/I_{off}$
TT simulation	0.40	1.40 <sup>†</sup>	290
SS simulation	0.15	1.29 <sup>†</sup>	110
Measurement	0.17	0.14	1210

<sup>†</sup> Dominated by junction leakage (overestimated).**Table 5.3.** Comparison of ULP ring-oscillator performances in 0.13  $\mu\text{m}$  bulk technology ( $W/L = 0.15/0.13 [\mu\text{m}]$ ,  $V_{dd}=0.5 \text{ V}$ ,  $T=25^\circ\text{C}$ )

Source	$V_{IL} [\text{V}]$	$V_{IH} [\text{V}]$	$P_{stat}^* [\text{pW}]$	$T_{del}^* [\mu\text{s}]$
TT simulation	0.07	0.40	1.20 <sup>†</sup>	1.69
SS simulation	0.07	0.41	1.12 <sup>†</sup>	4.15
Measurement	0.09	0.38	0.065	11.2

<sup>†</sup> Dominated by junction leakage (overestimated).

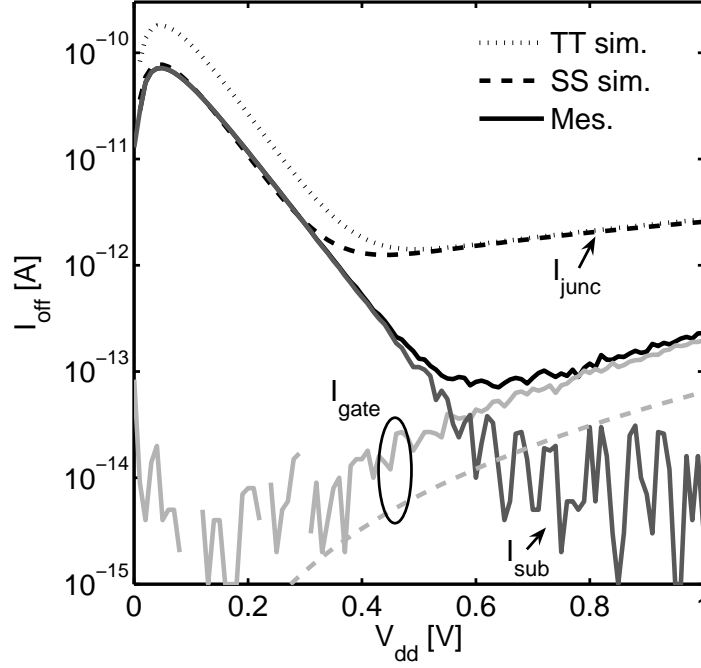
\*Expressed for one inverter.

### 5.6.1 Measurement of ring-oscillator test vehicle

A ring oscillator based on ULP inverters has been manufactured in 0.13  $\mu\text{m}$  bulk technology to demonstrate the feasibility of ULP logic style. Notice that only room-temperature measurements were carried out as the considered bulk technology cannot operate at high temperature due to latch-up issues.

Let us first present the characteristics of manufactured ULP transistors. Table 5.2 shows the  $I_{on}$  and  $I_{off}$  currents of ULP transistor from both measurements and simulations (industrial BSIM3 compact models). It shows that measured  $I_{on}$  current is close to simulation at SS corner, while measured  $I_{off}$  is much lower than simulated one. As shown in Fig. 5.12, this comes from junction leakage, which is dramatically overestimated in the MOSFET compact models of this technology. Gate leakage thus dominates  $I_{off}$  at  $V_{dd}$  values above 0.5 V.

DC measurement of ULP inverters has been carried out. Measured voltage-transfer curve is presented in Fig. 5.13, demonstrating the hysteresis property of ULP logic style. Finally, measured performances of a 53-stage ULP-inverter ring oscillator are given. Once more, static power is much lower in measurement than in simulation because of junction-leakage overestimation in MOSFET compact models of this technology. Ultra-low-leakage property is demonstrated. Measured delay is longer than simulated one due to routing capacitances not taken into account in the simulations.



**Fig. 5.12.** Measured  $I_{off}$  of ULP transistor in  $0.13\ \mu\text{m}$  bulk technology ( $W/L = 1/0.13\ [\mu\text{m}]$ ,  $V_{dd}=0.5\ \text{V}$ ,  $T=25^\circ\text{C}$ ). The fabricated lot fits simulations at SS corner, junction leakage is dramatically overestimated in simulations, the dominating contribution to  $I_{off}$  of ULP transistor for  $V_{dd} > 0.5\ \text{V}$  is gate leakage.

### 5.6.2 Simulation of a benchmark multiplier

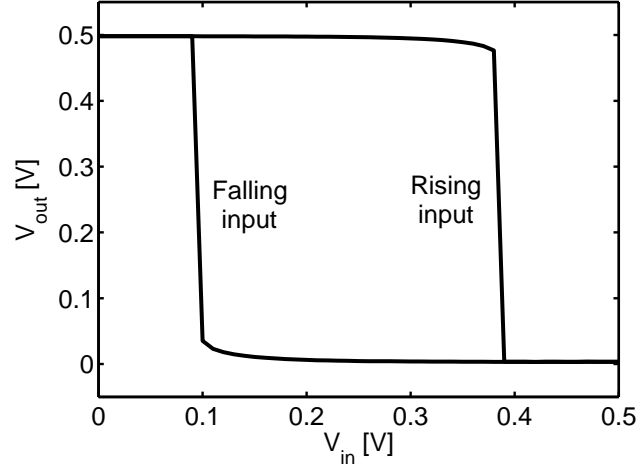
In order to validate the results from Section 5.4.4 at circuit level, we carried out simulations of the 8-bit RCA benchmark multiplier from previous chapters build with ULP logic gates in  $0.13\ \mu\text{m}$  PD SOI technology at  $200^\circ\text{C}$ , with the compact models validated in Section 5.3.2. Simulated performances are given in Table 5.4. This fully validates the static-power reduction by 3 orders of magnitude as well as the operation up to  $800\ \text{kOp/s}$ .

### 5.6.3 Comparison with other leakage-reduction techniques

Let us compare ULP logic style to reverse body biasing and power-supply gating in the context of high-temperature applications.

First, we showed in [CP1] that reverse body biasing is less efficient than ULP logic style at room temperature in  $0.13\ \mu\text{m}$  bulk technology. Moreover, in SOI technology, the application of a reverse body bias is not as straightforward as





**Fig. 5.13.** Measured voltage transfer curve of ULP inverter in 0.13  $\mu\text{m}$  bulk technology ( $W/L = 0.15/0.13 [\mu\text{m}]$ ,  $V_{dd}=0.5\text{ V}$ ,  $T=25^\circ\text{C}$ )

**Table 5.4.** Simulation of the 8-bit RCA multiplier in 0.13  $\mu\text{m}$  SOI technology ( $T=200^\circ\text{C}$ )

Logic style	$V_{dd} [\text{V}]$	$T_{del}$	$P_{stat}$	$P_{inst}$ @10kOp/s	$P_{inst}$ @100kOp/s
ULP low- $V_t$	1	$1.22 \mu\text{s}$	$14.5 \text{ nW}$	$240 \text{ nW}$	$380 \text{ nW}$
CMOS high- $V_t$	1	$3.55 \text{ ns}$	$15.9 \mu\text{W}$	$19.2 \mu\text{W}$	$20.8 \mu\text{W}$
CMOS high- $V_t$	0.5	$16.5 \text{ ns}$	$4.2 \mu\text{W}$	$5.1 \mu\text{W}$	$5.4 \mu\text{W}$

in bulk, because the body of each device has to be biased by an independent connection with a serious global routing overhead. At high temperature, reverse body biasing also leads to prohibitive junction leakage overhead, according to its high-temperature dependence ( $50\text{-}100\times/100^\circ\text{C}$ ) [12]. This makes reverse body biasing not practical in high-temperature circuits.

Secondly, sleep-mode power-supply gating [13], rely on the availability of high- $V_t$  devices. As high- $V_t$  devices are already considered for standard CMOS gates in the context of high-temperature applications, power gating is less efficient: it would only rely on a leakage limitation from the smaller width of the sleep device. Moreover, sequential elements (latches and flip-flops) have to hold circuit state and thus cannot enter sleep mode. Therefore, even if static power of combinatorial logic is reduced to the level of ULP logic, static power of sequential elements, which is higher by 3 orders of magnitude, would completely mask the associated benefit.

## 5.7 CONCLUSION

In this chapter, we proposed to build an ULP transistor with 2 standard CMOS devices to achieve ultra-low  $I_{off}$  current, even at high-temperature. In  $0.13\mu\text{m}$  SOI technology, the use of ULP transistor reduces  $I_{off}$  by more than 4 and 3 orders of magnitude at room temperature and  $200^\circ\text{C}$ , respectively. In order for a standard NMOS device to provide an  $I_{off}$  current as low as an ULP inverter, its  $V_t$  has to be raised up to  $0.8\text{V}$ , which is hardly proposed in scaled standard technologies.

The building of an ULP logic style based on the ULP transistor was proposed to drastically reduce power consumption of digital circuits for ULP high-temperature applications. The ULP inverter exhibit a power consumption at low clock frequency reduced by 3 orders of magnitude, as compared to standard CMOS inverter with high- $V_t$  devices, at the expense of delay and area overheads, which was demonstrated by measurement of a test vehicle. This huge power benefit is kept even when considering process, voltage and temperature variations. Moreover, ULP logic gates feature excellent noise robustness thanks to SNM higher than  $V_{dd}/2$ , which is never achieved in standard CMOS logic style.

The ULP logic style allows digital circuits to benefit from small die area and small dynamic power of scaled standard technologies while keeping ultra-low-leakage, even at high temperature. It is thus a unique and straightforward technique to design ultra-low-power circuits for high-temperature applications, without neither extra mask nor process cost.

Finally, notice that hysteresis is an interesting property for improving static noise margins (SNM) of SRAM cells [15]. The building of a 12-transistor SRAM cell based on ULP transistors and ULP inverters has been proposed in [CO7] and [16], with leakage current comparable to the 7-transistor SRAM cell based on ULP diodes from [8, 9] and SNM as high as  $0.7\text{V}$  at room temperature, under  $1\text{V}$   $V_{dd}$  in  $0.13\mu\text{m}$  bulk technology. The use of an internal read buffer makes its speed performance independent from the long delay of ULP inverters, the drawback being die area penalty.

## REFERENCES

1. D. Flandre, "Silicon-on-insulator technology for high temperature metal oxide semiconductor devices and circuits", in *High-Temperature Electronics*, IEEE Press, Ed. R. Kirschman, pp. 303-308, 1998.
2. L. Vancaillie, V. Kilchytska, P. Delatte, L. Demeus, H. Matsushashi, F. Ichikawa and Denis Flandre, "Peculiarities of the temperature behavior of SOI MOSFETs in the deep submicron area", in *Proc. IEEE Int. SOI Conf.*, pp.78-79, 2003.
3. D. Flandre *et al.*, "Fully depleted SOI CMOS technology for heterogeneous micropower, high-temperature or RF microsystems", in *Solid-State Electronics*, vol. 45, no. 4, pp. 541-549, 2001.
4. B. W. Ohme, "Development of SOI CMOS electronics under the DeepTrek program", in *Proc. Int. Conf. High Temperature Electronics*, pp. 25-33, 2003.
5. V. Dessard, *SOI Specific Analog Techniques for Low-Noise, High-Temperature or Ultra-Low Power Circuits*, Ph.D dissertation, Université catholique de Louvain, 2001.
6. A. Adriaensen V. Dessard and D. Flandre, "25 to 300°C ultra-low-power voltage reference compatible with standard SOI CMOS process", in *Electronics Letters*, vol. 38, no. 19, pp. 1103-1104, Sep. 2002.
7. D. Levacq, C. Liber, V. Dessard and D. Flandre, "Composite ULP diode fabrication, modelling and applications in multi- $V_{th}$  FD SOI CMOS technology", in *Solid-State Electronics*, vol. 48, no. 6, pp. 1017-1025, Jun. 2004.
8. D. Levacq, *Low leakage SOI CMOS circuits based on the ultra-low power diode concept*, Ph.D dissertation, Université catholique de Louvain, 2006.
9. D. Levacq, V. Dessard and D. Flandre, "Low leakage SOI CMOS static memory cell with ultra-low power diode", in *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 689-702, Mar. 2007.
10. A. Asenov, A.R. Brown, J.H. Davies, S. Kaya and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 50, no. 9, pp. 1837-1852, Sep. 2003.
11. B. Zhai, S. Hanson, D. Blauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *Proc. IEEE/ACM Int. Symp. Low-Power Electronics Des.*, pp. 20-25, 1995.
12. L. T. Clark, R. Patel and T. S. Beatty, "Managing standby and active mode leakage power in deep sub-micron design," in *Proc. IEEE/ACM Int. Symp. Low-Power Electronics Des.*, pp. 274-279, 2004.
13. S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu and J. Yamada, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS", in *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 847-854, Aug. 1995.
14. K. Bowman, S. Duval and J. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration", in *IEEE J. Solid-State Circuits*, vol. 37, no. 2, pp. 183-190, Feb. 2002.

15. J. P. Kulkarni, K. Kim and K. Roy, "A 160 mV robust Schmitt trigger based subthreshold SRAM", in *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2303-2312, Oct. 2007.
16. J. De Vos, *Développement de circuits mémoire à ultra-basse consommation pour applications portables en technologie bulk 130 nm*, M.Sc. dissertation, Université catholique de Louvain, 2008.

## CONCLUSIONS AND PERSPECTIVES

---

Ultra-low-power (ULP) applications such as RFID tags, biomedical devices and sensor networks are an emerging field of the IC market. Thanks to low computational load, their energy consumption can be tremendously reduced by jointly scaling the clock frequency  $f_{clk}$  and the supply voltage  $V_{dd}$  down to the limits drawn by robustness and throughput constraints dictated by the application. In such a frequency/voltage-scaled (FVS) CMOS circuit, when neglecting these constraints, energy per operation  $E_{op}$  is minimized when operating at a particular  $V_{dd}/f_{clk}$  point that balances dynamic and static contributions to  $E_{op}$ . This minimum-energy point often lies in MOSFET subthreshold region where both on- and off-state currents exponentially depends on gate bias and on the threshold voltage  $V_t$ .

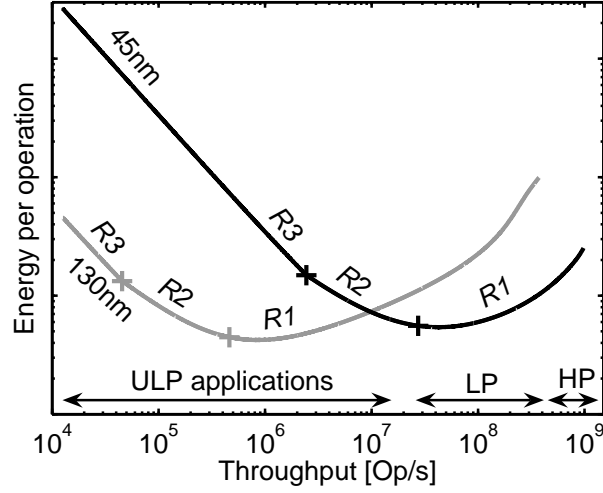
CMOS technology scaling driven by Moore's law leads to nanometer MOSFETs with reduced capacitances and thus reduced dynamic energy per operation, at the expense of increased leakage currents and device variability. The impact of these drawbacks is magnified when considering FVS subthreshold circuits for ULP applications. First, their low  $f_{clk}$  makes leakage-induced static energy proportionally more important. Second, their low  $V_{dd}$  increases the sensitivity against variability due to the exponential dependence of subthreshold current on  $V_t$ . In this dissertation, we therefore raised two questions:

- What is the impact of nanometer CMOS technology scaling on ultra-low-power digital circuits ?
- How to benefit from the circuit size reduction while keeping robustness and energy consumption under control ?

Trying to answer these questions is a whole expedition that we would like to summarize here.

### *Studying the trail map*

In order to clarify the situation, we first set up a strong theoretical framework to support the analysis of practical energy efficiency of FVS circuits under robustness and throughput constraints. It allows a unified representation from nominal- $V_{dd}$  circuits for high-performance (HP) applications to subthreshold circuits for ULP applications. The throughput range associated to the application spectrum can be divided into three regions: energy-efficient  $R1$  region where dynamic component dominates, energy-inefficient  $R2$  region where static component dominates and minimum  $V_{dd}$  is throughput-limited, and energy-inefficient  $R3$  region where static component dominates and minimum  $V_{dd}$  is robustness-limited. As shown in Fig. O.1 in a  $0.13\ \mu\text{m}$  deep-submicron technology, circuits for HP and



**Fig. O.1.** Impact of technology scaling on practical energy per operation under robustness and throughput constraints (at room temperature)

LP applications lies in  $R1$  region, while circuits for ULP applications may lie in any of these regions as the throughput range of ULP applications is quite broad ( $\approx 10\text{ k} - 10\text{ MOp/s}$ ). We used this framework throughout the dissertation as a compass to guide our analysis of technology/device/circuit considerations.

As an itinerary-planning analysis, we then thoroughly investigated the impact of technology scaling on subthreshold logic in two steps. First at device level, we analyzed its impact on MOSFET subthreshold operation. It shows that worst-case subthreshold  $I_{on}$  increases with constant-field scaling trend until 90 nm node and then saturates because of subthreshold swing, drain-induced barrier lowering (DIBL) and variability increase. Fringing capacitances due to slow scaling of gate-stack height also exhibit a worrying increase. Second, at circuit level, we showed that robustness-limited minimum- $V_{dd}$  dramatically increases while throughput-limited minimum- $V_{dd}$  decreases. The consequence for a given application with fixed robustness and throughput constraints is a jump of minimum  $V_{dd}$  from throughput to robustness limitation i.e. from  $R1/R2$  to  $R3$  throughput region, when migrating to nanometer technologies. We reported that minimum-energy level  $E_{min}$  is reduced when reaching 90 nm node thanks to dynamic energy reduction but then increases as static energy does at 65/45 nm nodes. As shown in Fig. O.1, technology scaling shifts minimum-energy point towards higher throughput values, which tends to enable minimum-energy operation beyond the restricted scope of pure ULP applications. From this figure, we also observed that, although technology scaling highly benefits to HP and LP applications where dynamic energy dominates, it is severely detrimental for ULP

applications. Indeed, in 45 nm technology, circuits for ULP applications mainly lie in energy-inefficient  $R2/R3$  regions with orders-of-magnitude energy penalty. We indicated that there is an optimum technology nodes that minimizes practical energy at a given target application throughput. Furthermore, high-temperature operation ( $> 150^\circ\text{C}$ ) in industrial environment makes things much worst by increasing the leakage currents by orders of magnitude, even when considering an SOI 0.13  $\mu\text{m}$  technology.

### *Climbing the rocks*

From these observations, we tried to cross three obstacles: the increase of minimum-energy level at 65/45 nm node, the bad energy efficiency in nanometer  $R3/R2$ -regions subthreshold circuits and the dramatical leakage increase at high temperature.

First, to understand the increase of minimum-energy level in nanometer technologies, we analyzed in depth the effects of nanometer MOSFETs on minimum-energy point. We showed that beyond previously-reported subthreshold swing, load capacitance and variability effects, minimum-energy level also suffers from an extra penalty up to 50% due to the increased DIBL effect and high  $I_{gate}/I_{sub}$  ratio in nanometer MOSFETs. To solve the  $E_{min}$  increase between 90 nm and 45 nm nodes that we reported, we proposed an optimum MOSFET selection at 45 nm node, which favors thin-oxide low- $V_t$  devices with an upsized gate length by 15/25 nm, with slight device area overheads leading to negligible die area overhead at circuit level. This selection yields 40%  $E_{min}$  reduction, brings  $E_{min}$  at 45 nm node back to its corresponding level at 90 nm node. This study reveals a new - *a priori* counter-intuitive - paradigm in device optimization towards ultimate minimum-energy subthreshold circuits. It indicates that efforts should be devoted to minimizing subthreshold swing, DIBL and variability, while tolerating gate leakage increase provided that it remains below the subthreshold leakage level. Moreover, we showed that undoped-channel MOSFETs in fully-depleted SOI technology can bring 60%  $E_{min}$  improvement at minimum gate length, which is much more than the typical claim of 15-30% energy reduction of SOI technology for nominal- $V_{dd}$  high-performance circuits. This makes fully-depleted SOI a highly energy-efficient technology.

Second, we revisited typical circuit design choices, in the light of nanometer subthreshold circuits, with the goal to make practical  $E_{op}$  meet  $E_{min}$ , i.e. to shift minimum-energy point to the application target throughput. At 45 nm node, we showed that, thanks to the versatility of nanometer technologies (multiple technology flavors with multi- $V_t$  devices), an appropriate technology/device selection is able to shift the minimum-energy point over nearly the whole throughput range of ULP applications. Nevertheless, we demonstrated that an independent device-type assignment to logic gates (dual- $V_t/T_{ox}/L_g$  techniques) is not feasible in nanometer subthreshold circuits because of high variability of short paths and large delay difference between subthreshold logic gates with different device types. We also showed that adaptive reverse body biasing (with a negative voltage) can be used for compensation of modeling errors of global pro-

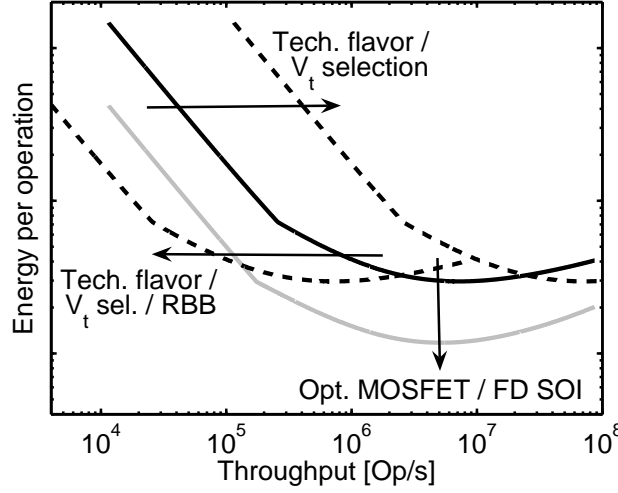
cess/temperature variations to keep minimum-energy point at target throughput and avoiding prohibitive design margins. On the contrary, forward body biasing should be avoided because of  $E_{min}$  penalty and bad behavior with discrete bias voltage values. Moreover at 45 nm node, we pointed out that reverse body biasing is only efficient in low-power technology flavor and we suggested that at next nodes it may no longer be practical because of decreasing body-bias coefficient and increasing band-to-band tunneling leakage. This re-emphasizes the need for technologies with low sensitivity against process/temperature variations in nanometer subthreshold logic. Finally, we showed that sleep-mode techniques (dynamic reverse body biasing and power gating) are inefficient to reduce active leakage when the circuit lies in  $R2/R3$  regions due to mode-transition energy overheads. However, power gating is very efficient to cut off leakage during stand-by periods of nanometer subthreshold circuits although special care has to be taken to ensure circuit robustness when engineering the power switch.

Third, we reported that the two-orders-of-magnitude higher leakage currents in high-temperature environments ( $> 150^\circ\text{C}$ ) completely dominates power consumption in low-to-medium-throughput ULP applications, even when using a  $0.13\ \mu\text{m}$  SOI (partially-depleted) technology. No technology option nor existing leakage-reduction techniques can solve this issue and we thus proposed a new logic style named Ultra-Low-Power (ULP) to reduce leakage by negative  $V_{gs}$  self-biasing. It allows to benefit from the small area and low dynamic power of scaled technologies while keeping leakage currents under control, even at high temperature. In  $0.13\ \mu\text{m}$  partially-depleted SOI technology, ULP logic style reduces static power consumption at  $200^\circ\text{C}$  by three orders of magnitude at the expense of increased circuit delay and die area, with good robustness against process variations. Moreover, ULP logic gates feature excellent noise robustness thanks to SNM higher than  $V_{dd}/2$ , which is never achieved in standard CMOS logic style. Functionality of ULP logic style was demonstrated by measurement results of ULP-inverter ring oscillators in  $0.13\ \mu\text{m}$  technology.

#### *Sight from the hill*

Previous considerations and results are nicely summed up when considering the curve of practical energy vs. throughput, as sketched in Fig. O.2. In order to minimize practical energy at the application target throughput, we can act on  $E_{op}$  in two ways: shifting the whole curve down to achieve the lowest possible  $E_{min}$  and translating it leftward or rightward to make minimum-energy point match the target throughput. In this dissertation, we showed that  $E_{min}$  downward shifting can be achieved at technology/device level by optimum MOSFET selection or new technology/device architectures such as FD SOI. Circuit designers can then manage leftward/rightward translation of  $E_{op}$  curve at design time by technology flavor selection, provided that the technology menu is sufficiently versatile, i.e. the devices come in different  $I_0$  versions. Reverse body biasing can finally be used in bulk technology at test and/or run time for post-Silicon fine adaptation by leftward translation of  $E_{op}$  curve. Circuit designers should then take margins to ensure that no rightward translation will be needed, i.e. that minimum-energy point is at a higher throughput than the application target.





**Fig. O.2.** Minimization of energy per operation

Additionally, we use these results to derive, in Appendix A, the technology and circuit specifications for nanometer subthreshold circuits. Given these specifications, we present the author's recommendations for present and future ULP circuits into a technology/circuit roadmap between  $0.13\ \mu\text{m}$  and  $22\ \text{nm}$  nodes.

#### *On the skyline*

The results from this dissertation hopefully not only give clues to answer the motivating questions but also reveal new horizons for further researches, that we would like to report here.

- As shown in Fig. O.1, technology scaling shifts minimum-energy point to the boundary between target throughputs of ULP and LP applications. This points to a possible extension of the market of minimum-energy subthreshold circuits to consumer products. The mass production market of consumer electronics could definitively release the cost-induced technology locker for the niche market of ULP applications. Indeed in this dissertation, we only considered device modifications available to circuit designers by making the assumption that process modifications are prohibitively expensive for the niche market of ULP applications. The new optimum-device paradigm we revealed for nanometer subthreshold logic should thus be used as route for conducting studies of process features such as [1] to manufacture subthreshold-optimized devices, the ultimate goal being to motivate large IC foundries to develop a process dedicated to subthreshold operation.

- We showed that technology versatility is an important tool for minimizing energy consumption. This versatility is nowadays widespread in industrial nanometer bulk technologies. An important technological challenge will be to implement this versatility in new technologies. As nanometer fully-depleted SOI and FinFET technologies feature an undoped channel, technology designers cannot rely on doping level to adapt the threshold voltage. Gate-work function engineering can be used but it seems unlikely that foundries will optimize many different gate stacks to meet the versatility requirements. In this light, ultra-thin-buried-oxide [2] and double-gate fully-depleted SOI devices [3] are very promising to provide  $V_t$ -tuning capability. Moreover, as we showed that independent dual- $V_t$  assignment is not practical in nanometer subthreshold circuits, a common tuning of the threshold voltage of all NMOS and all PMOS devices can be used, thereby considerably simplifying the problem. The same issues are present in future non-Si technologies and early-stage studies on the versatility of these technologies such as [4] are thus highly desirable.
- The inefficiency of independent dual- $V_t$  assignment we reported suggests that the design of nanometer subthreshold circuits highly benefits from circuit regularity. Besides that, layout regularity has been advocated in the design-for-manufacturability flow to push scaling to next technology nodes [5]. It is thus reasonable to suggest that nanometer subthreshold circuits will further benefit from layout regularity, even more than nominal- $V_{dd}$  circuits, so that investigation of subthreshold-optimized layout should also be carried out.
- For VLSI circuits, design techniques are useless unless integrated in design automation flow. As subthreshold logic seems promising for consumer electronics market, a first step to enable widespread adoption of subthreshold design would be for standard-cell library vendors to characterize their libraries at supply voltages below the traditional range (typical characterization range goes from 0.8 to 1.2V in 45 nm technology). Up to know, the considered circuit  $V_{dd}$  is a decision from the circuit designer: even in low-power design flow with multiple supply-voltage domains or dynamic voltage scaling, the supply voltage is set in the designer's power intent. The second step would thus be to adapt logic synthesis tools to unlock  $V_{dd}$  parameter. Rather than picking up logic gates in a multi- $V_t$  fixed- $V_{dd}$  library, a new mode could be added for the tool to automatically pick up one single- $V_t$ /single- $V_{dd}$  library in a wide range of multiple  $V_t/V_{dd}$  libraries.

To conclude this dissertation, let us mention that this work falls in the field of technology/circuit interaction and technology-aware design, which has become a hot topic in the IC community, illustrated by the introduction in 2008 of special technology/circuit-dedicated sessions at both the *IEEE International Electron Devices Meeting (IEDM)* and the joint *European Solid-State Device Research Conference/European Solid-State Circuits Conference (ESSDERC/ESSCIRC)*.

## REFERENCES

1. B. C. Paul, A. Raychowdhury and K. Roy, "Device optimization for digital sub-threshold logic operation", in *IEEE Trans. Electron Dev.*, vol. 52, no. 2, pp. 237-247, Feb. 2005.
2. R. Tsuchiya *et al.*, "Controllable inverter delay and suppressing  $V_{th}$  fluctuation technology in Silicon on thin BOX featuring dual back-gate bias architecture", in *Dig. IEEE Int. Electron Dev. Meeting*, pp. 475-478, 2007.
3. M. Masahara *et al.*, "Demonstration, analysis and device design considerations for independent DG MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 52, no. 9, pp. 2046-2053, Sep. 2005.
4. A. Raychowdhury and K. Roy, "Carbon nanotube electronics: design of high-performance and low-power digital circuits", in *IEEE Trans. Circuits Syst.-I: Reg. Papers*, vol. 54, no. 11, pp. 2391-2401, Nov. 2007.
5. B. H. Calhoun, Y. Cao, X. Li, K. Mai, L. T. Pileggi, R. A. Rutenbar and K. L. Shepard, "Digital circuit design challenges and opportunities in the era of nanoscale CMOS", in *Proc. IEEE*, vol. 96, no. 2, pp. 343-365, Feb. 2008.



## POSTFACE

---

In the very last lines of the conclusion, we mention that this work falls in the “technology-aware design [field], which has become a hot topic in the IC community”. In fact, I personally experienced the increasing importance of technology awareness during my Ph.D cursus. My original research topic was high-performance arithmetic circuits. As a young Ph.D student, I was seduced by the intrinsic power of two exotic logics, beyond classical Boolean logic: signed-digit (ternary) and threshold logics. Working in this field lead us to propose new logic gates to implement corresponding logic operations [UP1-UP4]. Nevertheless, after two years of research, this direction turned out to be a dead end because of technological issues. Indeed, the logic gates we proposed rely on the availability of depletion-mode MOSFETs (with negative threshold voltage), which can hardly be manufactured in deep-submicron bulk and partially-depleted SOI technologies as the requirement of low channel doping implies prohibitive short-channel effects. Moreover, when reaching the nanometer era, these logic gates would face major issues due to their high sensitivity against intrinsic process variability. In the next decade, undoped-channel fully-depleted SOI technology may solve these problems and renew the interest for exotic logic styles. However, back in the end of 2006, availability of such a technology at nanometer nodes was only a long-term hope.

For that reason, we decided to adopt another approach, more fundamental: analyzing the issues of mainstream technology scaling and proposing solutions at abstraction levels we have access to as circuit designers. The application field we selected was ultra-low-power circuits for biomedical devices, given my taste for non-mainstream stuffs and my (utopian) will to bring something more helpful than adding new 3-D graphics capability to smart phones. In January 2007, we came up with the ULP logic style for cutting off leakages (Chapter 5). Although, it is proposed here for high-temperature industrial applications, the original motivation came from low-leakage requirements of biomedical applications [CP1]. In August 2007, we decided to move to subthreshold logic as it seemed very promising for ultra-low-power applications. This time again, the research direction was dictated by the technology. Indeed, both the technology scaling and the increasing versatility of nanometer technologies raised the question of technology selection for a given application, after having defined the important figures of merit. This is how Chapters 2 to 4 started.

This Ph.D experience bears witness to the importance of technology awareness for efficient circuit design. It is now my belief that technology selection is an inherent part of the circuit design process and probably the most important step as it strongly impacts the performances and, consequently, the design choices.

D.B.



# APPENDIX A

## ROADMAP FOR NANOMETER ULTRA-LOW-POWER CIRCUITS

---

In Chapter 2, we pointed out several issues for nanometer subthreshold circuits that we tried to fix in Chapters 3 to 5. In this appendix, we would like to use these results to derive technology and circuit specifications for optimum subthreshold circuits operating at minimum-energy point under robustness and throughput constraints. We then propose a possible roadmap for meeting these specifications in ultra-low-power applications from  $0.13\mu\text{m}$  to 22 nm node.

### A.1 TECHNOLOGY/CIRCUIT SPECIFICATIONS FOR OPTIMUM SUBTHRESHOLD CIRCUITS

In this dissertation, we showed that the increase of practical energy of ULP subthreshold circuits under robustness and throughput constraints can be dealt with in two steps: reducing the minimum-energy level  $E_{min}$  as much as possible and then reaching  $E_{min}$  in practice when taking robustness and throughput constraints into account. As illustrated in Fig. A.1, we proposed to make technology optimizations for reducing  $E_{min}$  and to use circuit techniques for reaching  $E_{min}$ . The resulting technological targets as well as the circuit techniques requirement are detailed hereafter.

#### *Reducing $E_{min}$*

In Chapter 3, we showed that there are multiple effects in nanometer technologies that contributes to  $E_{min}$ . Limiting these effects leads to five key targets for optimum subthreshold MOSFET design at technology/device level in order to reduce  $E_{min}$ :

- low subthreshold swing  $S$ ,
- low DIBL effect,
- low variability of subthreshold reference current  $I_0$  (and thus low  $V_t$  variability),
- low mean load capacitance  $C_L$  including the device parasitic capacitances,
- gate and junction leakages,  $I_{gate}$  and  $I_{junc}$ , below the level of subthreshold leakage  $I_{sub}$ .

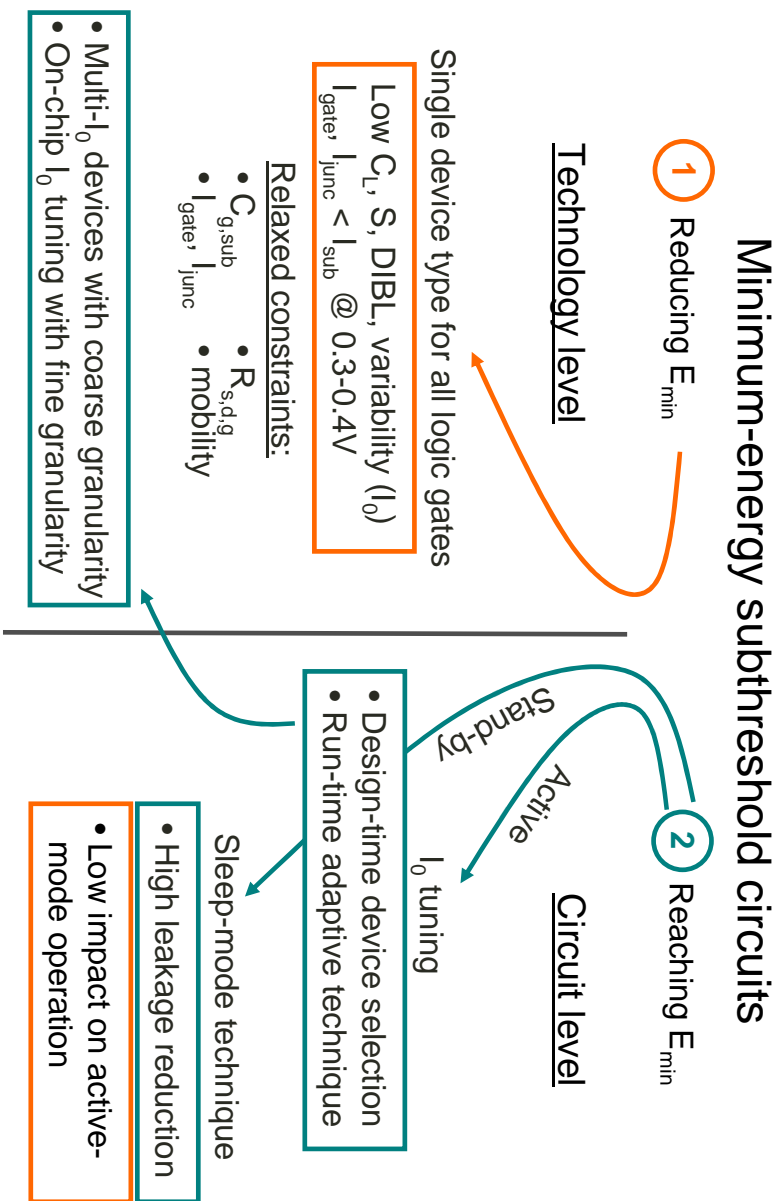


Fig. A.1.1. Specifications for optimum subthreshold circuits



These technological targets are quite general. Indeed, they are also valid when designing devices for nominal- $V_{dd}$  circuits, beyond the scope of ULP circuits. However, in ULP subthreshold circuits, their importance is magnified. The good point is that several technological constraints of nominal- $V_{dd}$  circuits are relaxed when considering subthreshold circuits. First, the intrinsic gate capacitance in subthreshold regime  $C_{g,sub}$  is less important due to the addition of the channel-depletion capacitance  $C_{dep}$  in series with the oxide capacitance  $C_{ox}$ . Therefore,  $C_{g,sub}$  contributes less to  $C_L$  than in nominal- $V_{dd}$  circuits. It can thus be increased to achieve the targets of  $S$ , DIBL and variability minimization. Second, as the on-state subthreshold current is quite low, the equivalent channel resistance is large even in on-state. Therefore, the parasitic resistances associated to the device accesses  $R_s$ ,  $R_d$  and  $R_g$  are proportionally less important. They can thus be increased without speed penalty in order to meet the other technological targets. Third, the subthreshold reference current  $I_0$  depends exponentially on  $V_t$  and linearly on the carrier mobility. A mobility degradation can be tolerated as it is easily compensated by slight  $V_t$  reduction. Finally,  $I_{gate}$  and  $I_{junc}$  leakage currents do not have to be minimized. Technology designers should only prevent them from becoming higher than the subthreshold leakage. These relaxed constraints give space for device optimization to meet the five key targets.

Notice that a single device type common to all logic gates can be used as dual- $V_t/T_{ox}/L_g$  assignments are not practical in subthreshold logic circuits. Only SRAM circuits may require different devices for leakage concern. This may possibly reduce the number of masks and process steps, and thereby save associated manufacturing costs.

### Reaching $E_{min}$

Making the minimum-energy point meet the application target throughput, during active periods, requires  $I_0$  tuning capability. First, at design time, circuit designers should choose the technology with an  $I_0$  value that brings minimum-energy point close to the target throughput. This means that the technology should be versatile and come with multi- $I_0$  devices with a coarse granularity (e.g. three or four  $I_0$  values in a wide range from 10 pA/ $\mu\text{m}$  to 10 nA/ $\mu\text{m}$ ). Moreover, a low-cost circuit technique is required for post-Silicon  $I_0$  tuning: at test time for compensating modeling errors or extrinsic global process variations, and/or at run time for compensating device aging, variations of the environment temperature or a dynamically-varying workload. This implies that such a technique should be enabled at technology level for fine-grain tuning (a smaller  $I_0$  range, e.g. between  $0.2\times$  and  $5\times$  the nominal value).

When the application features stand-by periods, a sleep-mode technique should be used at circuit level with strong leakage-reduction capability when in sleep mode. This technique should feature a low impact on delay and robustness when in active mode to avoid ruining the  $E_{min}$  level.

## A.2 A POSSIBLE TECHNOLOGY/CIRCUIT ROADMAP FOR NANOMETER ULTRA-LOW-POWER CIRCUITS

We now would like to give the author's personal view on the valuable technologies and circuit techniques to achieve the specifications we derived in previous section. As the application spectrum of ULP circuits is quite wide, we divided it in three categories depending on their requirements for the circuits: high-temperature ULP applications in industrial environment (oil drilling, process monitoring), standard ULP applications (RFID tags, biomedical devices and sensor networks) and ULP modes (low-performance mode for background computation or mid-performance with massive parallelization) in low-power/wireless consumer applications (smart phones, blackberries). As illustrated in Fig. A.2, we also divided the roadmap in three groups related to technology nodes: deep-submicron 130/90 nm, present nanometer 65/45 nm and future nanometer 32/22 nm nodes.




### *High-temperature ULP applications*

In Chapter 5, we showed that the proposed ULP logic style is the only technique capable of keeping subthreshold leakage under control in deep-submicron technologies (130/90 nm nodes) when the operating temperature is above 150°C, as no technology option provides  $V_t$  values high enough ( $\sim 0.8$  V). This requires the use of an SOI technology either partially- or fully-depleted (PD or FD, respectively) to achieve compact layout and prevent high junction leakage. A general-purpose (GP) technology flavor should be used to provide sufficient drive current to ULP logic gates. In this case, the maximum reachable throughput is in the order of 1 MOp/s.

At high temperature, reliability issues are magnified. It comes from effects such as negative-bias temperature instability and electromigration, which are higher in nanometer technologies [1]. Therefore, we believe that these reliability issues will prevent from implementing circuits for high-temperature operation in nanometer (present and future) technologies.

### *Standard ULP applications*

In Chapter 2, we showed that subthreshold logic at deep-submicron 130/90 nm nodes is well adapted to standard ULP applications even in bulk technology. At these nodes, foundries usually provide only a GP technology flavor. Fortunately, this is the most appropriate for providing  $I_0$  values that make minimum-energy point meet target throughputs in the range of ULP applications (10 k to 10 MOp/s). Fully-depleted SOI technology or adaptive reverse body biasing (RBB) technique are interesting options for further energy saving by improving MOSFET subthreshold characteristics (technology targets for reducing  $E_{min}$ ) and removing design margins (circuit techniques requirements for reaching  $E_{min}$ ), respectively. However, efficient subthreshold circuits have already been demonstrated at these nodes without these options [2], which shows that they are not compulsory.

Node Applications	130 / 90 nm	65 / 45 nm	32 / 22 nm
High-temperature ULP industrial applications 	ULP logic style <ul style="list-style-type: none"> <li>• PD SOI</li> <li>• (FD SOI)</li> </ul> @ GP flavor	Reliability issues	Reliability issues
Standard ULP applications 	Subthreshold logic <ul style="list-style-type: none"> <li>• Bulk (+ adapt. RBB)</li> <li>• (FD SOI)</li> </ul> @ GP flavor	Subthreshold logic <ul style="list-style-type: none"> <li>• Bulk + adapt. RBB</li> <li>• FD SOI</li> </ul> @ LP flavor	Economical issues
ULP mode in LP applications 	Performance issues	Subthreshold logic <ul style="list-style-type: none"> <li>• Bulk opt. + adapt. RBB</li> <li>• FD SOI</li> </ul> @ HP/GP flavor	Subthreshold logic <ul style="list-style-type: none"> <li>• FD SOI + UTBOX/DG</li> <li>+ adapt. dual-BG bias</li> </ul> @ dedicated flavor

Architectural techniques (//, pipe)  
for meeting throughput constraint

Fig. A.2. Roadmap for nanometer ultra-low-power circuits

When it comes to nanometer technologies at present 65/45 nm nodes, we showed in Chapter 4 that only a low-power (LP) technology flavor features  $I_0$  values compatible with throughput range of ULP applications. Without such an LP flavor, we showed in Chapter 2 that the benefit of die area reduction is waived by an energy increase from 130/90 to 65/45 nm nodes and there is thus no interest in migrating to nanometer technologies. Moreover, in order to avoid prohibitive design margins, which may prevent from reaching  $E_{min}$  in practice, an adaptive RBB technique is needed for test-and/or run-time circuit adaptation. Alternatively, an undoped-channel ultra-thin-body FD SOI technology may be used to remove the need for circuit adaptation, thanks to its lower sensitivity against extrinsic global process variations [3].

If the application features stand-by periods, sleep-mode power-gating technique should be considered for saving leakage energy. At all technology nodes, special care has to be taken when engineering the power switch for subthreshold operation, as shown in Chapter 4.

Regarding future scaling, it is the author's belief that economical reasons will prevent from porting subthreshold circuits for ULP applications beyond 45 nm node. Although further scaling will reduce the costs of raw material from die area reduction, we think that the increasing costs associated to the manufacturing process will no longer be supported by the niche market of ULP applications (low-volume production or low chip selling price), when reaching 32/22 nm nodes.

#### *ULP mode in low-power/wireless applications*

In Chapter 2, we showed that the minimum-energy point in 65/45 nm high-performance/general-purpose (HP/GP) technologies is shifted towards throughput values close to the range of low-power/wireless consumer applications. This may create new opportunities for minimum-energy subthreshold circuits. In HP/GP flavor, the  $E_{min}$  level is higher at 65/45 than at 130/90 nm node. As shown in Chapter 3, an optimum MOSFET selection (low  $V_t$  and upsized gate length) allows to improve the key technological targets and fixes this problem. It has thus to be used in nanometer subthreshold circuits for low-power/wireless applications. Adaptive RBB has to be used too for circuit adaptation even if the bias voltages may be larger in HP/GP than in LP flavor because of reduced body-bias coefficient. Nevertheless, in low-power/wireless applications, the area/volume constraints are somewhat relaxed as compared to ULP applications. Generator or external supplies of largest bias voltages can thus be tolerated. Alternatively, an undoped-channel ultra-thin-body FD SOI technology may be used for further  $E_{min}$  improvement thanks to better subthreshold MOSFET characteristics. It also limits the need for circuit adaptation, which may lead to cost savings. Notice that depending on the  $I_0$  value, architectural techniques such as parallelization or pipelining may be required for meeting the throughput constraints.

To the author's point of view, subthreshold circuits at future 32/22 nm nodes will not be feasible - or at least inefficient - in bulk technology. Indeed, bulk MOSFET are likely to feature bad subthreshold characteristics at these nodes so that

bulk technology will no longer be able to meet the key technological targets for low  $E_{min}$ . Therefore, undoped-channel ultra-thin-body FD SOI technology will be compulsory for nanometer subthreshold circuits at 32/22 nm. Moreover, as the market of consumer low-power/wireless applications is a mass production market, we think that it may motivate IC foundries to develop a process dedicated to subthreshold operation. This process should thus target optimum MOSFET characteristics to implement the specifications we reported in this appendix.

At these nodes, it is likely that circuit adaptation will be needed to avoid prohibitive design margins, even in FD SOI technology. In order to provide adaptation opportunity, the technology should thus come with an ultra-thin-buried-oxide (UTBOX) and dual back-gate (BG) bias [5], or with a double gate (DG) in independent-gate configuration [5].

Additionally, from the author's point of view, multiple-gate devices such as FinFETs, MuGFETs or double-gate MOSFETs in common-gate configuration are less valuable for subthreshold logic for two reasons. First, back-gate biasing has a weak impact on these devices [6] and circuit adaptation through  $I_0$  tuning can thus hardly be achieved. Second, the parasitic capacitances associated to the multiple gates [7] may increase the mean load capacitance  $C_L$  proportionally more than at nominal  $V_{dd}$  given the low intrinsic gate capacitance in subthreshold regime  $C_{g,sub}$ .

## REFERENCES

1. J. Srinivasan, S. V. Adve, P. Bose and J. A. Rivers, "The impact of technology scaling on lifetime reliability", in *Proc. Int. Conf. Dependable Systems and Networks*, pp. 177-188, 2004.
2. B. Zhai, L. Nazhandali, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, D. Blaauw, T. Austin: "A 2.60pJ/inst subthreshold sensor processor for optimal energy efficiency", in *Dig. Tech. Papers VLSI Circuits*, pp. 154-155, 2006.
3. O. Weber *et al.*, "High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding", in *Dig. IEEE Int. Electron Dev. Meeting*, in press, 4 p., 2008.
4. R. Tsuchiya *et al.*, "Controllable inverter delay and suppressing  $V_{th}$  fluctuation technology in Silicon on thin BOX featuring dual back-gate bias architecture", in *Dig. IEEE Int. Electron Dev. Meeting*, pp. 475-478, 2007.
5. M. Masahara *et al.*, "Demonstration, analysis and device design considerations for independent DG MOSFETs", in *IEEE Trans. Electron Dev.*, vol. 52, no. 9, pp. 2046-2053, Sep. 2005.
6. F. Dauge, J. Pretet, S. Cristoloveanu, A. Vandooren, L. Mathew, J. Jomaah, B.-Y. Nguyen, "Coupling effects and channels separation in FinFETs", in *Solid-State Electronics*, vol. 48, no. 4, pp. 535-542, April 2004.
7. J.-P. Raskin, T. M. Chung, V. Kilchytska, D. Lederer and D. Flandre, "Analog/RF performance of multiple gate SOI devices: wideband simulations and characterization", in *IEEE Trans. Electron Dev.*, vol. 53, no. 5, pp. 1088-1095, May 2006.



## APPENDIX B

### DESCRIPTION OF THE CIRCUIT

### SIMULATION BENCHMARK

---

In this dissertation, we based most of our simulation results on a common benchmark circuit. In this appendix, we give a brief description of this circuit and of the simulation setup.

#### B.1 8-BIT RCA BENCHMARK MULTIPLIER

The considered benchmark circuit is a standard 8-bit ripple-carry-array (RCA) multiplier. The 8-bit word width was selected to get representative results while keeping affordable simulation time. The multiplier architecture is represented in Fig. B.1. In this figure, AND cells are made of a 2-input NAND gate with an inverter in series. FA (resp. HA) cells are made of an AND cell to generate a partial product, which is then fed into a full (resp. half) adder. Carry out/in paths are horizontally routed and Sum paths are vertically routed.

As shown in Table B.1, there are 180 logic gates in the multiplier (inverters, 2-input NAND, full and half adders). The full adder implementation is the standard static CMOS 28-transistor one [1] and the half adder has the corresponding 14-transistor implementation. The total transistor count is 1828. The logic depth is 23 (including two stages for Sum output in full/half adders).

All devices feature the minimum width of the considered technology, except:

- NAND gates, where the width of stacked NMOS devices is doubled to mitigate the unbalance between pull-up PMOS and pull-down NMOS networks,
- Carry in/out path in full/half adders, where the width of NMOS and PMOS devices is increased by 50% for limiting the critical-path delay.

#### B.2 SIMULATION SETUP

For each input, the simulation setup includes a two-stage-inverter buffer operating under the same conditions as the benchmark circuit for feeding the circuit with realistic input signals (levels and rise/fall times). Moreover, each output drives a two-stage NAND load, as realistic loads. Routing capacitances are not

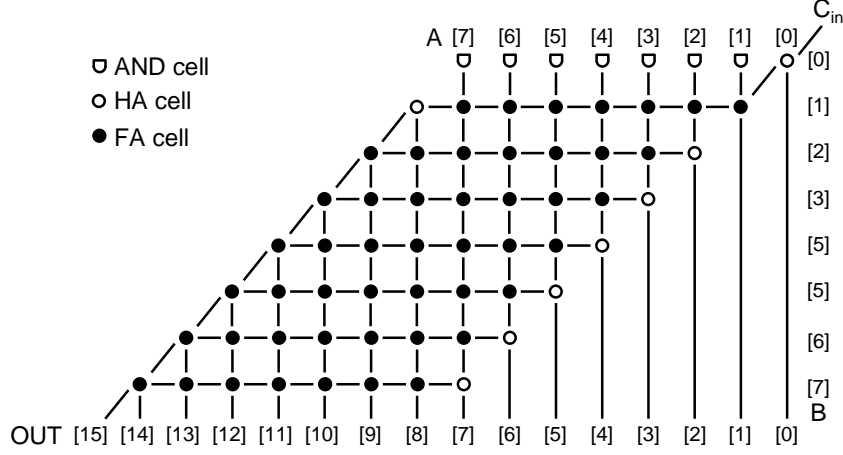


Fig. B.1. Architecture of the benchmark multiplier

Table B.1. Complexity of the benchmark multiplier

Gate count	Transistor count	Logic depth
180	1828	23

included in the simulations. Simulations are carried out with Eldo tool, the Spice-like simulator from Mentor Graphics. The backward-Euler convergence method is used for transient simulations as it yields good stability in low currents.

The critical path delay  $T_{del}$  is statistically-extracted from Monte-Carlo simulations with 100 runs. Notice that for simulation time concern, a single critical path is simulated (from A[0] input to OUT[15] output). This may slightly underestimate the mean delay from statistical variations as other paths may become critical with intrinsic variability, which is uncorrelated between logic gates. However, this gives representative results, which allows comparing technology and circuit techniques as the same simulation methodology is kept in all cases. The worst case of rising and falling transitions (measured between 50%-50% input/output transitions) is considered as critical path delay for each Monte-Carlo run. The typical simulation time for the 100 runs on a recent UNIX workstation (1.6 GHz Ultra-Sparc IIIi with 2 Gb memory) is 1.5 hour.

From the extracted mean and standard deviation of the delay, the  $3\sigma$  worst-case  $T_{del}$  (99.9% confidence interval) is computed with different expressions depending on the considered  $V_{dd}$ . For nominal  $V_{dd}$ ,  $T_{del}$  is normally-distributed and the  $3\sigma$  worst case is simply calculated as  $T_{del,3\sigma} = Mean(T_{del}) + 3 StdDev(T_{del})$ . For subthreshold  $V_{dd}$ ,  $T_{del}$  follows a lognormal distribution [2] and the  $3\sigma$  worst-



case delay is calculated as:

$$T_{del,3\sigma} = e^{\mu+3\sigma}$$

$$\text{where } \sigma = \sqrt{\ln \left( \frac{StdDev(T_{del})^2}{Mean(T_{del})^2} + 1 \right)}$$

$$\text{and } \mu = \ln(Mean(T_{del})) - \frac{1}{2} \ln \left( 1 + \frac{StdDev(T_{del})^2}{Mean(T_{del})^2} \right).$$

Static power/energy is calculated from statistical extraction of the mean total leakage in Monte-Carlo DC simulations with 100 runs. The mean leakage current is larger than the typical one as subthreshold leakage is a lognormal distribution.

Typical dynamic power/energy is calculated by subtracting the static power/energy from the total power. In this case, both static and total power are extracted by simulating the benchmark circuit with pseudo-random input pattern. Total power is measured by the integration of instantaneous power over 10 data periods while static power is the mean of 10 measurements of instantaneous power consumption at the end of each data period (when internal nodes are stable). The period duration is chosen as  $10 \times$  the typical delay as a trade-off between high dynamic/static power ratio and stable conditions when measuring the static power at the end of each data period.

## REFERENCES

1. A. Tisserand, "Low-power arithmetic operators", in *Low-Power Electronics Design*, Ch. Piguat Ed., CRC Press, pp. 9.1-15, 2005.
2. B. Zhai, S. Hanson, D. Blauw and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Proc. IEEE/ACM Int. Symp. on Low-Power Electron. Des.*, pp. 20-25, 2005.



## APPENDIX C

# BSIM4 PRE-SILICON NANOMETER MOSFET MODEL CARDS

---

In this dissertation, we set up a methodology for generating pre-Silicon BSIM4 model cards for subthreshold circuit simulation, starting from the Predictive Technology Model from Arizona State University<sup>1</sup>. The methodologies for generating bulk and fully-depleted SOI MOSFET model cards are described in Sections 3.3 and 3.7, respectively. In this appendix, we present the generated model cards at 45 nm node as an example. Notice that the model cards are intended for Eldo simulator from Mentor Graphics. In order to use it in other Spice-like simulators, the `level` parameter has to be updated. Some syntax modifications may be required too.

The model cards are enclosed in sub-circuit instances and have to be used that way (N\_STD model card):

```
xM1 VD VG VS VB N_STD W=wn L=lg
.param lg_nom=45n
.param DEV_LG = 0.0 DEV/gauss = 1.4n
.param lg=lg_nom+(DEV_LG*MISMATCH_L)
.param wn=1.5*45n
.param MISMATCH_L=1
.param MISMATCH_VT=1
```

The length to assign to the devices is the 45 nm drawn  $L_g$ , which corresponds to a 35 nm printed  $L_g$ . The printed  $L_g$  has to be modeled as a normally-distributed variable to include extrinsic process variability. Therefore, we add a normal distribution to the drawn  $L_g$ , which implies a corresponding distribution of the printed  $L_g$ . This is done in the netlist in order to set the same printed  $L_g$  to all devices, as we assume full correlation of  $L_g$  amongst devices. The switch to activate extrinsic global  $L_g$  variability is `MISMATCH_L` and the switch to activate intrinsic local  $V_t$  variability due to random dopinf fluctuations is `MISMATCH_VT`.

As we target subthreshold simulation, we did not calibrate the parameters associated to irrelevant effects such as the access resistances.

<sup>1</sup>Models are available on-line at <http://www.eas.asu.edu/~ptm>.

*Bulk MOSFET model cards*

This is the model card for the std- $V_t$  (0.35V) NMOS device for high-performance flavor in bulk technology:

```
.subckt N_STD D G S B
M1 D G S B N_STD w=w l=l AS='122.5e-9*w' AD='122.5e-9*w'
PD='w+2*122.5e-9' PS='w+2*122.5e-9' NF=1

* Vt variability parameters
.param sig_vt_RDF=1.50e-9/((w*(1-27.5n))^0.5)
.param sig_vt_Other=0.5e-9/((w*(1-27.5n))^0.5)
.param DEV_VTH0_RDF = 0.0 DEV/gauss = sig_vt_RDF
.param DEV_VTH0_Other = 0.0 DEV/gauss = sig_vt_Other
.param DEV_VTH0 = (DEV_VTH0_RDF+DEV_VTH0_Other)*MISMATCH_VT

.model N_STD nmos level = 60

+version = 4.4 binunit = 1 paramchk= 1 mobmod = 0
+capmod = 2 igcmod = 1 igbmod = 0 geomod = 1
+diomod = 1 rdsmod = 0 rbodysmod= 1 rgatemod= 1
+permod = 1 acnqsmode= 0 trnqsmode= 0

* basic parameters
+tnom = 27 epsrox = 3.9
+wint = 0e-09
+x1 = -2e-08

* parameters customized by the user from PTM original model
+toxe = 1.75e-09 toxp = 1.1e-09 toxm = 1.75e-09 toxref = 1.8e-09
+dtox = 6.5e-10 lint = 3.75e-09
+vt0 = '0.513+DEV_VTH0' k1 = 0.566 u0 = 0.04077 vsat = 147390
+rdsw = 155 ndep = 3.77e+18 xj = 1.4e-08

* parasitic capacitance parameters
+cjsws = '1.225e-08*0.0024' cjswd = '1.225e-8*0.0024'
+cjs = 0.0024 cjd = 0.0024
+cgso = '(1/3)*0.246e-9' cgdo = '(1/3)*0.246e-9'
+cgbo = 0 cgdl = 0
+cgs1 = 0
+cf = '(1/3)*0.874e-9'

* parameters for S vs Lg fitting (no SCE nor RSCE)
+nfactor = 1 cdsc = 0.00017 dvt0 = 0 dvt1 = 0.1
```

```

* parameters for DIBL vs Lg fitting
+eta0 = 0.0056 dsub = 0.1

* parameters for gate leakage
+aigc = '0.9*0.011' bigc = '0.95*0.0045'
+cigc = '1.2*0.66' nigc = '1.68'
+aigsd = 0.011 bigsd = 0.0027 cigsd = 0.24 pigcd = 9
+DLCIG=3.5e-9 poxedg = 1 ntox = 22.5

* don't care because igbmod=0
+aigbacc = 0 bigbacc = 0 cigbacc = 0 nigbacc = 0
+aigbinv = 0 bigbinv = 0 cigbinv = 0 nigbinv = 0
+eigbinv = 0

* parameters for junction leakage: not taken into account
+tjss = 1e-15 jsws = 1e-15 jswgs = 0 njs = 1
+tjssd = 1e-15 jsswd = 1e-15 jssgd = 0 njd = 1

.ends N_STD

```

This is the model card for the corresponding PMOS device:

```

.subckt P_STD D G S B
M1 D G S B P_STD w=w l=l AS='122.5e-9*w' AD='122.5e-9*w'
PD='w+2*122.5e-9' PS='w+2*122.5e-9' NF=1

* Vt variability parameters
.param sig_vt_RDF=1.41e-9/((w*(1-27.5n))^0.5)
.param sig_vt_Other=0.5e-9/((w*(1-27.5n))^0.5)
.param DEV_VTH0_RDF = 0.0 DEV/gauss = sig_vt_RDF
.param DEV_VTH0_Other = 0.0 DEV/gauss = sig_vt_Other
.param DEV_VTH0 = (DEV_VTH0_RDF+DEV_VTH0_Other)*MISMATCH_VT

.model P_STD pmos level = 60

+version = 4.4 binunit = 1 paramchk= 1 mobmod = 0
+capmod = 2 igcmod = 1 igbmod = 0 geomod = 1
+diomod = 1 rdsmod = 0 rbodmod= 1 rgatemod= 1
+permod = 1 acnqsmo= 0 trnqsmo= 0

* basic parameters
+tnom = 27 epsrox = 3.9
+wint = 0e-09
+xl = -2e-08

```

```

* parameters customized by the user from PTM original model
+toxe = 1.85e-09 toxp = 1.1e-09 toxm = 1.85e-09 toxref = 1.9e-09
+dtox = 7.5e-10 lint = 3.75e-09
+vth0 = '-0.45+DEV_VTH0' k1 = 0.515 u0 = 0.00392 vsat = 70000
+rdsw = 155 ndep = 2.79e+18 xj = 1.4e-08

* parasitic capacitance parameters
+cjsws = '1.225e-08*0.0024' cjswd = '1.225e-8*0.0024'
+cjs = 0.0024 cjd = 0.0024
+cgso = '(1/3)*0.246e-9' cgdo = '(1/3)*0.246e-9'
+cgbo = 0 cgdl = 0
+cgs1 = 0
+cf = '(1/3)*0.874e-9'

* parameters for S vs Lg fitting (no SCE nor RSCE)
+nfactor = 1 cdsc = 0.00017 dvt0 = 0 dvt1 = 0.1

* parameters for DIBL vs Lg fitting
+eta0 = 0.0056 dsub = 0.1

* parameters for gate leakage
+aigc = '0.99*0.0089' bigc = '1*0.0015'
+cigc = '0.03' nigc = '1*2.44'
+aigsd = 0.0068 bigsd = 0.00047 cigsd = 0.098 pigcd = 9.1
+DLCIG=6.65e-9 poxedge = 1 ntox = 23

* don't care because igbmod=0
+aigbacc = 0 bigbacc = 0 cigbacc = 0 nigbacc = 0
+aigbinv = 0 bigbinv = 0 cigbinv = 0 nigbinv = 0
+eigbinv = 0

* parameters for junction leakage: not taken into account
+jss = 1e-15 jsws = 1e-15 jswgs = 0 njs = 1
+jjsd = 1e-15 jsd = 1e-15 jsdgd = 0 njd = 1

.ends P_STD

```

### FD SOI MOSFET model cards

This is the model card for the std- $V_t$  (0.26V) NMOS device for high-performance flavor in undoped-channel fully-depleted SOI technology with a midgap metal gate:

```
.subckt N_STD D G S BG

.param Cbox=3.9*8.85e-12/145n
.param Cj_val=Cbox*w*45n*2.5

Cjd D BG Cj_val
Cjs S BG Cj_val

M1 D G S S N_STD w=w l=l AS='122.5e-9*w' AD='122.5e-9*w'
PD='w+2*122.5e-9' PS='w+2*122.5e-9' NF=1

.param sig_vt_RDF=0.067e-9/((w*(1-27.5n))^0.5)
.param sig_vt_Tsi=0e-3
.param sig_vt_Other=0.5e-9/((w*(1-27.5n))^0.5)
.param DEV_VTH0_RDF = 0.0 DEV/gauss = sig_vt_RDF
.param DEV_VTH0_Tsi = 0.0 DEV/gauss = sig_vt_Tsi
.param DEV_VTH0_Other = 0.0 DEV/gauss = sig_vt_Other
.param DEV_VTH0 = (DEV_VTH0_RDF+DEV_VTH0_Tsi+DEV_VTH0_Other)*MISMATCH_VT

.model N_STD nmos level = 60

+version = 4.4 binunit = 1 paramchk= 1 mobmod = 0
+capmod = 2 igcmod = 1 igbmod = 0 geomod = 1
+diomod = 1 rdsmod = 0 rbodysmod= 1 rgatemod= 1
+permod = 1 acnqsmode= 0 trnqsmode= 0

* basic parameters
+tnom = 27 epsrox = 3.9
+wint = 0e-09
+x1 = -2e-08

* parameters customized by the user from PTM original model
+toxe = 1.45e-09 toxp = 1.1e-09 toxm = 1.45e-09 toxref = 1.32e-09
+dtox = 3.5e-10 lint = 3.75e-09
+vth0 = '0.414+DEV_VTH0' k1 = 0 u0 = 0.04077 vsat = 147390
+rdsw = 155 ndep = 3.77e+18 xj = 1.4e-08
```

```

* parasitic capacitance parameters
+cjsws = 0 cjswd = 0
+cjs = 0.0002 cjd = 0.0002
+cgso = '(1/3)*0.3e-9' cgdo = '(1/3)*0.3e-9'
+cgbo = 0 cgdl = 0
+cgs1 = 0
+cf = '(1/3)*0.859e-9'

* parameters for S vs Lg fitting (no SCE nor RSCE)
+nfactor = 0.04 cdsc = 0.00016 dvt0 = 0 dvt1 = 0.1

* parameters for DIBL vs Lg fitting
+eta0 = 0.0063 dsub = 0.1

* parameters for gate leakage
+aigc = '0.9*0.011' bigc = '0.95*0.0045'
+cigc = '1.2*0.66' nigc = '1.68'
+aigsd = 0.011 bigsd = 0.0027 cigsd = 0.24 pigcd = 9
+DLCIG=3.5e-9 poxedge = 1 ntox = 22.5

* don't care because igbmod=0
+aigbacc = 0 bigbacc = 0 cigbacc = 0 nigbacc = 0
+aigbinv = 0 bigbinv = 0 cigbinv = 0 nigbinv = 0
+eigbinv = 0

* parameters for junction leakage: not taken into account
+jss = 1e-15 jsws = 1e-15 jswgs = 0 njs = 1
+jsd = 1e-15 jswd = 1e-15 jswgd = 0 njd = 1

.ends N_STD

```

This is the model card for the corresponding PMOS device:

```

.subckt P_STD D G S BG

.param Cbox=3.9*8.85e-12/145n
.param Cj_val=Cbox*w*45n*2.5

Cjd D BG Cj_val
Cjs S BG Cj_val

M1 D G S S P_STD w=w l=1 AS='122.5e-9*w' AD='122.5e-9*w'
PD='w+2*122.5e-9' PS='w+2*122.5e-9' NF=1

```



```

.param sig_vt_RDF=0.063e-9/((w*(1-27.5n))^0.5)
.param sig_vt_Tsi=0e-3
.param sig_vt_Other=0.5e-9/((w*(1-27.5n))^0.5)
.param DEV_VTH0_RDF = 0.0 DEV/gauss = sig_vt_RDF
.param DEV_VTH0_Tsi = 0.0 DEV/gauss = sig_vt_Tsi
.param DEV_VTH0_Other = 0.0 DEV/gauss = sig_vt_Other
.param DEV_VTH0 = (DEV_VTH0_RDF+DEV_VTH0_Tsi+DEV_VTH0_Other)*MISMATCH_VT

.model P_STD pmos level = 60

+version = 4.4 binunit = 1 paramchk= 1 mobmod = 0
+capmod = 2 igcmod = 1 igbmod = 0 geomod = 1
+diomod = 1 rdsmod = 0 rbodmod= 1 rgatemod= 1
+permod = 1 acnqsmode= 0 trnqsmode= 0

* basic parameters
+tnom = 27 epsrox = 3.9
+wint = 0e-09
+xl = -2e-08

* parameters customized by the user from PTM original model
+toxe = 1.55e-09 toxp = 1.1e-09 toxm = 1.55e-09 toxref = 1.42e-09
+dtex = 4.5e-10 lint = 3.75e-09
+vt0 = '-0.371+DEV_VTH0' k1 = 0 u0 = 0.00392 vsat = 70000
+rdsw = 155 ndep = 2.79e+18 xj = 1.4e-08

* parasitic capacitance parameters
+cjsws = 0' cjswd = 0
+cjs = 0.0002 cjd = 0.0002
+cgso = '(1/3)*0.3e-9' cgdo = '(1/3)*0.3e-9'
+cgbo = 0 cgdl = 0
+cgs1 = 0
+cf = '(1/3)*0.859e-9'

* parameters for S vs Lg fitting (no SCE nor RSCE)
+nfactor = 0.04 cdsc = 0.00016 dvt0 = 0 dvt1 = 0.1

* parameters for DIBL vs Lg fitting
+eta0 = 0.0063 dsub = 0.1

* parameters for gate leakage
+taigc = '0.99*0.0089' bigc = '1*0.0015'
+cigc = '0.03' nigc = '1*2.44'
+taigsd = 0.0068 bigsd = 0.00047 cigsd = 0.098 pigcd = 9.1
+DLCIG=6.65e-9 poxedg = 1 ntox = 23

```

```

* don't care because igbmod=0
+aigbacc = 0 bigbacc = 0 cigbacc = 0 nigbacc = 0
+aigbinv = 0 bigbinv = 0 cigbinv = 0 nigbinv = 0
+eigbinv = 0

* parameters for junction leakage: not taken into account
+jss = 1e-15 jsws = 1e-15 jswgs = 0 njs = 1
+jsd = 1e-15 jswd = 1e-15 jswgd = 0 njd = 1

.ends P_STD

```

### *Secondary parameters*

All model cards (bulk and FD SOI, NMOS and PMOS) share a common set of parameters, which have original values from PTM model cards:

```

* secondary parameters: default values from PTM original model
+ll = 0 wl = 0 llm = 1 wlm = 1
+lw = 0 ww = 0 lwm = 1 wwm = 1
+lwl = 0 ww1 = 0 xpart = 0
+k2 = 0.01 k3 = 0
+k3b = 0 w0 = 2.5e-006
+dvt2 = -0.032 dvt0w = 0 dvt1w = 0 dvt2w = 0
+minv = 0.05 voff1 = 0 dvtp0 = 1.0e-009
+dvtp1 = 0.1 lpe0 = 0 lpeb = 0
+ngate = 2e+020 nsd = 2e+020 phin = 0
+cdsch = 0 cdsd = 0 cit = 0
+voff = -0.13 etab = 0
+vfb = -0.55 ua = 6e-010 ub = 1.2e-018
+uc = 0 a0 = 1.0 ags = 1e-020
+a1 = 0 a2 = 1.0 b0 = 0 b1 = 0
+keta = 0.04 dwg = 0 dwb = 0 pclm = 0.04
+pdiblc1 = 0.001 pdiblc2 = 0.001 pdiblc3 = -0.005 drout = 0.5
+pvag = 1e-020 delta = 0.01 pscbe1 = 8.14e+008 pscbe2 = 1e-007
+fprout = 0.2 pdits = 0.08 pditsd = 0.23 pditsl = 2.3e+006
+rsh = 5 rsw = 85 rdw = 85
+rdswmin = 0 rdwmin = 0 rswmin = 0 prwg = 0
+prwb = 6.8e-011 wr = 1 alpha0 = 0.074 alpha1 = 0.005
+beta0 = 30 agidl = 0.0002 bgidl = 2.1e+009 cgidl = 0.0002
+egidl = 0.8
+xrcrg1 = 12 xrcrg2 = 5
+ckappas = 0.03 ckappad = 0.03 acde = 1
+moin = 15 noff = 0.9 voffcv = 0.02
+kt1 = -0.11 kt1l = 0 kt2 = 0.022 ute = -1.5

```

```

+ua1 = 4.31e-009 ub1 = 7.61e-018 uc1 = -5.6e-011 prt = 0
+at = 33000
+fnoimod = 1 tnoimod = 0
+ijthsfwd= 0.01 ijthsrev= 0.001 bvs = 10 xjbvs = 1
+ijthdfwd= 0.01 ijthdrev= 0.001 bvd = 10 xjbvd = 1
+pbs = 1 mjs = 0.5 pbsws = 1
+mjsws = 0.33 pbswgs = 1 cjswgs = 0
+mjswgs = 0.33 pbd = 1 mjd = 0.5
+pbswd = 1 mjswd = 0.33 pbswgd = 1
+cjswgd = 0 mjswgd = 0.33 tpb = 0.005 tcj = 0.001
+tpbsw = 0.005 tcjsw = 0.001 tpbswg = 0.005 tcjswg = 0.001
+xtis = 3 xtid = 3
+dmcg = 0e-006 dmci = 0e-006 dmdg = 0e-006 dmcgt = 0e-007
+dwj = 0.0e-008 xgw = 0e-007 xgl = 0e-008
+trshg = 0.4 gbmin = 1e-010 rbpb = 5 rbpd = 15
+rbps = 15 rbdb = 15 rbsb = 15 ngcon = 1

```