Improvements towards Optimal Design of Reliable Subthreshold Digital CMOS with applications in Logic and Memory.

Hans Kristian Otnes Berge

July 15, 2015

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1658

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Contents

Nomenclature

Notation

6T	Short hand for 6-transistor (circuit)
E$[X]$	Estimate (arithmetic mean) of random variable X
SD$[X]$	Standard deviation of random variable X
SD$^2[X]$	Variance of random variable X

Variables and Constants

α	Activity factor
A_{VT}	Mismatch factor for determination of SD$[V_{th}]$. Typically expressed in units of mVµm
β	The ratio of width to length for the transistor. $(\frac{W}{L})$
C_D	Depletion (channel-bulk) capacitance
C_L	Load capacitance.
C_{ox}	Gate oxide capacitance
E_{min}	The minimum achievable energy per operation
E_{dyn}	Dynamic energy
E_{leak}	Leakage (static) energy
E_{op}	Energy per operation
E_{sc}	Excess short-circuit energy
ϵ_{ox}	Permittivity of Oxide
ϵ_{Si}	Permittivity of Silicon
I_{ds}	Drain to source current
I_{off}	Drain to source current when $V_{gs} = 0, V_{ds} = V_{DD}$

I_{on}	Drain to source current when $V_{gs} = V_{ds} = V_{DD}$
k	Boltzmann's constant 8.617×10^{-5} eV K^{-1}
L	Transistor length
L_{eff}	Transistor effective length
M	Propagation delay scaling factor
μ_0	Electron mobility at low V_{ds} bias.
n	Subthreshold slope factor
P_α	Switching probability
P_{dyn}	Dynamic power consumption (switching power).
P_{stat}	Static power consumption (leakage power).
q	A quantile, $q \in (0, 1)$
q	Electron charge in Coloumbe or electronVolts. $1.602 \times 10^{-19}\,\mathrm{C} = 1\,\mathrm{eV}$
Q, Q_R	Number of standard deviations
T	Temperature, typically in Kelvin
t_{ox}	Gate oxide thickness
t_p	Intrinsic gate propagation delay
U_T	The thermal voltage (kT/q)
V_{DD}	The supply voltage
V_{ds}	Drain-source voltage
V_{gs}	Gate-source voltage
V_{min}	The minimum supply voltage before failure occurs
V_{opt}	The optimum supply voltage achieving E_{min}
V_{th}	The transistor threshold voltage
$V_{th,n}$	Threshold voltage for NMOS device
$V_{th,p}$	Threshold voltage for PMOS device
W	Transistor width
wc	Index signifying worst-case condition
W_{eff}	Transistor effective width

Acronyms

ABB	Adaptive Body Bias.
ASIC	Application Specific Integrated Circuit.
BL	Bitline.
BSIM	Berkeley Short-channel IGFET Model.
CMOS	Complementary Metal-Oxide-Semiconductor.
CPI	Clocks per Instruction.
DIBL	Drain Induced Barrier Lowering.
DITS	Drain Induced Threshold Shift.
DVS	Dynamic Voltage Scaling.
DVFS	Dynamic Voltage and Frequency Scaling.
EDA	Electronic Design Automation.
EKV	Enz, Krummenacher, Vittoz. A MOSFET model.
FIFO	First in, First Out (memory buffer for queuing/flow control).
INWE	Inverse Narrow Width Effect.
MEP	Minimum Energy (operating) Point.
MOEA	Multi-Objective Evolutionary Algorithm.
MOO	Multi-Objective Optimization.
MOOP	Multi-Objective Optimization Problem.
MOS	Metal-Oxide-Semiconductor (Transistor).
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor.
MST	Minimum-Split Transistor. Array of parallel transistors using minimum width.
NBTI	Negative Bias Temperature Instability.

NMOS	N-type Metal-Oxide-Semiconductor.
NWE	Narrow Width Effect.
PMOS	P-type Metal-Oxide-Semiconductor.
PDP	Power-Delay Product.
PVT	Process, Voltage and Temperature.
RDF	Random Dopant Fluctuations.
SRAM	Static Random Access Memory.
SCE	Short Channel Effect.
SNM	Static Noise Margin.
STI	Shallow Trench Isolation.
RISC	Reduced Instruction Set Computing.
RMSE	Root Mean Square Error.
RSCE	Reverse Short Channel Effect.
RV	Random Variable.
ULV	Ultra-Low Voltage.
ULP	Ultra-Low Power.
VLIW	Very Large Instruction Word.
WL	Wordline.
WT	Wide transistor. Used when comparing with MST.

List of Figures

Abstract

This dissertation is organized as a collection of papers, where each paper represents original research contributions relating to the design and analysis of ultra low power CMOS, with a particular emphasis on ultra low voltage and subthreshold operation. The individual papers represent advancements particularily within methods and practices related to the design of both digital logic and memory circuits in the presence of severe process variation. At the device level it is demonstrated how the use of multiple minimum-width gates can exploit the inverse narrow-width subthreshold device effect to improve performance and power-delay products. Measurement results from a 90 nm prototype confirm the effect. Multi-objective optimization strategies are developed and applied to allow exploration of the Pareto optimal design space for reliable logic at 150 mV. Targeting operation at 300 mV, the design of a 9-transistor SRAM memory cell employing multi-Vt and virtual power techniques is presented. A multi-objective optimization strategy is developed and applied to achieve an optimal trade-off for an efficient and reliable sizing of the SRAM cell. Based on the 9-transistor cell, measured results from an ultra low voltage 64 × 32 SRAM module operating down to 273 mV in a 65 nm technology indicate good yield and competitive performance metrics (17.8 fJ/access/bit at averages of 761 kHz @ 321 mV supply). Finally, the behavior of subthreshold logic circuits under the influence of adverse fluctuations in the transistor threshold voltages is treated analytically, with specific emphasis on minimum-energy operation and yield constraints. The analysis can suggest optimal choices for supply voltage and device sizing, prior to simulation.

Preface

During my studies at the University of Oslo I had the pleasure of making many acquaintances, both local and across the world. First and foremost I would like to thank my main thesis supervisor Professor Snorre Aunet, who is now with the Norwegian University for Science and Technology (NTNU), for his genuineness, his friendship, and for a lot of helpful advice and encouragement throughout the course of my Ph.D. Also thanks to my co-supervisors Professor Tor Sverre Lande, and Professor Emeritus Oddvar Søråsen. Thank you to the Department for Computer Science for funding my Ph.D. from 2009-2012, and to the Nanoelectronics Research Group for funding 90 nm chip production, and thank you to Nano Network for funding my trip to the DDECS conference in 2013.

I would also like to thank Associate Professor Sigbjørn Næss, whom I had the pleasure of working with from 2009 to 2011, during the early development of the courses INF1410 and later INF1411. Here I got the chance to hone my skills on teaching, creating laboratory exercises, tutorials, and solutions to hundreds of undergraduate EE problems. During the semester 2011/12 I also had the pleasure of co-supervising Martin Haugland on his M.Sc. thesis work, culminating in a tiny functional subthreshold standard cell library, including sample synthesized layouts.

From the research groups Cognitronics and Sensor Systems Group at the Center of Excellence Cognitive Interaction Technology, Bielefeld University, and Systems and Circuit Technology, Heinz-Nixdorf Institute, University of Paderborn, I would like to especially express my thanks to Professor Dr.-Ing. Ulrich Rückert, Dr.-Ing. Sven Lütkemeier, and Dr.-Ing. Mathias Blesken, for our successful collaboration on two papers, as well as being excellent hosts, and for ensuring interesting talks during our research visit in Paderborn in 2010, and 2011, as well as funding of the 65 nm chip presented in Paper V. Thanks also go to the German Academic Exchange Service DAAD, and the Norwegian Research Council, for funding our exchange visits at Paderborn University in the project “Robust Ultra-Low-Power Circuits for Nano-Scale CMOS Technologies”.

To my all my former office cohabitants, Amir Hasanbegovic, Dr. Kin-Keung Lee, Dr. Jørgen Andreas Michaelsen, Ali Zaher and his family, Kristian Gjertsen Kjelgård, and Dr. Jan Erik Ramstad, thank you all for your friendship, your openness, and the times that we shared. Especially thanks to Amir Hasanbegovic for the collaboration on Paper III as well as taking the lead role on our tutorial on standard cell characterization. And special thanks also to Dr. Jørgen Andreas Michaelsen for great collaborative work on our project “A Low-Voltage Low-Power and Low-Noise Signal Amplification and Activity Detector”.

Thanks also to all my other colleagues at the Nanoelectronics Research Group of whom I had the pleasure of meeting there, to my former M.Sc. thesis advisor Associate Professor

Philipp Häfliger for his efforts then and continued support now, to Malihe Z. Doogahbadi, Thanh Trung Nguyen, Srinivasa Reddy Kuppi Reddi, Dr. Shanthi Sudalaiyandi, thanks. Thanks also to Head Engineer Olav Stanly Kyrvestad, for many years of friendship and for keeping the labs fully operational, to Dr. Øivind Næss, and to Professor Yngvar Berg, for very interesting discussions, but perhaps mostly for being an excellent teacher which I had the good fortune of experiencing as an undergraduate student, to Håkon Andre Hjortland, for sharing his love and passion for Linux-based free software tools, and to Associate Professor Joar Martin Østby, for his dedication in arranging interesting technical meetings in Analog Asic Forum, and everyone else if I have missed anyone.

Thanks goes also to all my colleagues at Integrated Detector Electronics for accepting me so quickly. Especially to the general manager Dr. Gunnar Mæhlum for allowing me time to complete my thesis, but also to Cand. real. Alf Olsen, Dr. Dirk Meier, Jahanzad Talebi, Mehmet Altan, Jörg Ackermann, Suleyman Azman, Bahram Najafiuchevler, Codin Gheorge, Tor Magnus Johansen, Philip Påhlsson, Petter Øya, David Steenari, Aage Kalsæg, Willy Dang, Arkadiusz Edward Dlugolecki and Lyusine Shakbazhyan.

Finally, I would like to thank my family, in particular my parents, Erling Berge and Berit Otnes, for always being loving and supportive. And especially my beautiful soul-mate and wife, Samantha Kay Kelly Berge, for her continuing, radiant, and unabating love and support, never ceasing to believe in me during my work towards this Ph.D.

Chapter 1

Introduction

1.1 Why low voltage CMOS ?

Historically, in commercial CMOS technologies, the core supply voltage has been scaled down along with transistor dimensions. The main reason for doing so can be tied to device reliability issues, as high electric fields can cause damage and reduce the lifetime of nanometer scale transistors [1]. Additionally, the reduced supply voltage facilitates a power reduction. Beneficially this reduces power dissipation, and very importantly it reduces self-heating which can be a major concern in densely packed high-performance devices.

For any electronic digital circuit technology scaling down the power supply voltage is beneficial in terms of reducing both the dynamic (switching) and passive (leakage) power consumption. For circuits dominated by active (switching) power consumption these gains are roughly proportional to the frequency reduction and the square of the voltage reduction [2]. For circuits dominated by static power (standby leakage power), gains in deep submicron processes can have an exponential relationship to the supply voltage, due to effects such as drain-induced barrier lowering (DIBL) [3]. When the supply voltage is reduced circuits dominated by switching power will also benefit from a reduction in the energy per computation figure. However, as the maximum operating speed drops, leakage power increases in proportion, thus leading to a minima condition for the energy per computation. When the nominal supply voltage was around 5 V one study [4] indicated that it could be possible to reduce power consumption by several orders of magnitude by reducing the power supply voltage. As nominal core supply voltages are soon creeping below 0.8 V [5], potential gains from supply voltage reduction are reduced in magnitude, but still one can achieve very significant gains.

It is important to note that reducing the supply voltage consequentially decreases the operating speed. For applications in the low to medium performance region, or for applications where high performance is only required occasionally, there are however few reasons to maintain a higher supply voltage than neccessary, as this would simply lead to wasted power. Many applications could thus take advantage of reducing the voltage supply to reduce power consumption or the energy per computation. As reductions in power and energy consumption are beneficial in extending the battery time of battery-operated devices, applications within handheld and portable devices can easily be imagined. Additionally, ultra low power consumption may be an enabler for new classes of devices, powered for instance by energy harvesting mechanisms. Savings on energy/computation figures may also provide substantial savings on the electricity bill of large scale computing farms. A potential example is wireless sensor nodes. Wireless sensor nodes could be used for a wide variety of purposes, with many examples such as [6]: humidity monitoring within agriculture, early forest fire detection within environmental monitoring, or gas leak detection for oil and gas industries.

1.2 Challenges for ultra low voltage CMOS

While scaling down the supply voltage is very effective in reducing the overall power consumption of a circuit, there are several factors that can limit effectiveness and reliability of circuits in nanometer scale CMOS.

When the supply voltage is reduced to below the transistors inherent threshold voltage, the transistors operate in the subthreshold region. In the subthreshold domain it is normal to see great variation in the on-currents of devices, particularily for small transistors. The dominant cause of this is random dopant fluctuations (RDF), i.e. fluctuations in the number and placement of dopant atoms, which are implanted during fabrication in order to set the transistor threshold voltage. The overall effects of RDF can however be tedious to model and simulate, and works focused on subthreshold design can suffer a strong bias if it's disregarded. For minimum size devices RDF can cause variations in the device on-current of several orders of magnitude. For digital circuits this may lead to timing variations of similar magnitude. For synchronous systems the worst case implication is fatal timing violations, as current design practices for determining safe hold times may not be adequate in subthreshold design. Although careful design can limit fatal errors, the main effect of RDF is that the maximum clock speed is drastically reduced, leading to a significant performance drop.

When scaling down the supply voltage we simultaneously reduce the noise margins. At lower voltages the difference between the device on and off currents are also reduced. Combined with the increased current variation induced by random dopant fluctuations this can ultimately lead to an inability of simple gates to yield the correct output voltage.

The current in devices operated in the subthreshold region is limited by the diffusion of available carriers in the channel [7]. Therefore subthreshold devices show a very strong response to changes in operating temperature. At lower temperatures there are fewer carriers available, and the current is greatly reduced, while at higher temperatures an increase in current is seen. This is the opposite of what is seen in superthreshold. When operating a transistor in the superthreshold region, one typically estimates a $\pm 20\%$ change in the device on-current, in a typical range from -20°C to $+85$°C . In the same temperature range a subthreshold current may exhibit variation of several orders of magnitude. While this topic is not given too much time in this thesis, the global variation resulting from temperature change can be handled by several techniques, such as adaptive body biasing, or dynamic scaling of the voltage supply.

Traditional superthreshold sizing strategy, using $L = L_{\min}$ while scaling W, is not necessarily the best approach for nanoscale subthreshold transistors. Specifically, the short channel effect (SCE), the narrow width effect (NWE), and the inverse narrow width effect (INWE) may yield counter-intuitive results. Near the smallest dimensions, the device I_{ds} current may increase with increasing L due to SCE, or even decrease with increasing W due to INWE. If we additionally consider the impact that sizing has on RDF, optimal subthreshold device sizing becomes a rather complex problem.

1.3 A roadmap for this thesis

This thesis is a collection of papers that all relate to ultra-low voltage and subthreshold circuits. The original contributions perhaps most central theme is, specific to subthreshold logic gates and memory, how to mitigate or utilize certain device and processing effects that are typically difficult to handle during design and optimization. Reprints of the individual works are included in Part II. Paper contributions included in this thesis are listed as follows :

Paper I H. K. O. Berge and S. Aunet, "Benefits of decomposing wide CMOS transistors into minimum-size gates" in *NORCHIP, 2009*, pp. 1 –4, Nov. 2009.

Paper II H. K. O. Berge and S. Aunet, "Multi-objective optimization of minority-3 functions for ultra-low voltage supplies", in *Proc. IEEE Int. Circuits and Systems (ISCAS) Symp.*, pp. 2313–2316, 2011.

Paper III H. K. O. Berge, A. Hasanbegovic, and S. Aunet, "Muller c-elements based on minority-3 functions for ultra low voltage supplies", in *Design and Diagnostics of Electronic Circuits Systems (DDECS), 2011 IEEE 14th International Symposium on*, pp. 195–200, April 2011.

Paper IV H. K. O. Berge, M. Blesken, S. Aunet, and U. Rückert, "Design of 9T SRAM for dynamic voltage supplies by a multiobjective optimization approach", in *Proc. 17th IEEE Int Electronics, Circuits, and Systems (ICECS) Conf*, pp. 319–322, 2010.

Paper V S. Lütkemeier, T. Jungeblut, H. K. O. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65 nm 32 b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control," *Solid-State Circuits, IEEE Journal of*, vol. PP, pp. 1 –12, Jan 2013.

Paper VI H. K. O. Berge and S. Aunet, "Yield-oriented energy and performance model for subthreshold circuits with Vth variations," in *Design and Diagnostics of Electronic Circuits Systems (DDECS), 2013 IEEE 16th International Symposium on*, pp. 193–198, April 2013.

All paper contributions (I–VI) relate to ultra-low voltage and subthreshold circuits. Paper I concerns itself with a new opportunity for device sizing that may arise for devices that display the inverse narrow width effect (INWE), allowing two minimum-width transistor to operate with improved characteristics compared to a single wide equivalent. Papers II, III, and IV investigates the application of multi-objective optimization strategies to improve the performance and reliability/yield of subthreshold circuits. Papers II and III explore the subthreshold design space for several implementations of minority-3 gates and Muller C-elements by using multi-objective optimization to uncover the Pareto Fronts of these circuits. Paper IV and V relate to the design and measurement results of subthreshold SRAM. Paper IV covers design improvements and design space exploration for a 9T multi-V_t SRAM cell making it suitable for subthreshold applications at 300 mV. Paper V describes measurement results from a subthreshold VLIW processor, as well as measured results from an SRAM module based on the 9T cell of Paper IV. Paper VI approaches the problem of subthreshold device sizing and voltage selection for minimum energy consumption analytically. This analysis is done taking into account the desired yield, and RDF as the dominant source of variation.

The papers are organized in chronological order, with the exemption of Paper IV, due to it's close thematic and introductory relation to Paper V. For Paper V, SRAM related content has the emphasis in this thesis. To learn more about the subthrehold processor of Paper V, please refer to section 3.5

Additionally during my Ph.D. work I co-supervised the M.Sc. thesis work of Martin Haugland [8]. This work can be considered related to this thesis as it employs multi-objective optimization to size standard cells targeting a subthreshold standard cell library, including layout, synthesis and trial place and route.

The rest of this thesis is organized as follows: In Chapter 2 a brief introduction to the most central topics of this thesis is given. In Chapter 3 each paper contribution is introduced, and a summary of the results are given. Chapter 4 is devoted to a discussion of the thesis contributions, providing further perspectives. The conclusion and recommendations for future work are presented in Chapter 5.

Chapter 2

Background

This chapter very briefly introduces basic concepts central to the paper contributions of this thesis. It is not intended as an exhaustive review, but rather serve as a convenience to the reader providing a more general scope. Short introductions are also provided in the papers.

2.1 Ultra Low Power Design

A short historical account for low power design up until 2003 has been given in [9], of which I give an even briefer, slightly modified account in this first paragraph. In the early days of computing vacuum tubes were used to do calculations and power consumption was a concern. The ENIAC used 18,000 vacuum tubes and consumed 150 kW. By comparison transistors dissipate much less power, typically at least by a factor 1000, and a much greater power reduction is achieved in modern ICs. Since the invention of the bipolar transistors in 1947/1948, and later integrated ICs in 1958/1959, power consumption in computational circuits was for a long time rarely a concern for circuit designers. Ultra low power (ULP) circuit design was however pioneered in the 1960s-1970s, when Swiss watchmakers decided to make an electronic watch. While their first circuits were made in bipolar, they became early adopters of CMOS in 1964. To allow the watch to operate for 1 year on a small battery it had to consume only microwatts. Fortunately it did; their Beta wristwatch operated at 1.3 V and drew only 13 μWof power, while the battery could supply 18 μW. Around 1990-1992, the semiconductor industry became aware that it would be necessary to pay more attention to the power consumption in designs and that cooling might be necessary. The power consumption was continually increasing, along with the speed and complexity of digital processors. Additionally, the market demand for more complex portable devices was growing. Low Power conferences and workshops started appearing around 1993. Many concepts that were discussed then were not really new but were in part the reuse of old techniques with the purpose of achieving low power, e.g. pipelining, parallelism, asynchronous circuits, selection of states for finite state machines, reduced swing and transistor sizing. The Harvard Architecture (which is used in the now popular AVR architecture) was designed in 1939, and almost all early computers used RISC-like instruction sets, achieving a low clocks per instruction (CPI) figure. Pipelining and parallelism for low power were introduced in [10]. Pipelining shortens delay paths and thus allows one to reduce the supply voltage at the same frequency.

Parallelism allows a reduction in frequency for the same throughput, thus also allowing a reduction in the supply voltage. Asynchronous circuits promise the removal of the clock tree, which often is responsible for a major part of the power consumption. Some new, but perhaps obvious concepts were introduced, such as gated clocks and activity reduction. Another new concept was dynamic voltage scaling (DVS), e.g. changing the supply voltage and frequency dynamically to suit the required throughput. Today, many of the more dramatic issues discussed derive from the use of deep submicron and nanometer scale processes, e.g. leakage, delay variations, very low supply voltages, cross-talk, and soft errors.

The Swiss watchmakers continued their work from the late 1960s into the 1970s. In [11] it is described how in particular Dr. Eric A. Vittoz did pioneering work on operating MOSFETs in the weak inversion region [12]. Although analytical expressions for the current in weak inversion, or the diffusion current, was in basic principles derived independently by 1966 [13], Vittoz showed how to utilize this region of operation, using what was then unheard of low supply voltages, achieving a remarkably low power consumption, with applications in miniature portable devices such as hearing aids, wrist-watches and biomedical devices [11]. His work on micropower techniques and near- and subthreshold operation has continued such as in [2, 14, 15] with contributions also to [16].

During the early 2000s attention was again raised when the Massachusetts Institute of Technology's Subthreshold Research Group produced several relatively complex digital ICs capable of operation at very low V_{DD} , such as a 175 mV multiply-accumulate unit [17], and a 180 mV FFT processor [18]. In more recent years a few companies and startups have appeared with either a direct or partial goal of taking advantage of the subthreshold domain. To name a few, Ambiq Micro founded in 2010 advertises a microcontroller with 30 µA/MHz, as well as another product, a real-time clock that can operate with as little as 42 nW [1]. PsiKick founded in 2012 is currently developing "Ultra-Low-Power Wireless Platforms"[2]. Iridium Technologies LLC founded in 2006 is working on producing high-reliability and radiation-hardened circuits capable of reliable subthreshold operation[3].

As subthreshold operation allows the lowest supply voltage it is easy to understand it's allure. This is easy to see when looking at the dynamic power in a switching circuit which can be expressed as [2]:

$$P_{\text{dyn}} = f \alpha C_L V_{DD} V_{\text{swing}} \tag{2.1}$$

Here f is the frequency, C_L is the total capacitive load, V_{DD} is the power supply voltage, V_{swing} is the logic voltage swing (often equal to V_{DD}), and α is an activity factor – a number between 0 and 1 specifying the proportion of how much of the total load is being switched on average in a cycle. For a fixed system configuration with a low frequency capable of operating at 200 mV we can calculate that we can save 96 % of the dynamic power compared to operating at 1 V. In addition to switching power, the total active power may also be considered to contain a component of short-circuit power [2]. This contribution is fairly often considered to be negligible [19, 20], although it can depend on design and application. According to [2] the short circuit contribution should remain below 20 %. For the purpose of this brief

[1] See ambiqmicro.com

[2] See www.iridumtec.com

[3] See iridumtec.com

introduction it will also be neglected. The static leakage power P_{stat} can be expressed as:

$$P_{\text{stat}} = I_{\text{leak}}(V_{DD})V_{DD} \tag{2.2}$$

Here the leakage I_{leak} is expressed as a function of the supply voltage. P_{stat} scales linearly with V_{DD} , only if $I_{\text{leak}}(V_{DD})$ is constant with V_{DD} . However, in deep submicron and nanometer processes $I_{\text{leak}}(V_{DD})$ typically increases exponentially with V_{DD} , which will be further explained in Section2.2. From these two power equations it is easy to see that it is very useful to operate at a low V_{DD} as long as your system requirements otherwise allow this. This is the primary driver for interest in the field of subthreshold CMOS or ultra low voltage (ULV) circuits in general – its promise of delivering extremely low power consumption and energy per computation.

The energy per operation E_{op} can in sequential circuits be calculated as the power dissipated during the clock period $t_{clk} = 1/f_{clk}$ [2]:

$$E_{op} = t_{clk}P_{\text{dyn}} + t_{clk}P_{\text{stat}} \tag{2.3}$$

$$E_{op} = \alpha C_L V_{DD}^2 + t_{clk}I_{\text{leak}}(V_{DD})V_{DD} \tag{2.4}$$

With respect to adjustments of the supply voltage, we can find the supply voltage where minimum energy per operation occurs (V_{opt}), by solving $\frac{\partial E_{op}}{\partial V_{DD}} = 0$ [16]. Finding the minimum energy operating point when taking into account RDF is also considered in Paper VI.

Figure 2.1 shows qualitatively how the components P_{dyn} , P_{stat} , cycle period, E_{op}, and static and dynamic energy components typically scale. This figure is based on simulations in a 65 nm LP technology with typical conditions and no statistical variation, of an 11-stage ring oscillator where the dynamic power.has been scaled down in post-processing, to simulate an activity factor of $\alpha = 0.05$. We can see that the power consumption varies with over 6 orders of magnitude, while the cycle period varies with a little more than 5 orders of magnitude until leakage dominates the power consumption. For the energy consumed per cycle E_{op}, the minimum energy point occurs around 0.4 V, and for high performance duty at 1.2 V the maximum energy occurs at a factor approximately 7× larger. This corresponds to an energy saving of over 85 % if it suits the demands of the application. The activity factor α is important when estimating how low the energy minimum will occur. For $\alpha = 0.2$ the MEP occurs at 0.25 V and saves 92.75 % compared to maximum performance. However, the maximum energy then consumes a factor 4 more, so to keep any gains the increased switching should result in increased throughput.

According to [21] the minimum theoretical operating voltage for CMOS switching circuits for a fan-in of 3 and maximum gain larger than 4 is 83 mV at room temperature, while the practical limit due to PVT variability was estimated at 200 mV. In [22] 100 mV is suggested as a lower practical limit for V_{DD}. Necessarily, with higher reliability demands and increasing system complexity, variability may lead to more adverse results than in [21]. In [23] it is argued that variability and high yield targets may make it impossible to reach the target V_{min} or V_{opt}. That discussion has been given a more quantitative basis in Paper VI.

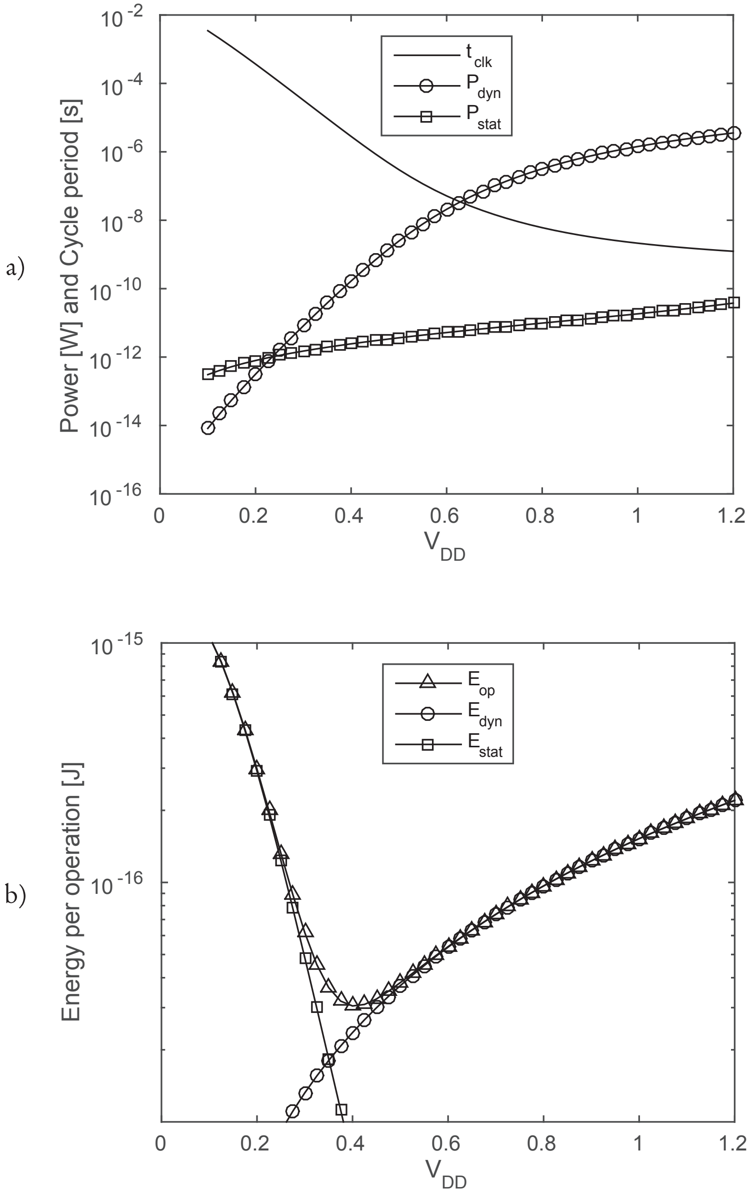


Figure 2.1: Qualitative illustration of power, cycle period, and energy per cycle as a function of V_{DD} , in a switching circuit.

2.2 CMOS Subthreshold operation

Several good introductions to subthreshold operation and design have already been written, for instance [2, 14, 16, 23, 24]. I will however introduce a few concepts central to this thesis

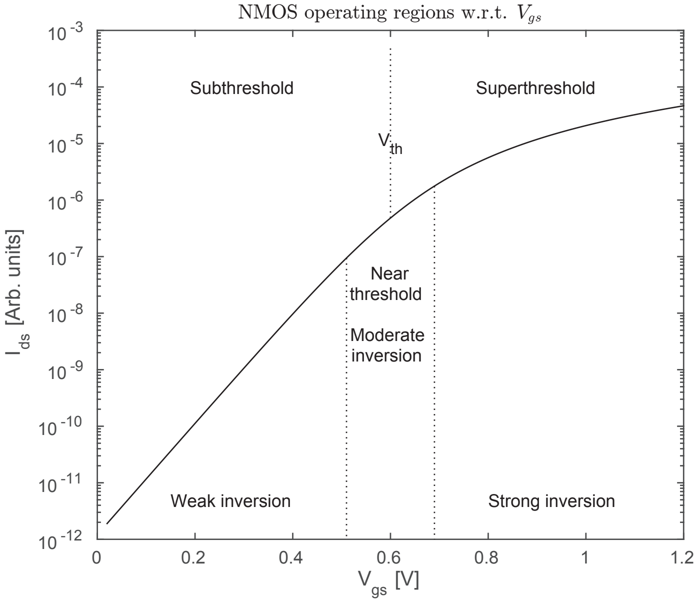


Figure 2.2: Plot of I_{ds} for an arbitrary NMOS device, displaying regions of operation with respect to V_{gs}.

here, as a convenience to the reader.

Figure 2.2 displays the subdivision of operating regions for an NMOS with respect to varying the gate to source voltage V_{gs}. The subthreshold region can be defined as when V_{gs} is smaller than the threshold voltage ($V_{gs} < V_{th}$). The weak inversion region is where the transistor drain current develops exponentially with the gate to source voltage (V_{gs}). However, often in the literature, the subthreshold current is actually referring to the current in the weak inversion region ($V_{gs} < V_{th} - X$), where X is a suitable value allowing the approximation to remain reasonable. In this thesis I also keep this simplification, referring to the weak inversion current simply as the subthreshold current, and I explicitly mention near-threshold or moderate inversion operation when necessary.

The current in weak inversion, or the diffusion current, was in basic principles derived analytically by 1966 [13]. Later, several MOSFET models have added more detail. One model that has been popular with analog circuit designers is the EKV model [15], and it is continuous and differentiable over all regions. In the following we shall however exclusively concern ourselves with weak inversion. A slight reformulation of the expression for the drain to source subthreshold current, excluding moderate inversion, that perhaps is particularly suitable for circuit design and analysis was expressed in [23] as:

$$I_{ds} = \beta I_0 e^{\frac{V_{gs}+\lambda_{ds}V_{ds}}{nU_T}} \left(1 - e^{\frac{-V_{ds}}{U_T}}\right) \tag{2.5}$$

Here V_{gs} is the gate-to-source voltage, V_{ds} is the drain-to-source voltage, U_T is the thermal voltage (kT/q), and λ_{ds} represents the shift of the threshold voltage due to DIBL. The slope factor, n, is given by $(1 + \frac{C_d}{C_{ox}})$ where C_d is the depletion layer capacitance per unit area. Experimental data show that n is also affected by the geometric sizing of the transistor, particularily the length [25]. The subthreshold swing $n U_T \ln 10$ expresses the subthreshold slope in terms of the V_{gs} necessary to increase the current by a decade. Although the ideal transistor would reach 60 mV/decade for $n = 1$, more moderate values for n will result in a larger subthreshold swing, e.g. 83.5 mV/decade for $n = 1.4$, or 101.4 mV/decade for $n = 1.7$. Deep submicron CMOS processes usually involve a poorer subthreshold swing.

β in equation (2.5) represents the tuneable transistor strength as typically seen by the circuit designer:

$$\beta = \frac{W}{L} e^{\frac{\lambda_{bs} V_{bs}}{n U_T}} \tag{2.6}$$

Here W and L are the transistors width and length, V_{bs} represents the body-to-source voltage, and λ_{bs} represents the body-effect on the transistor threshold voltage. The device characteristic current I_0 is given mainly by factors from the process, often outside the circuit designers direct influence:

$$I_0 = (n-1)\mu_0 C_{ox} U_T^2 e^{\frac{-V_{th}}{n U_T}} \tag{2.7}$$

Here μ_0 is the carrier mobility, C_{ox} the oxide capacitance, and V_{th} is the threshold voltage.

For long and wide gates with uniform doping, the threshold voltage V_{th} can be given by:

$$V_{th} = V_{th0} + \gamma_{bs} \sqrt{\Phi_s - V_b s} \tag{2.8}$$

Here V_{th0} is the long-channel threshold voltage for zero substrate bias, and Φ_s is the surface potential. Non-uniform doping effects can be modelled using λ_{bs} and γ_{bs}. Note that $\lambda_{bs} V_{bs}$ and $\lambda_{ds} V_{ds}$ in equation (2.6) is also typically considered a contribution the threshold voltage, and a detailed model is much more complicated. Also obscured by equations (2.6, 2.8, 2.7), V_{th0} is also a function of device W, L sizing. This is particularly relevant at small dimensions. In the next subsection I will however describe the influence these contributions to V_{th} in more detail.

2.2.1 Short and Narrow Channel Effects

When the channel length becomes smaller, or the drain voltage becomes larger, the electric field from the MOS drain terminal to the channel grows in importance. This phenomenon is called drained-induce barrier lowering (DIBL[4]). Eventually, when increasing V_{ds} and/or decreasing L, DIBL will lead to punch-through, as source and drain channel become merged. To counteract the effect of short channel effects such as DIBL and its associated V_{th} roll-off, one can use a larger doping near the source and drain edges of the channel. These implants are called pocket implants (or Halo implants) and are widely used in deep submicron pro-

[4] In an energy band diagram DIBL could be drawn as a drag on the energy bands of the channel (barrier).

cesses [26].

To allow a slightly deeper understand of the various effects that modulate the threshold voltage in deep submicron CMOS devices, we will indulge ourselves with a quick and shallow review of the main contributions to V_{th} as expressed in the BSIM4.6 model.

In the BSIM4.6 model the short channel effect (SCE) on the threshold voltage is modelled separately to DIBL and can be written as [25]:

$$\Delta V_{th,\mathrm{SCE}} = -\frac{0.5\mathrm{DVT0}}{\cosh\left(\mathrm{DVT1}\frac{L_{eff}}{l_{t0}(1+\mathrm{DVT2}\,V_{bs}}\right)-1}[V_{bi}-\phi_s] \tag{2.9}$$

Where L_{eff} is the effective length, l_t0 is the characteristic length, V_{bi} is the built-in voltage of the source and drain junctions. The model parameters DVT0, DVT1, DVT2 are respectively the first, second, and body-bias coefficients for the short channel effect. The effect of DIBL on the threshold voltage is modelled as [25]:

$$\Delta V_{th,\mathrm{DIBL}} = -\frac{0.5}{\cosh\left(\mathrm{DSUB}\frac{L_{eff}}{l_{t0}}\right)-1}[\mathrm{ETA0}+\mathrm{ETAB}\,V_{bs}]\,V_{ds} \tag{2.10}$$

Where the model parameters ETA0 is the DIBL coefficient in subthreshold region, and ETAB is the body-bias coefficient for the subthreshold DIBL effect.

We notice that both the above effects are strongly dependent on short lengths as when L_{eff} approaches zero the cosh function approaches 1. We can also see that the DIBL contribution, that is dependent on V_{ds}, is a weaker effect in subthreshold compared to nominal supply voltage.

For short channels the length dependent effect of pocket (Halo) implants modulates the effect of body bias. This is modelled as [25]:

$$\Delta V_{th,\mathrm{RSCE}} = K1\left(\sqrt{\phi_s - V_{bs}} - \sqrt{phi_s}\right)\sqrt{1+\frac{\mathrm{LPEB}}{L_{eff}}} \tag{2.11}$$

$$+K1\left(\sqrt{1+\frac{\mathrm{LPE0}}{L_{eff}}}-1\right)\sqrt{\phi_s} - K2\,V_{bs} \tag{2.12}$$

Here the parameters K1 and K2 are called the first-order and second-order body bias coefficient (written as γ_{bs} and λ_{bs} in the previous section), LPE0 and LPEB are respectively the zero body bias, and body bias dependent, lateral non-uniform doping parameters. For long channel devices pocket implants can also cause a significant drain induced threshold shift (DITS) [26]. In the BSIM4.6 model the effect of DITS is modelled as:

$$\Delta V_{th,\mathrm{DITS}} = -n\,U_T \ln\left(\frac{L_{eff}\left(1-e^{-\frac{V_{ds}}{U_T}}\right)}{L_{eff}+\mathrm{DVTP0}\,(1+e^{-\mathrm{DVTP1}\,V_{ds}})}\right) \tag{2.13}$$

Here $n\,U_T$ is the subthreshold swing, and DVTP0 and DVTP1 are the first and second coefficient for DITS due to long channels with pocket implants. This effect approaches zero slowly for increasing L_{eff}.

For long narrow-width devices there is a notable contribution from fringing fields along the edge of the channel. This effect depends on the isolation technology, and cause a threshold shift which is modelled as [25]:

$$\Delta V_{th,\mathrm{NWE1}} = (\mathrm{K3} + \mathrm{K3B}\, V_{bs}) \frac{\mathrm{TOXE}}{W'_{eff} + \mathrm{W0}} \phi_s \tag{2.14}$$

Here K3 is called the narrow width coefficient and K3B is the body effect coefficient of K3, while W0 is the narrow width parameter, and TOXE is the equivalent oxide thickness. The main effect of equation (2.14) is an increase in V_{th} at low channel widths. Alternatively the expression can be used for a decrease in V_{th} if we allow for a negative sign for K3, and K3B. For short and narrow devices a reverse width-dependent effect is also modelled as [25]:

$$\Delta V_{th,\mathrm{NWE2}} = -\frac{0.5\mathrm{DVT0W}}{\cosh\left(\mathrm{DVT1W} \frac{L_{eff} W'_{eff}}{l_{t0}(1+\mathrm{DVT2W}\, V_{bs})}\right) - 1} [V_{bi} - \phi_s] \tag{2.15}$$

Here DVT0W and DVT1W and DVT2W are respectively the first, second and body-bias coefficients for narrow width effects on V_{th} in short channels. We see that this contribution grows for small widths, and decreases with V_{bs} (for default negative values of DVT2W).

To summarize, the various short and narrow channel effects contribution to a V_{th} shift in deep submicron devices are numerous and relatively complicated. Partly this is due to non-uniform halo doping which is introduced to counter V_{th} roll-off and punch-through. To visualize what reverse and short channel effects on the I_{ds} current of a 65 nm device may look like, simulation results are displayed in a mesh plot in Figure 2.3. At the lower bounds for both W and L we see the effect of V_{th} roll-off on the drain current. The normal and reverse short channel effect and it's impact on design in subthreshold is discussed in [27], while utilization of the reverse (inverse) narrow width effect is a topic of Paper I in this thesis.

2.2.2 Random Dopant Fluctuations

Although variations in the effective width and length of the transistor contributes to variations in the threshold voltage [28], the dominant source of random variations in the threshold voltage is often a result of the implantation process. Typically for CMOS, the threshold voltage of a transistor is typically set by implanting dopant atoms near or just below the channel-oxide interface [29]. While many aspects of the implantation process is well controlled, some parameters, such as the number of dopants, and their geometrical distribution, relies on random processes. Therefore each transistor will be slightly different from the next and in a macroscopic model they have slightly different threshold voltages.

Threshold voltage fluctuations are often considered to follow the Gaussian (normal) distribution after experimental evidence such as in [30], and “atomistic” simulations such as in [31]. However, one might encounter some skewness depending on the exact nature of the device. More recent experimental results on Vth shifts induced by NBTI such as in [32] displays comparatively fairly strong skewness after stressing devices.

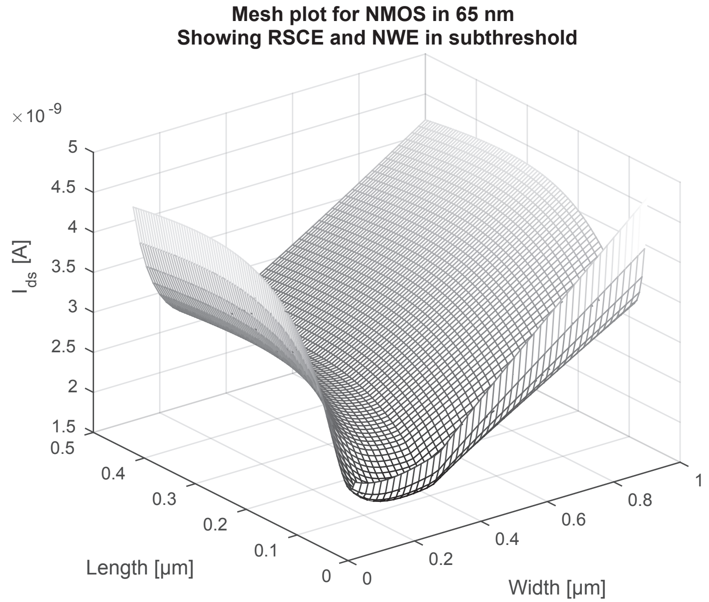


Figure 2.3: Plot of I_{ds} for a LP 65 nm NMOS device that displays both RSCE (along length axis) and NWE (along width axis).

For circuit design the standard deviation of V_{th} , $\sigma(V_t)$, is typically given as [33]:

$$\sigma(V_t) = \mathrm{SD}[V_{th}] = \frac{A_{VT}}{\sqrt{W_{eff}L_{eff}}} \tag{2.16}$$

Here W_{eff} and L_{eff} are the transistor effective width and length. A_{VT} is given by the technology as [34]:

$$A_{VT} = \frac{1}{2}\sqrt[4]{4q^3 N_d \epsilon_{Si} \phi_B}\frac{t_{ox}}{\epsilon_{ox}} \tag{2.17}$$

Where q is the elementary charge, N_d is the number of channel dopants, ϵ_{Si} and ϵ_{ox} are the permittivity of the silicon and the oxide, ϕ_B is the work function, and t_{ox} is the oxide thickness. Technology scaling to smaller geometry processes usually involves reducing the oxide thickness or increasing the oxide permittivity (using high-K dielectric), in an effort to enhance the channel control. Therefore more modern processes usually means a reduced A_{VT}, resulting in that a same-geometry transistor will have a reduced $\sigma(V_t)$in a smaller scale technology. However, if the geometry is scaled down as well, more modern technologies usually display more variation in V_{th} as a result of RDF and downscaling.

For subthreshold operation the variation in current due to RDF can be very large in a minimum size device. Since the subthreshold current is exponentially dependent on the threshold voltage the impact of V_{th} variations due to RDF are much more severe than in the

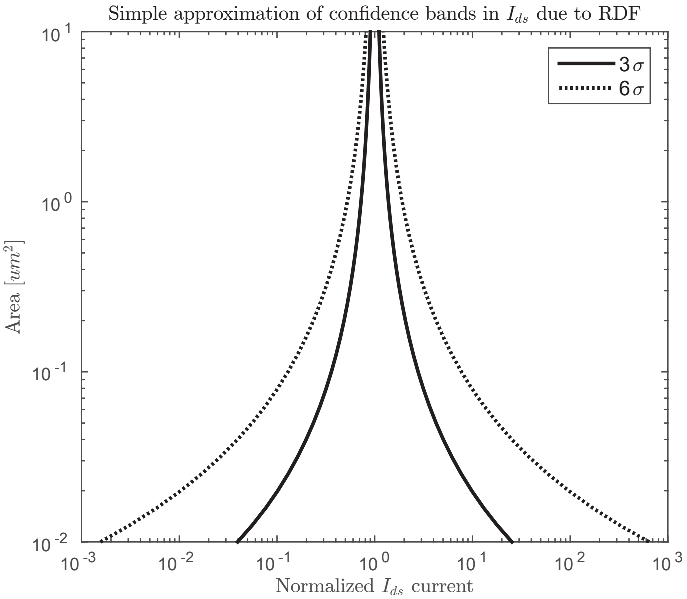


Figure 2.4: Plot of confidence bands using simple theory, for subthreshold I_{ds} normalized to the typical case. ($A_{VT} = 4$ mVµm , $n = 1.7$, T = -20°C)

superthreshold domain. This can have a severe negative impact on the realization of large synchronous digital circuits, and large delay increases may occur. To visualize the impact a plot of the $\pm 3\sigma$ and $\pm 6\sigma$ confidence bands of I_{ds} has been made based on equation (2.5) and (2.16), shown in Figure 2.4.

Naturally it would be welcome if devices could offer less variation in V_{th} . New gate stacks with high-K dielectrics and metals and alloys to set the threshold voltage offer a significant improvement and were introduced in commercial processes in 2007 at the 45 nm node [35]. While the technique is difficult, most technology nodes below 45 nm utilize such techniques today. At moderate technology nodes there are however few if any advertised processes that offers this.

2.3 Multi-objective Optimization

The classical approach to sizing standard CMOS logic cells is to set the length at minimum to minimize input capacitance while maximizing drive strength and focus on symmetric DC curves to simultaneously optimize noise margin and propagation delay [36]. This leaves one free parameter. Knowing that subtle and complex effects cause the DC behaviour of subthreshold MOS devices to vary substantially, a question that may arise is *how can a circuit designer take into account this multitude of effects to optimize a subthreshold circuit?*. One possible answer may be to exploit a multi-objective optimization (MOO) method. Paper II, III and IV apply MOO to circuit design optimization. The papers primarily focus on the application and results, and do not go into details of the algorithms, therefore only a brief introduction is included here.

Multi-objective optimization can be used for problems when optimal decisions need to be taken in the presence of conflicting objectives. This type of problem is inherent to engineering where one seeks to balance performance, cost, risk, and schedule [37]. Many algorithm approaches exist to solving multi-objective optimization problems (MOOPs), such as genetic algorithms, simulated annealing, the complex method, random search, taboo search, and hybrid methods [38]. In the last decade research interest has however increasingly focused on using multi-objective evolutionary algorithms (MOEAs) [39], as they can work on populations of solutions.

Multi-objective optimization seeks to optimize multiple objective functions, under a set of constraints. Mathematically, a multi-objective optimization problem (MOOP) can be described as a minimization problem:

$$\min_{\{x_1,\ldots,x_n\}\in S} \{f_1(x),\ldots,f_k(x)\} \tag{2.18}$$

Here the parameters $\{x_1,\ldots,x_n\}$ form a point in the n-dimensional search space S, also known as the decision space. The objective functions $\{f_1(x),\ldots,f_k(x)\}$ form a k-dimensional function F, the objective space. The solution to the MOOP is called the Pareto[5] set P, while the image of P in F is called the Pareto front. P consists of all non-dominated solutions p in S. A solution is called non-dominated if there exists no other solution that can improve any objective without at the same time worsening another. Conversely, a solution is called dominated if there exists another solution that is not worse for any objective and at the same time improves at least one objectives. For more complex MOO problems, it becomes increasingly hard to find the *true Pareto front*. Therefore, MOO algorithms typically only approximate P.

To explain the above an example Pareto front is shown in Fig. 2.5. Here the objectives delay and power are subject to be minimized with respect to underlying design parameters. In the figure the line represents the true Pareto front, and the points represent the evaluation of solutions p, q and r. In the figure q is dominated by p since both delay and power is improved for p, thus q is not part of the Pareto set or front. The solution p is not dominated by either q or r, and r is not dominated by either p or q. A MOO algorithm can not necessarily see that p is part of the true Pareto front or that r is not. The solution r would therefore also be considered non-dominated and part of the approximated Pareto front until the algorithm can find a new solution that dominates it.

It is fairly common to use MOEAs to solve complex MOOPs by approximating the Pareto Front. Genetic algorithms such as the *Non-dominated Sorting Genetic Algorithm-II* (NSGA-II) [40] and *Strength Pareto Evolutionary Algorithm 2* (SPEA-2) [41] have become standard approaches, and many other algorithms and variants exist.

For standard cell optimization the GAIO software package was used in [42]. There, after an initial search in a grid, the search space is iteratively subdivided in smaller and smaller subspaces, so only subspaces that contained Pareto points is searched. This approach seems

[5] Historically, Francis Y. Edgewick and Vilfredo Pareto are credited with the introduction of the concept of non-inferiority in the context of economy [37].

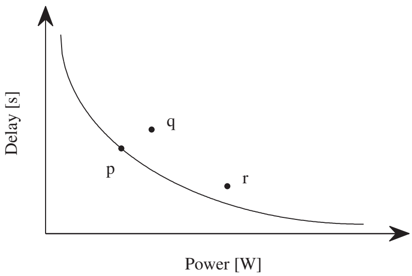


Figure 2.5: An example Pareto front for two conflicting objectives, delay and power, and the tentative solutions p, q and r.

effective when there is an expected relationship between parameters, and will have some obvious advantages when the number of solutions are large.

Naturally, it makes sense to employ MOO only for conflicting objectives. If two objectives are related such that improving one objective will always improve the other, then it is sufficient to include only one objective in the MOOP. In the context of optimizing fuzzy control systems [43] quotes MOO evolutionary algorithms as usually being very good at handling two or three objective functions, whereas when the number of objectives increases, almost all solutions become non-dominated, thus their search capacity worsens. To handle this one may choose to ignore some objectives, or integrate several aspects into one objective. These methods often work well if some objectives are statistically insignificant, or if they are related.

For the purpose of optimizing CMOS logic circuits multiple performance criteria exist. In broad terms we often seek to optimize on criteria such as delay, active power, leakage power, layout area, and reliability. There are many ways to express these criteria into objective functions for a MOOP, and several of these objectives have some relation. E.g. minimizing the area for a given delay will typically simultaneously minimize the active power, unless leakage is a significant contributor. Fairly recently MOO was used for resource efficient design of standard cell libraries, both for standard and subthreshold operation [36, 42, 44]. The approach allows for a well-informed, and balanced selection of optimal sizings for standard cells. In [36] the objectives noise margin, dynamic energy and propagation delay are optimized. However noise margin and propagation delay appear related, thus one could perhaps get similar results with one less objective or by combining the objectives.

Naturally also the number of parameters influence the complexity and hence the search capacity and execution time. If there exists algebraic combinations of parameters that evaluate to equal and Pareto optimal objective values, and the parameters are continuous, then there are also infinitely many solutions to the MOOP. On the other hand, if such an algebraic combination is known, then one may set this as a constraint and reduce the number of parameters thus simplifying the search.

To summarize, multi-objective optimization problems typically have more than one solution. MOEAs represent an efficient method to approximate all the best available resource-efficient tradeoffs in an engineering design. At todays status MOEAs can be effective in handling multi-objective optimization problems with up to four objectives.

2.4 Ultra Low Voltage SRAM

Static random access memory (SRAM) plays a key role in many digital systems, supporting volatile storage in applications such as instruction memory, data memory, cache, FIFOs, register files and scratchpad memories. The ability to reduce the supply voltage of SRAM modules is interesting for several reasons; to reduce leakage during inactive standby modes while retaining the contents of memory, to reduce access energy when only low throughput is required, and/or to operate at the same supply voltage as other intra-die ULV circuits. When scaling down the supply voltage of digital circuits, the SRAM minimum operating voltage $V_{\min}$, is however often considered the limiting factor [45].

The typical organization of an SRAM module is depicted in Figure 2.6. The SRAM 1-bit memory cells (bit cells) are organized in an array with rows and columns. Typical control signals include chip select (CS), write enable (WE), a clock (CLK), and an address (ADDR). When a read or write access is performed the address is split up into two parts, a row address and a word address. The row decoder decodes the row address and enables a signal along the appropriate row, this signal is typically called the wordline (WL). For read access data is output by each cell of the active row using the bitline (BL) signals, oriented along the columns. The value from each cell is typically amplified by a sense amplifier. A multiplexer (before or after the sense amplifiers) uses the word address to select a subset of the columns to output as a data word. During write access the bitlines are actively driven in order to overpower the cell and write a new value, logic ’1’ or ’0’.

The energy per access in an SRAM module depends on the number of rows and columns in the bit cell array. When a full row is accessed for read, the switching energy per access in a cycle can be estimated as:

$$E_{read} \approx E_{ctl\&dec,r} + N_C C_{WL,bit} V_{DD}^2 + N_C \left[N_R C_{BL,bit} V_{DD} \Delta V_{BL} + E_{sense\&output} \right] \qquad (2.19)$$

Here N_R and N_C are the numbers of rows and columns, $E_{ctl\&dec,r}$ is the switching energy from the decoder and control circuitry, $C_{WL,bit}$ is the wordline capacitance per bit, $C_{WL,bit}$ is the bitline capacitance per bit, ΔV_{BL} is the bitline swing, and $E_{sense\&output}$ represent the energy from the sense amplifier and other output stages. We see in this expression that when both N_R and N_C become large the bitline capacitance of the bit cell will become the dominant term. To reduce the impact of this term it is customary to use a thin cell layout, and to reduce the bitline swing by employing sense amplifiers that can amplify a smaller differential voltage.

Figure 2.7 shows the conventional 6T SRAM cell, which uses two back-to-back inverters to store 1 bit on the complementary retention nodes (Q and $\overline{\text{Q}}$). Writing is performed by raising the wordline (WL), while forcing complementary value on the bitlines, BL and $\overline{\text{BL}}$. In order for the write to be successful, primarily the access transistors $\text{M}_{5,6}$ need to be stronger than $\text{M}_{3,4}$. A read can be performed by raising both BL and $\overline{\text{BL}}$ to the supply voltage, and then raising WL. The current through $\text{M}_{5,6}$ is integrated on the bitline capacitance and produces a smaller differential voltage. This differential voltage is then amplified by the sense amplifier and then latched as a digital value. Usually the bitline capacitance is large, thus the access

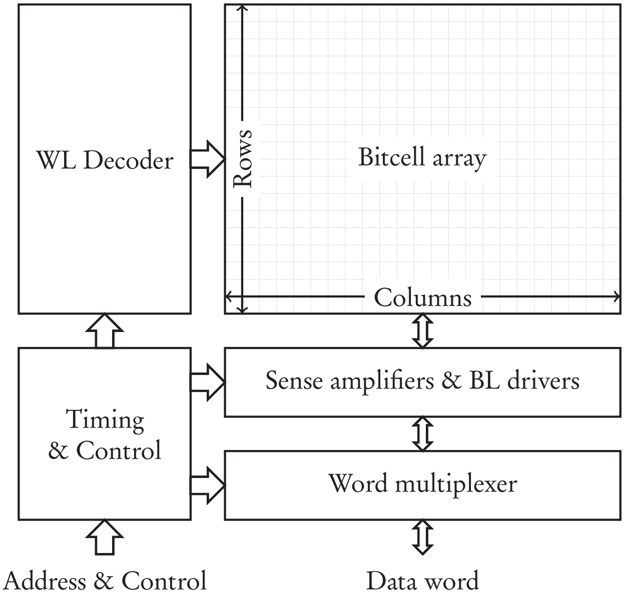


Figure 2.6: Overview of typical SRAM module organization, components and layout.

transistors $M_{5,6}$ will for a while be able pull the drain of $M_{1,2}$. Therefore to avoid overwriting the cell during read, the retention transistors $M_{1,2}$ must be stronger than $M_{5,6}$. The fact that read stability depends on weak $M_{5,6}$ and write stability depend on strong $M_{5,6}$, represents a conflict. When the margins are reduced, such as when reducing the cell area or reducing the voltage supply, it eventually becomes difficult or impossible to retain reliable access to the cell.

When the DC transfer function for the two back-to-back inverters of the 6T cell are plotted against each other, such as in Figure 2.8, the static noise margin of an SRAM cell can be defined as the diagonal of the maximum square that can be fitted within the two curves, on both sides of the tripping point. If there is no such square on either side, the cell is unstable and will not be able to retain data.

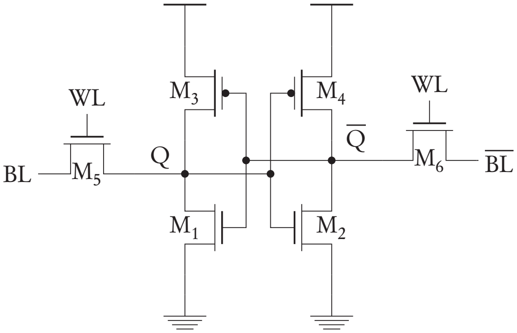


Figure 2.7: Conventional CMOS 6-transistor (6T) SRAM

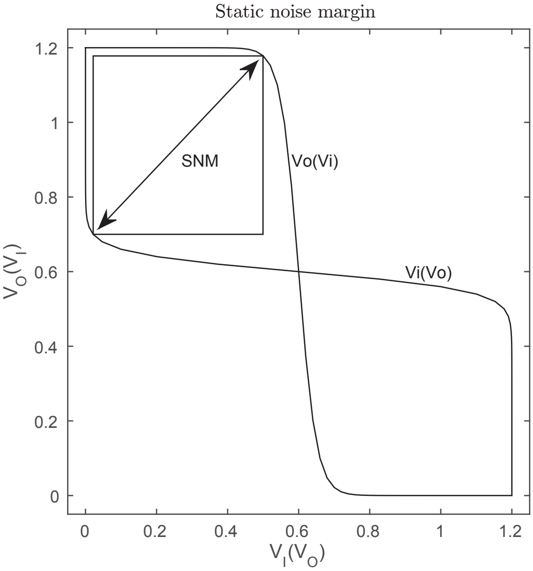


Figure 2.8: Static noise margin (SNM)

An effective way to simulate SNM can be found in [46]. The same method can also be used to simulate an SNM for the read and write operation. SNM_{read} is found when both wordlines and bitlines are forced to the supply voltage. $\text{SNM}_{\text{write}}$ is found when wordlines are at the supply voltage and bitlines are forced to complementary logic values that should cause the cell to be overwritten. In the case of $\text{SNM}_{\text{write}}$ it should not be possible to fit a square within the two DC curves, therefore $\text{SNM}_{\text{write}}$ is often represented as a negative value.

Techniques to improve writeability of the 6T cell include boosting the WL voltage [47], or through a virtual supply rail for the retention inverters, either collapsing V_{DD} or boosting ground [48]. Similarly dual-rail schemes can also allow enhanced readability if a virtual V_{DD} is raised prior to access, or a virtual ground is lowered; also a WL underdrive scheme will result in improved read stability [48].

To circumvent the conflict between read and write stability it is also possible to introduce extra transistors. As seen in Figure 2.9 the additional read transistors in the 8T SRAM cell allows for a much improved read stability at low voltages [49, 50]. In [51] this was used together with sizing utilizing RSCE, to achieve a V_{min} of 260 mV. Another consideration that can improve margins is optimizing device threshold voltages, and utilizing multiple threshold voltages [52].

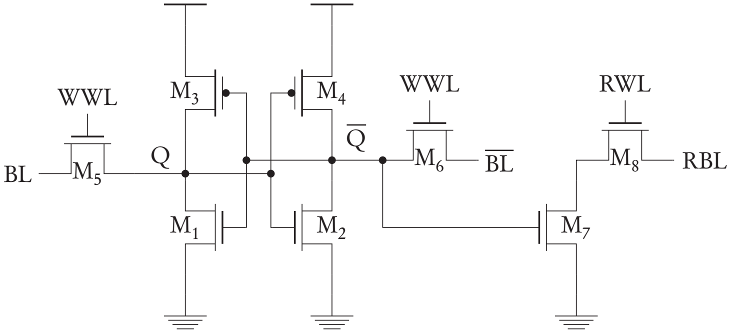


Figure 2.9: 8T SRAM cell with separate read and write wordlines (RWL,WWL) and separate read and write bitlines (BL, RBL)

Chapter 3

Summary of paper contributions

3.1 Paper I : Benefits of Decomposing Wide CMOS Transistors into Minimum Size Gates

3.1.1 Introduction

Paper I [53] (see reprint on p. 53) focuses on exploiting the inverse narrow width effect (INWE) of MOSFETs for subthreshold operation. NWE and INWE is typically a concern in narrow devices made using shallow trench isolation (STI). INWE is quoted to cause a field-enhancement that reduces the threshold voltage of the device [54]. For superthreshold operation this is a concern as it can cause a considerable increase in the subthreshold leakage current [55]. For pure subthreshold operation the effect can however be beneficial, as both leakage and drive current see an increase while parasitic capacitances are reduced. The reduction in parasitic capacitance allows a substantial reduction in the switching energy or Power-Delay Product (PDP), for logic gates drawn using minimum-width devices of a type that displays the INWE effect. The combination of increased currents and reduced parasitic capacitances allows faster operation, but with an increase in leakage.

To highlight potential benefits and drawbacks of using minimum width transistors the paper investigates the properties of drive strength multiples through parallel coupling of minimum width transistors. In the paper this is called a minimum-split transistor (MST). The MST is compared to an iso-area design where the drive strength is adjusted by increasing the width of a single transistor. A Monte Carlo experiment conducted in MatLab features some statistical properties of the drive current of MSTs.

3.1.2 Summary of results

Main results of Paper I include an inverter design that for iso-areas exhibited reduced delays (by 35 %–40 %) and reduced PDP (by 40 %–43 %) for the switching of an inverter in a ring oscillator experiment. The Monte Carlo experiment shows that multiple MSTs have similar or improved worst-case on-currents when compared to iso-width WTs.

3.1.3 Errata

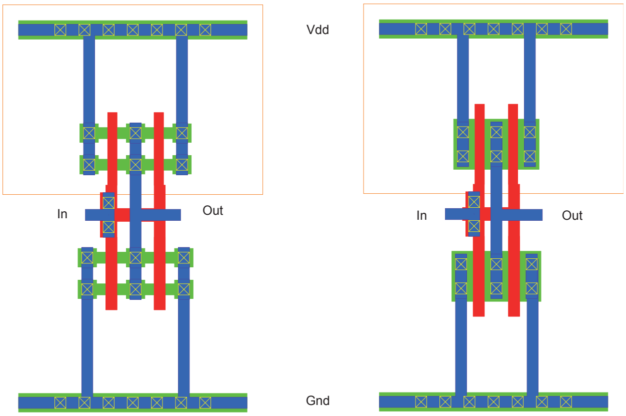
1. In Paper I, Fig. 2, 0.6m should have read 0.6 µm .
2. In Paper I, In the legend of Fig. 4, the words mean and nominal were accidentally switched.

3.1.4 Postscript : Measurements

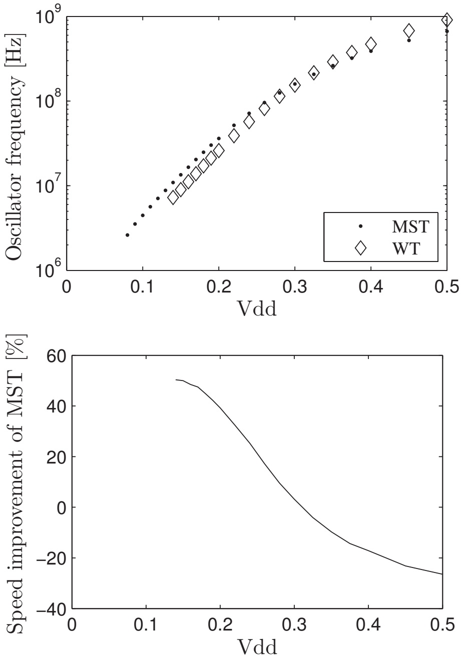
A few months prior to the writing of Paper I, two ring oscillators were sent to manufacturing. The transistor topologies are slightly different to those of Paper I. However, the measurement results still display positive results for the MST variant, at the lowest voltages.

In the experimental setup two ring oscillators were implemented each with 11 stages. For the MST inverter both PMOS and NMOS had 4 parallel transistors of minimum width. A layout of the two inverter cells can be seen in figure 3.1a. A low threshold voltage device was used for the PMOS and a standard threshold voltage device was used for the NMOS. The PMOS employed minimum length, while the length of the NMOS was slightly increased to 0.12 µm to allow equal pull-up and pull-down currents for the MST inverter at $V_{DD} =$ 300 mV. The wide gate structure had two gates 0.54 µm wide for both PMOS and NMOS, and a shared drain junction. In both oscillators the inverters were spaced equidistant to make wiring identical.

Measurements of the oscillator frequencies as a function V_{dd} is shown in the top part of figure 3.1b. The WT design was functional down to a minimum supply voltage of 140 mV. At this V_{DD} the WT oscillator frequency (f_{WT}) was 7.25 MHz while the MST design was 50.3 % faster operating at $f_{MST} =$ 10.9 MHz at the same operating voltage. The minimum V_{DD} for the MST design was about 80 mV, and at this voltage it was running at $f_{MST} =$ 2.62 MHz. The relative speed difference between the two designs decreased when the power supply voltage was increased. At about 311 mV the two oscillators were operating at the same frequency of about 178 MHz. At 500 mV the WT design was running at 908 MHz while the MST design was about 26.4 % slower, running at 668 MHz. Unfortunately it was not possible to do measurements of the power consumption as for this multi-project chip, the power supply was required to be connected to other circuits through the pad ring.



(a) Layout of manufactured inverters. MST (left) and WT (right).



(b) MST and WT operating frequency (top) and relative speed difference (bottom).

Figure 3.1: Layout and measurements of MST and WT ring oscillator.

3.2 Paper II : Multi-Objective Optimization of Minority-3 Functions for Ultra Low Voltage Supplies

3.2.1 Introduction

When optimizing multiple and different performance goals, it can become a time-consuming task to find good and appropriate trade-offs for several conflicting performance goals. Many traditional circuit optimization techniques focus on one performance variable, while other performance measures are regarded as constraints which a designer must set from a priori knowledge of the circuit and the environment it is expected to operate in. Multi-objective optimization takes a step back and allows search for multiple solutions in a multidimensional performance space. Only Pareto optimal solutions are kept, that is to say the result of algorithm will only be efficient trade-offs where each potential solution is not dominated by any other potential solution on all performance measures. Following a multi-objective optimization run, a designer or perhaps a place and route algorithm, may from the results choose a candidate particularly suited to the application it is intended for.

Paper II [56] (see reprint p. 59) presents a method used to explore the performance space of three implementations of the minority-3 logic function operating at a supply voltage of 150 mV. Three conflicting design goals; area, leakage power, and a measure for robustness, are evaluated for each circuit. These performance goals are co-optimized using multi-objective optimization so the optimization result for each circuit is a Pareto Front. At specific performance values we further evaluate each circuit with a ring oscillator experiment, where we include measures of the ring oscillators frequency and energy consumption.

3.2.2 Summary of results

The presented multi-objective optimization method was developed to co-optimize three conflicting design objectives; circuit area, static leakage power, and robustness. The robustness measure is novel and was developed primarily as an optimization goal to compare between the circuits abilities to present the ideal logic output voltage. All performance measures are evaluated several times for the same circuit in a Monte Carlo run. The approximated Pareto front resulting from the multi-objective optimization are plotted for all circuits. We found that for the supply voltage of 150 mV the performance space was dominated by the 10T mirrored gate implementation [57] at small circuit areas. At larger areas a 22T standard CMOS implementation using inverters and 2-input NAND and NOR gates, was potentially more robust. While the optimization method did not take speed into account, the 12T mirrored gate implementation [58] was slightly faster than the 10T implementation in the iso-area ring oscillator experiment.

3.2.3 Postscript

Following the presentation at ISCAS further analysis of the data in this paper was performed. Specifically we investigated the width and length sizing parameters of the circuits that were part of the Pareto front. For series transistors, the width parameters W_{dp}, W_{dn}, and the

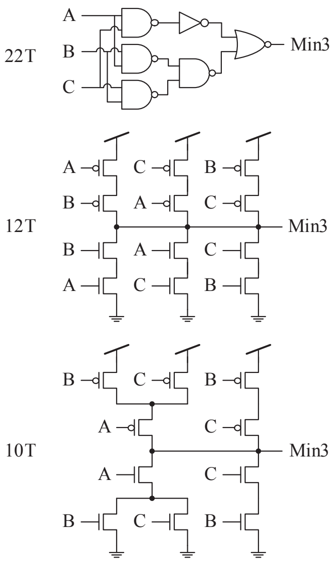


Figure 3.2: Minority-3 functions investigated in Paper I. Named after the number of transistors in the circuit: 22T, 12T, 10T.

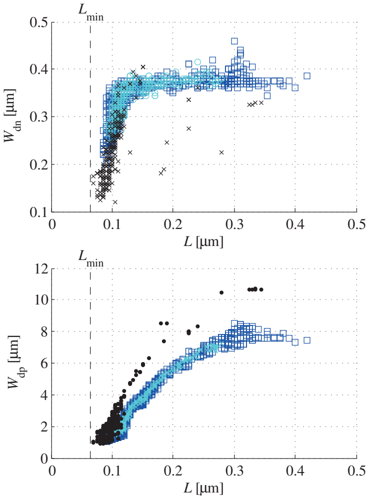


Figure 3.3: Pareto sets showing width and length parameters W_{dn}, W_{dp}, L corresponding to the Pareto Front in Fig. 2 of Paper II for areas less than 40 µm^2 and RMSEs less than 4 mV. (Marker symbols purely for visibility).

common length parameter L, are shown in Figure 3.3. The particular optimization problem, as stated in the paper, tries to optimize gates for maximum stability (minimizing “RMSE”) while simultaneously minimizing power and area. This figure is a good example to show that the optimization method is able to choose parameter values that would perhaps deviate from a conventional sizing method, incorporating short- and narrow-channel effects, as well as variability based on results from the Monte Carlo simulation. The algorithm has preferred to scale NMOS in primarily two regimes; first for small area transistors L is kept slightly above L_{min} while W is scaled until it reaches approximately 370 nm, second for larger area transistor, it keeps W constant around 370 nm and increases L. For PMOS minimizing RMSE, power and area results in a W_{min} just under 1 µm . For larger areas both W and L are then both scaled, although the ratio decays somewhat. Ratios W_{dp}/W_{dn} are typically in the range 5 × –20×. This is not an untypical ratio for subthreshold sizing according to [59].

3.2.4 Errata/Note

1. In Fig. 2 the Pareto fronts are referred to as Pareto sets. This is not entirely uncommon, however it differs from the terminology applied elsewhere in this thesis.

3.3 Paper III : Muller C-elements based on Minority-3 Functions for Ultra Low Voltage Supplies

3.3.1 Introduction

Paper III [60] (see reprint p. 65) expands on the results of Paper II. Results from the multi-objective optimization of Paper II are given for two additional implementations of the minority-3 function. This includes a 6T 'ratioed-fight' implementation, and a variation over the 12T implementation of Paper II. From the five different minority-3 implementations, a 10T and 22T implementation is selected to form two-input Muller C-elements which the paper refers to as the 12T and the 24T Muller C-element, as the Muller C-elements are formed by adding a two-transistor inverter and a feedback connection. Muller C-elements, named after D. E. Muller who introduced them in [61], is a logic function that switches state only when all its inputs hold the same logic value. Muller-C elements can therefore be used as memory elements. One common use is to facilitate correct timing for asynchronous logic [62]. As such Muller-C elements are common building blocks of Null Convention Logic[63].

3.3.2 Summary of results

The two candidates (12T and 24T) for an Ultra-Low Voltage Muller C-element were resized to accommodate equal rise and fall times, and were further investigated under operation at 150 mV and 300 mV and at three different temperatures, -20°C, 27°Cand 85°C. The influence of process variations was investigated using Monte Carlo simulations and characteristics for switching energy, power consumption and propagation delay are evaluated.

At room temperature and a supply voltage of 150 mV the 12T implementation of the Muller C-element had a switching delay of approximately 16.22 μs , approximately 10 % faster than the 24T implementation. The 12T's static power consumption was on average 2.62 pW, or just 35 % of that of the 24T implementation. The switching energy was also approximately 44 % lower in the 12T implementation. The relative comparison was fairly constant at the other temperatures and at the increased a power supply of 300 mV. At 300 mV the absolute switching energy was however approximately quadrupled, while propagation delay was reduced with a factor of about 5 ×. Process variations and the temperature had a very strong influence on the propagation delays, ranging over 3 orders of magnitude from worst-case at -20°Cto best-case at +85°C.

Overall the 12T implementation, also known as the Sutherland implementation [64] appeared to be the superior candidate for an ULV Muller C-element although the results of the multi-objective optimization indicate through the robustness measure that the 24T implement may be a more robust solution at large circuit areas.

3.4 Paper IV : Design of 9T SRAM for Dynamic Voltage Supplies by a Multiobjective Optimization Approach

3.4.1 Introduction

Memory is a central component in modern microsystems and SRAM is a typical choice for applications such as cache or FIFOs. Low voltage operation of SRAM is desirable to reduce leakage power in standby and low power modes, or simply to allow compatibility with other circuits operating at low voltages without the need for interface circuitry such as level shifters. It is also possible to reduce access energy in SRAM by operating at lower voltages, this is however limited to small to moderately sized SRAM blocks (cell arrays) as the leakage energy per access proportion grows with memory size.

For operation at very low supply voltages (below approximately 600 mV for 65 nm and 90 nm technologies) the traditional 6T SRAM cell quickly runs into reliability issues as read and write reliability are in conflict by design, i.e. increasing read reliability will lower write reliability, and vice versa [65]. This conflict can be solved by using additional transistors to decouple the read access from the cell, as for instance in [65, 66].

Paper IV [67] (see p. 73) focuses on the design of a 9T SRAM cell, with an emphasis on reliable operation at 300 mV. The 9T SRAM cell solves the read-write reliability conflict by decoupling the read-out signal so current is provided externally to the back-to-back memory-retaining inverters. Thus the 9T cell also features independent read and write access. Compared to the 8T SRAM cell of [66] the readout signal is differential, thus allowing traditional sensing techniques for the bitline.

In the paper a multi-objective optimization approach is used for transistor sizing. This method provides a global picture of all Pareto optimal solutions and allows us to present a variety of resource efficient implementation choices. A sample Pareto optimal cell is investigated further in the paper, using Monte Carlo simulation.

Paper IV is a paper written in collaboration with the Schaltungstechnik research group from the Heinz-Nixdorf Institute, University of Paderborn. Additional and comprehensive information on the multiobjective optimization method related to this paper is available in German in Matthias W. Blesken's Ph.D thesis [42].

3.4.2 Summary of Results

A main result in the paper is the multi-objective optimization method that was designed to facilitate a choice of resource efficient sizings for the 9T SRAM cell. The method optimizes $\sigma(\mathrm{SNM}_{\mathrm{hold}})$, cell area, $\mathrm{SNM}_{\mathrm{write}}$, and leakage power and read and write delay times at supply voltages of 300 mV and 1.2 V. A sample resource efficient sizing was investigated more thoroughly. Here, the leakage per cell was very low at 0.26 pW. This is two orders of magnitude lower than the design references that were used for comparison [47, 66, 68, 69], although the references include leakage from peripheral access circuits. Stability was indicated as good with the typical $\mathrm{SNM}_{\mathrm{hold}}$ value located at $8.56\,\sigma$ away from the typical simulation failure criterion. Read and write delay optimization goals have typical values at 0.268 µs indicating an expected performance comparable and/or improved with respect to [47, 66, 68, 69].

3.4.3 Supplement: Deriving equation (2) of paper IV

We assume that variability in the threshold voltage (V_{th}) of a transistor can be modeled as a random variable following a normal distribution, with the standard variation of the threshold voltage modeled using equation (2.16):

$$\sigma(V_{th}) = \frac{A_{VT}}{\sqrt{W_{eff}L_{eff}}}$$

We also assume that the static noise margin (SNM) can be expressed on the linear form:

$$\text{SNM} = R_p V_{th,p} + R_n V_{th,n} + M \tag{3.1}$$

where R_p, R_n and M are constants. Since the means are constants we can rewrite this as:

$$\text{SNM} = R_p \Delta V_{th,p} + R_n \Delta V_{th,n} + M' \tag{3.2}$$

Where M' is a constant that incorporates the means, and the ΔV_{th} values are random variables. The variance of the sum of two independent RVs taken from the normal distributions $a\mathcal{N}(0,\sigma_1)$ and $b\mathcal{N}(0,\sigma_2)$, where a and b are constants (linearly transforming the normal distribution), can be given by:

$$\sigma_\Sigma^2 = a^2\sigma_1^2 + b^2\sigma_2^2 \tag{3.3}$$

The variance of the sum of the RVs ($R_p\Delta V_{th,p} + R_n\Delta V_{th,n}$) can thus be expressed:

$$\sigma_\Sigma^2 = R_p^2(\sigma(V_{th,p}))^2 + R_n^2(\sigma(V_{th,n}))^2 \tag{3.4}$$

Using equation 2.16 we then insert for $\sigma(V_{th})$ and get:

$$\sigma_\Sigma^2 = \frac{R_p^2 A_{VT,p}^2}{W_p L_p} + \frac{R_n^2 A_{VT,n}^2}{W_n L_n} \tag{3.5}$$

Comparing this to equation (2) we see that we should have:

$$\begin{aligned} \sigma_n^2 &= \frac{R_n^2 A_{VT,n}^2}{W_{0,n} L_{0,n}} \\ \sigma_p^2 &= \frac{R_p^2 A_{VT,p}^2}{W_{0,p} L_{0,p}} \end{aligned} \tag{3.6}$$

To find σ_n we fix $W_{0,n}, L_{0,n}$, $W_{0,p}$ and $L_{0,p}$, and determine σ_n by setting $A_{VT,p} = 0$, running a Monte-Carlo simulation and calculating the resulting standard deviation of SNM_{hold}. We determine σ_p in an equivalent manner by setting $A_{VT,n} = 0$. Since the lengths $L_1 = L_{0,n} = L_3 = L_{0,p}$ are constant and equal in the application in Paper IV we can simplify equation (3.5) to equation (2).

3.5 Paper V : A 65 nm 32 b Subthreshold Processor with 9T Multi-Vt SRAM and Adaptive Supply Voltage Control

3.5.1 Introduction

Paper V [70] (see p. 79) presents overall design methods and measurement results from a 32 b very large instruction word (VLIW) 65 nm subthreshold processor [71], together with a 64 × 32 block of the 9T SRAM cell [67]. Paper V is a paper written in collaboration with the Cognitronics and Sensor System Group, University of Bielefeldt, and the Schaltungstechnik research group from the Heinz-Nixdorf Institute, University of Paderborn. While this thesis only concerns itself with content relating to the SRAM block of this paper, additional and comprehensive information on the subthreshold processor can be found, in German, in Sven Lütkemeier's Ph.D. dissertation [72].

Micropower processing in the subthreshold domain can be very beneficial, as reduced supply voltages often yield large savings in the energy per computational operation. Savings on the order 10×–20× [73] in the energy per operation compared to operation at nominal supply voltages of 1.2 V. A subthreshold processor could therefore be natural as part of many computational systems aiming at minimal energy and/or power consumption. The processor presented in the paper is a subthreshold implementation of the CoreVA architecture [74].

SRAM is an indispensable part of many digital systems, with important applications such as in cache, FIFO buffers or register files, and it is of interest to be able to operate SRAM at the same voltage as surrounding circuits. Design in the subthreshold domain is demanding as increased susceptibility to variations, particularly in the transistor threshold voltage [23], incurs large timing variations and present hazardous sources of error for hold and setup timing closure. One must therefore take into account this increased variation and mitigate the impact on failure rates and delay variations. The paper also details a performance and power management subsystem that provides dynamic voltage and frequency scaling (DVFS) combined with an adaptive supply voltage generation for dynamic PVT compensation.

3.5.2 Summary of results

The processor was synthesized using standard cells with high efficiency ensured by a multi-objective optimization strategy presented in [44]. In measurement the CoreVA processor achieves a minimum energy per instruction of 9.94 pJ at 325 mV. The lowest operating voltage is 200 mV for best samples with a clock frequency of 10 kHz, while mean speed is 94.32 MHz at 1.2 V. Average energy per cycle is 110.22 pJ at 1.2 V, a factor of 11.1 compared to the minimum energy operation point.

The 2 kb 9T SRAM macro presented in the paper achieves minimum energy per operation at averages of 321 mV (0.03 σ/μ), 0.57 pJ (0.037 σ/μ) and 730 kHz (0.184 σ/μ). Maximum operating frequencies at the minimum operating voltage fell in the range from 448 kHz to 1016 kHz. All 38 measured samples were functional from 1.2 V down to 280 mV. Best samples operated error-free down to 230 mV. At 171 mV 1 % of the bits experienced retention errors. Average leakage per cell was 17.8 pW at 0.3 V.

3.6 Paper VI : Yield-Oriented Energy and Performance Model for Subthreshold Circuits with V_{th} Variations

3.6.1 Introduction

Paper VI [75] (see p. 93) extends previous work [76, 77] to model also the impact of RDF on various performance metrics for subthreshold logic circuits. Although many sources of contribute to variability in nanoscale subthreshold circuits, RDF is often the dominant source of intra-die current and delay variation [23]. Investigating this topic is important as it helps in understanding how overall system performance degrades when scaling up the number of transistors in a sequential circuit.

The analysis consist of several steps. First fairly basic subthreshold current equations [23] are developed in order to express a worst-case current based on V_{th} variance and gate area. Thus later metrics can be calculated taking into account a targeted circuit complexity and yield, as well as gate sizing. In the analytical expressions the propagation delay is lognormally distributed. Using an approximation for the sum of iid. lognormals we evaluate a worst-case clock period as a function of number of gate delays in series and load capacitance. We then follow steps similar to [76] in order to find analytical expressions for optimal supply voltage for minimum energy operation.

3.6.2 Summary of results

The main results of Paper VI are the analytical expressions developed during the analysis. These enables a designer to fairly quickly evaluate relevant worst-case conditions in the subthreshold domain. The analysis derives worst-case expressions for: on-current (I_{on}), on/off ratios (I_{on}/I_{off}), propagation delay of simple gates (t_p), clock period for series-connection of identical gates (t_{clk}), and energy per cycle (E_{op}). Results are presented in contour plots, visualizing the subthreshold digital logic design space in a 90 nm process, with respect to gate area and supply voltage parameters. An analytical expression is found for the optimum voltage supply for minimum energy operation, given that it operates on the worst-case clock period t_{clk}:

$$VDD_{opt} = nU_T\left(2 - W_{-1}\left(-\frac{P_\alpha}{NMG(\cdot)}\exp\left(\frac{A_{VT}^2}{2WLn^2U_T^2} - Q\frac{A_{VT}}{\sqrt{WL}nU_T} - 2\right)\right)\right) \quad (3.7)$$

This expression takes into account threshold voltage variations which to our knowledge has not been done analytically in previous works, although it significantly affects the minimum energy operating point. Since the analysis relies on subthreshold current expressions, near- and super-threshold operation is not taken into account. For pure subthreshold circuits we find that V_{th} fluctuations significantly increases the energy per cycle, and also that V_{th} variation increases the required V_{DD} and gate area to achieve minimum energy operation. For the technology we evaluated, we found that gate sizes several times the minimum may be required for minimum energy designs, primarily as it reduces the impact of RDF on delay. Importantly the paper may help the reader gain an understanding and intuition for very significant yield-related effects applicable to design of circuits in the subthreshold domain.

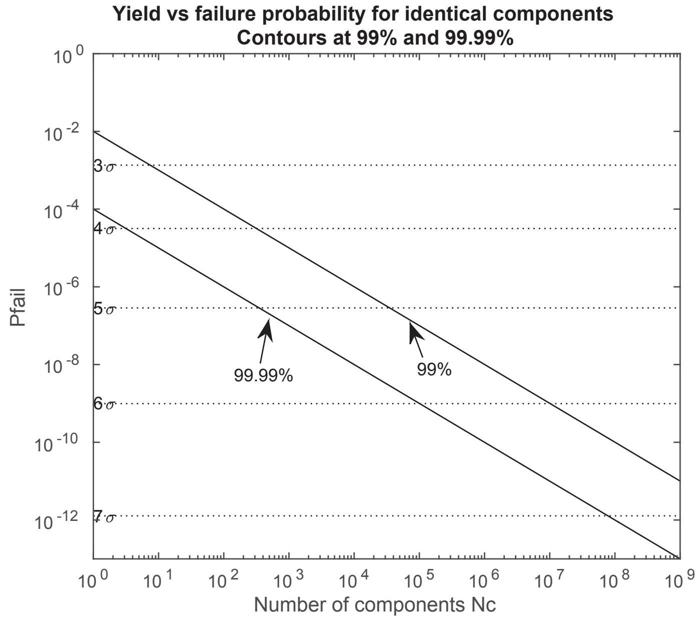


Figure 3.4: Yield contours for equation (1) of Paper VI.

3.6.3 Supplementary material

To supplement the paper and later discussion yield contours for equation (1) of Paper VI is plotted in Figure 3.4. Horizontal stapled lines are drawn for sigma values from the normal distribution, corresponding to the failure rate P_{fail}. We see that a fixed 6σ failure rate for 10 k components will result in a yield of 99.99 %.

Chapter 4

Discussion

4.1 Exploiting INWE

The use of minimum-width and minimum-length sized transistor arrays, rather than the conventional scaling of width may have applications in modern technologies, also in superthreshold. In a recent paper [78] it is reported on performance and energy gains of minimum sized transistor arrays based on a 16 nm predictive technology model operated at $V_{DD} = 0.8\,\mathrm{V}$. For instance, a two-input NAND shows a performance (speed) gain of 61%, and energy gains of 57% at a cost of a 22% area increase compared to conventional scaling. These are substantial gains indeed, still one may raise questions about how trustworthy these simulation results are. Specifically, while NWE is modeled in both BSIM3v3 and BSIM4, the inverse narrow width effect (INWE) is not modeled by either [79], although BSIM4 has a length-dependent reverse narrow width term for short channels. It is however sometimes seen that foundries use the NWE parameters to model effects that could include INWE, utilizing negative coefficients and/or wrapping equations around the central BSIM NWE model parameters (K3, K3B and DVT0W-DVT2W [25]). On the other hand, other foundry models do not describe the required parameters to include this effect at all, in effect setting these coefficients to zero.

The measurement results presented in Section 3.1.4, Figure 3.3 indicate that there could potentially be substantial gains at low V_{DD} if minimum-sized transistor arrays are employed, and INWE combined with reduced gate capacitance, could at least partially explain why. However these results do not match too well with simulated results. In order to fully utilize effects such as INWE in the weak or moderate inversion regions, improved models are required to properly quantify the effects. Needless to say, a combined speed and energy improvement of minimum-sized transistor arrays would be welcome news for many applications. However, these results are based only on a single sample, due to great difficulty in extracting the data, and should therefore not be considered to be conclusive.

Although the topic of minimum-sized arrays of transistors was set aside, partly due to poor transistor models in a circuit design context, others have developed such a concept further.The impact of INWE on subthreshold device sizing is investigated in [80]. In [81] the use of INWE aware sizing to construct a low voltage standard cell library is investigated. Applying this to a baseband processor they indicate gains of up to 20% less delay, up to 34% less power consumption and up to 47% less area. In [82] it is advocated to use arrays of optimally sized transistors, to improve delay, power and reliability, through W/L sizing allowing adjustment of V_{th} and $\sigma(V_t)$of individual unit transistors.

In subthreshold, the sum of iid. lognormally distributed currents may in some respects show superior statistical qualities compared to the single current of an iso-area transistor, even without a baseline shift of V_{th} by effects such as SCE, NWE and INWE, as illustrated in the numerical experiment of Paper I. How the statistical properties of the sum of iid. subthreshold currents behave is however strongly dictated by the magnitude of the ratio $\frac{\sigma(V_t)}{nU_t}$, i.e. for a subthreshold current the ratio $\frac{\sigma(V_t)}{nU_t}$ would correspond to σ as appearing in Fig 5. in Paper VI, which shows the right tail of a lognormal sum. Noteworthy, skewing effects on the sum of subthreshold iid. currents increase with decreasing temperature.

4.2 Multi-objective optimization of ULV circuits

In the context of the particular MOOP of Paper II, the objectives area and power were fairly closely related, particularily for areas between 10 µm^2 to 40 µm^{2}for the 10T and 12T gates. This should perhaps come as no surprise when noting that area is related to the capacitance, and capacitance is related to switching power. However, for small areas and for the 22T gate, the relationship with power was not quite that clear. In later analysis the area was also found to be related to the static leakage and switching energy, providing a near linear relationship in the 10 µm^2 to 40 µm^{2}range. For propagation delay the area relation was slightly non-linear but still monotonous. As a lesson learned, in the context of this MOOP, many aspects could be co-optimized by simply calculating the circuit area, or perhaps better in some respects, the total capacitance. The main conflicting objective for the area objective , was our objective for optimizing 'robustness' (reliability) – the RMSE of the output voltage. As can be seen in the Fig. 2 of paper II, the area objective has a very large effect on the RMSE for very small areas, with a decreasing influence when the area grows.

In paper II where 10T, 12T and 22T minority-3 gates are compared it may be of interest to note that the lower bounds on RMSE, for a given area, seem to covariate with the number of transistors in the output function. For the metrics for active power and RMSE, the relationship seems weak until RMSE approaches a limit where the power consumption rises fairly sharply.

As a final note on the circuits investigated in Paper II and III, for a fully comprehensive comparison of minority-3 and C-element circuits in the subthreshold domain, it is probably of interest to also include the topology of the van Berkel implementation [83]. Since the van Berkel implementation utilizes 3 transistors in series it's V_{min} is likely higher than that of the Sutherland implementation. Properties such as area, minimum energy and operating speed may however compete with the Sutherland implementation at moderate supply voltages.

As seen in Papers II, III, and IV, multi-objective optimization (MOO) can be used to approximate the Pareto Front, which in the context of the optimization problem represents all resource-efficient tradeoff alternatives for a circuit. Thus MOO represents a powerful method for exploration of the potential performance space of a circuit. Given a specific optimization problem, MOO can be used to compare circuit topologies and find the optimal implementation for a specific desired performance range. Visualization or metrics from a Pareto Front can quickly provide performance limits of a specific topology. For applications where specific performance requirements are available, selection of the best suited tradeoff alternative(s) is fairly easy. For applications where requirements are not too specific, the Pareto front can be used in a further cost-benefit analysis to suggest optimal allocation of resources in a larger context.

Thus, when the combined effects of subthreshold phenomena such as SCE, NWE and INWE, and RDF, may otherwise cause design difficulties due to the complex design space, MOO has, in principle, the potential to bypass such problems. Several challenges are however involved for the successful application of any MOO algorithm in the context of circuit design. Formulation of the optimization problem; the objective functions and constraints for the search space play a critical role for achieving quality results, within a reasonable computational time.

Another obvious concern for the application of multi-objective optimization is the evaluation time of the objective functions. For circuit optimization macro-modeling can reduce computational effort greatly, however this can not offer the same detail as circuit simulation. For large populations and long simulations parallelization can however be used to reduce the time spent in function evaluation. To incorporate variability in simulations one could consider to introduce corner simulations or Monte Carlo simulations, although both will increase the time spent in simulation. Corner simulations for mismatch typically fix all devices with the same parameter deviation, and may therefore likely yield unrealistic results. Monte Carlo on the other hand will introduce randomness to function evaluation which can cause convergence difficulties for the MOO algorithm. To reduce the impact of random errors one could consider to greatly increase the number of simulations. Another intermediate solution is to use the same seed or sampling as in Paper II and III, so any error is kept constant for each function evaluation, although one must keep in mind this can introduce a certain bias. A challenge related to using Monte-Carlo simulation in conjunction with MOO is thus how to sample the distribution effectively. Several approaches exist, with latin hypercube sampling and orthogonal sampling among the more common methods. Another way could be to tailor corners to the specific application, such as in [84]. The problem of sampling when optimizing a circuit is made more difficult by the influence of W/L scaling on the contributing distributions, and the lognormal distribution of currents in subthreshold. Circuit simulators also often provide DC sensitivity analysis. This can be used to a certain extent, however since it evaluates sensitivity in a fixed operating point, it could lead to an error when estimating the variance of aggregated sources of variation. When it is possible and practical, a solution may also be to analytically determine the variance, thus eliminating the need for Monte Carlo simulations.

Statistical simulations of standard cells could perhaps also be avoided during the optimization stage if all pull-up and pull-down network can be constrained to achieve equal or comparable worst-case conditions in their current distribution. Final evaluation of reliability can then be done in a second step. A constraint for the relative area of single PMOS and single NMOS that to a large degree would equalize the variance of their subthreshold I_{on} current can be found in equation (29) in Paper VI. Since the NMOS is typically smaller a cost-benefit analysis could also set a slightly larger area for the NMOS, improving the distribution further. For more complex gates further area constraints for equalizing the variability of multiple series PMOS and series NMOS would however need to be developed. The single constraint would remove one dimension from the search, and similar relationships for the area of series transistors could provide even greater benefit for more complex gates.

Compared to the approach with MATLAB's *'gamultiobj'* in Papers II and III, a different algorithmic approach was made for the work in [8]. There the Pareto fronts were found by a multi-dimensional bisection search. Initially the search space was divided in a coarse multidimensional grid and each grid point was evaluated and used to find a coarse approximation to the Pareto front. Next the search grid was refined enhancing the resolution for each parameter by a factor 2 in each iteration, however only neighbors in the current Pareto set were investigated in each new iteration. This resulted in a gradual refinement of the Pareto front. For problems that involve several minima this may of course miss solutions, depending on the starting grid. However given some a-priori knowledge of the circuit and kinks related to SCL and NWE effects, this can largely be avoided during the setup of the initial grid. Simulation speed was also improved by using DC approximations and running all simulations via RAM disk, offering an over 10× speed improvement in our environment. For problems with 3 objectives and up to 4 parameters we found this approach preferable to the use of *'gamultiobj'*, however there are still obvious problems with the starting grid and number of neighbors to search, when scaling the number of parameters.

The algorithmic approach in [8] has some similarities to the GAIO algorithm employed in [42]. There, after an initial search in a grid, the search space is iteratively divided in smaller and smaller subspaces, so only subspaces that contained Pareto points is searched. This approach seems effective when there is an expected relationship between parameters, and will have some obvious advantages when the number of solutions are large.

4.3 ULV SRAM

In the context of low power design, ULV memory is an interesting topic with particular challenges. For larger SRAM arrays, such as in a large cache, power consumption is often dominated by leakage due to a low activity factor for each bitcell. Thus it can be of great value to reduce leakage by operating the SRAM bitcell array at a reduced supply voltage. As a counterpoint, the leakage resulting from larger SRAM arrays will push the minimum energy operating point to a higher supply voltage, perhaps above the threshold voltage. Smaller, more active SRAM modules may however display minimum energy operation at voltages well below the threshold voltage of the process. This was the case with the 64×32 block presented in Paper V, where the minimum energy occurred at $V_{DD} = 0.3\,V$, and the threshold voltage was approximately 450 mV. Mainstream applications for an effective use of subthreshold SRAMs are then perhaps limited to smaller SRAM modules for use as FIFOs, L1 cache, register files and scratchpad memories. A strategy to remedy the situation for larger SRAMs is to design peripheral circuits either with a higher voltage supply or using low-V_{th} devices. This can improve access times for larger arrays by a significant amount, thus reducing the leakage energy. For cost and overhead reasons it may however be of interest to simply operate SRAM blocks at the same V_{DD} and/or frequency as the rest of the circuit. For special applications, such as circuits powered by a very limited power source, i.e. in an energy harvesting application, it may also be of benefit for the circuit to operate, even with an extremely low power supply voltage.

The first challenge that must be addressed when implementing an ULV SRAM cell is the decrease in noise margins. This is very prominent for the traditional 6T SRAM cell as there is a conflict when optimizing the cell for the noise margins during read and write access, $\mathrm{SNM}_{\mathrm{read}}$ and $\mathrm{SNM}_{\mathrm{write}}$. Similar to several other ULV SRAM solutions [47, 50, 51, 65, 66, 68, 69, 85], the 9T cell of Paper V solves this conflict by adding extra transistors for the read access. Thus the write access and read access transistors can be optimized independently. Another way of improving ULV operation is to break the feedback-loop during access [47]. Recently, in a comparative study including 4 SRAM cells the 9T cell of Paper IV was found to yield the best hold SNM [86].

Separating it from [87] the 9T cell of Paper V uses a virtual supply for the retention voltage, multiple threshold voltages and PMOS write access transistors as these showed improved characteristics with respect to variability in the target 65 nm LP process. The use of multiple threshold voltages in a very similar 9T topology was further investigated in [88] and triple-V_{th} versions of the 7T [89], 8T [49] and 9T [87] topologies are compared with improvements in other topologies in [52, 90]. In particular the layout of [88] is likely a major improvement to the layout described in Paper V, as the thincell layout will show superior qualities with respect to bitline capacitance thus lowering E_{min} significantly. On the other hand the thincell layout may have some difficulties in accommodating a virtual retention supply for the row, as was used in the cell of Paper V. In the context of [52] triple-V_{th} versions of the 8T and 9T topologies are the best with regards to minimum supply voltage and data stability, with the 9T shows somewhat less leakage. The 9T implementation incurs an area and thus a performance penalty over the 8T implementation. However, a major difference from the 8T cell, which perhaps is not highlighted in that comparison, is that also the 9T can benefit from splitting bitlines into separate read and write bitlines. This reduces bitline capacitance and thus it will enhance access times and lower dynamic power consumption. Additionally the 9T cell provides a differential output signal, circumventing the need for replica bitlines or other techniques to predict leakage when evaluating the read signal, likely this provides additional dynamic read stability to the 9T cell when compared to the 8T.

In Table 4.1 the manufactured SRAM block of of Paper V is compared to seven contemporary manufactured subthreshold SRAMs. For the minimum supply voltage metric V_{min} our SRAM is only the fifth best, although several of our best samples operated down to 230 mV. Notably, all the SRAM cells manufactured in a 130 nm technology were able to reach a lower V_{min} than those in a 65 nm technology. The reason for this is probably a combination of well controlled Vth and cell area, as area can be traded for enhanced stability. Body biasing is also used in [69, 91] to reduce V_{min} significantly. For the area metric our cell is the second smallest, between the publications where cell layout area was reported. In the context of Table 4.1 the 9T SRAM of Paper V was second best for the average minimum energy per operation metric, and it had the lowest minimum energy supply voltage, still with a reasonable average speed of 761 kHz. For operating speed our cell was mid-range at the minimum energy per operation operating point. However when comparing at $V_{DD} = 400$ mV our worst sample operated at 2 MHz, with average speed at 2.9 MHz.

The leakage per bit at 300 mV was 17.8 pW per bit, placing the 9T SRAM of Paper V in the middle. The effective leakage per bit depends not only on the cell topology, but also on the technology and on peripheral circuits. For larger SRAM arrays typically only a fractional part of the leakage derives from the peripheral circuit. However based on simulations of the SRAM module of Paper V a very substantial 72 % of the leakage can be attributed to the decoder. If the number of columns in the SRAM were increased, one could expect a significant reduction in the peripheral circuits leakage. Since average leakage per cell for the 64 × 32 block was 17.8 pW at 0.3 V, thus one could expect a 64 × 64 module to exhibit leakage/bit closer to 12 pW per bit. Although extending the number of columns could incur a speed reduction due to increased wordline capacitances, the throughput would normally increase.

At the time of making, the SRAM array was sized as 64 × 32. The rather limited size is primarily due to size constraints of the die. Although SRAMs with a fairly high number of rows have been demonstrated, such as in [66, 96] which achieve 1 k cells per bitline, here the number of rows was limited intentionally to safely avoid read access errors due to bitline leakage. In typical cases the worst-case maximum leakage from all non-accessed bitcells on the column can be considered a sum of iid. lognormally distributed currents and we can apply techniques such as that of Appendix A in Paper VI to find a worst-case value for the acceptable leakage given a reasonable yield. While the leakage on its own must be small enough for the duration of the bitline integration to not affect operation of the sense amplifier, also the difference between the accessed bitcell read current and the leakage must be able to produce some minimum voltage as determined by the requirements of the sense amplifier. Interestingly in [97] it is shown that for operation in the ULV domain energy efficiency can be better when the number of rows is kept low, when sizing the SRAM array, however they argue that this is the case particularily when the SRAM array is large.

Sense amplifier design in the subthreshold domain is perhaps the most challenging task of ULV SRAM design as RDF causes severe mismatch in the input sensing transistors, and speed and energy efficiency will suffer when upsizing the transistors to compensate. In the context of Paper VI, no special circuit techniques other than W/L sizing were used to mitigate subthreshold variability. Although multiple sense amplifier topologies were investigated, all struggled to meet the design requirements of operation from -20°C to +85°C with $V_{\mathrm{min}} = 200\,\mathrm{mV}$, with a sensing time allowing for an access time of 1 µs at $V_{DD} = 300\,\mathrm{mV}$. While the selected topology was the only one that came close to these requirements within an appreciable yield, the implementation displays increased detection delays when scaling to higher supply voltages. This is due to the read access precharge to ground and the need for an increased bitline integration time to provide sufficient current. For this application it was however acceptable as ULV operation was the main target for investigation.

One way of handling mismatch in the sense amplifiers is to calibrate each sense amplifier via multiple references, as seen in [98]. However in subthreshold a large temperature range may cause major grievance unless it is tackled. Body biasing [17, 99, 100, 101] can improve leakage and mitigate global variation in subthreshold currents due to overall V_{th} bias as well as temperature effects over a significant range [23]. Thus it could be a very sensible technique to apply to sense amplifiers, to boost their effectiveness when operated in the subthreshold domain. Body biasing for sense amplifiers was for instance used for calibration in [102] to achieve 2× less input offset after calibration. Adaptive body biasing was also used in the SRAM of [69, 91] to achieve a supply voltage as low as 193 mV for a 6T SRAM cell.

Table 4.1: Comparison of subthreshold SRAM modules.

Source	This work	[85]	[68, 92]	[65, 93]	[47]	[51, 94]	[66, 95]	[69, 91]	[96]
Technology	65nm	65nm	65nm	65nm	90nm	130nm	130nm	130nm	130nm
Cell type	9T	9T	8T	10T	10T	8T	10T	6T	A8T-RWD[5]
Block size	64×32	64×72	64×128	256×128	128×256	512×64	1024×256	16×16	512×32
Cell area [μm^2]	2.83	1.77			3.5[1]	6.36	7.50	4.79	
Eff. cell area [μm^2]	4.61	3.04				9.34		13.96	9.46
E_{min} Voltage [mV]	321[3]	500	400	400[2]		400[2]	400[2]	340	375[1]
E_{min} Frequency [kHz]	761[3](1016[4])	5000[1]	500[1]	475		6700	1900	400[1]	1900[1]
E_{min} [fJ/access/bit]	17.8[3]	62.5	82.5[1]	14.1[1]		25.2	32.9[1]	138	119[1]
Leakage [pW/bit] @ 300 mV	17.8[3]	40.7[1]	12.2[1]	11.8[1]	25	64.1[1]	7.3[1]	54.9	11.5
V_{min} [mV]	273 (230[4])	275	250	380		230	200	193	300 (250[4])

[1] Approximation based on diagram and/or calculation.

[2] Operating point chosen due to values available in publication, but not reported as true energy minimum.

[3] Average value.

[4] Value for best sample.

[5] Average-8T Read Write Decoupled

[†] This table is an expanded version, based on a table presented in [72]

Considering multi-V_{DD} circuits, another and more direct way of improving sense amplifiers is to operate these at a supply voltage allowing super-threshold operation. Although this would increase leakage in the sense amplifiers, the sensing time would be expected to drop drastically and could thus improve overall energy efficiency.

Although adaptive body biasing easily can compensate global between-die variation, local within-die variation is still severe. The analysis of measured bitcell retention failure in Paper V adds weight to the hypothesis that mismatch in the subthreshold domain is dominated by RDF, and thus relatively less impacted by strain and optical effects due to surrounding geometries.

4.4 Subthreshold energy modeling including RDF

The impact of RDF induced variations is important to investigate in order to achieve reasonable yield in subthreshold circuits. Monte-Carlo simulation can be helpful to investigate single transistors, or smaller circuits, but this approach often becomes unmanageable for larger circuits with high yield targets. In some cases an analytical approach can however replace Monte-Carlo simulation. Specific for subthreshold circuits individual gate delays follow strongly skewed distributions. Summing a fixed number of skewed distributions with unequal means and variances to find the distribution of total delay, is unfortunately rather difficult to treat analytically. Paper VI instead looks at the sum of lognormally iid. distributed delays. For circuit design this is perhaps rather a special case that is only valid if all gate delays are optimized to have the same delay distribution. Analysis of circuits where multiple distributions are known to be lognormal but have different means and variances could perhaps benefit from the approach in [103]. Nevertheless, several principles are well highlighted by this study. Perhaps most importantly, the sum of relatively few iid. lognormal distributions may show a strong response in the right-tail of the total, given that the variance of individual delays are large. I.e., for long paths and large variances the worst-case delay may not always be adequately ameliorated by 'averaging' effects, as is sometimes suggested.

The analysis derives worst-case expressions for: on-current (I_{on}), on/off ratios (I_{on}/I_{off}), propagation delay of simple gates (t_p), clock period for series-connection of identical gates (t_{clk}), and energy per cycle (E_{op}). Analytical expressions for the optimal V_{DD} to minimize energy per operations were also derived. A similar analysis can be found in [76, 77], although these do not include variability. Minimum energy-operation when simultaneously applying DVS and ABB techniques was also treated in [104]. While [77] states that minimum energy per operation theoretically is at minimum gate area. Paper VI finds that given the influence of RDF this may not always be the case, although the absence of RDF yields the same result.

Although Paper VI successfully incorporates RDF in a subthreshold minimum energy model there are several effects that may be interesting to add if one were to expand the analysis. In the analysis several threshold voltage shifts due to geometric variation of the gate are not taken into account. As mentioned already, short and narrow channel effects may shift the mean threshold voltage. Additionally DIBL may have some influence as it depends both on the geometry of the device as well as the drain voltage. Since the DIBL V_{th} shift is proportional to V_{DD} [7], it is however a weaker effect when comparing to superthreshold operation. The assumption that the threshold voltage follows the normal distribution may be challenged, perhaps particularily if effects such as NBTI modulate the threshold voltage during operation. Experimental results in [32] seem to indicate that Vth shifts induced by NBTI follow a Skellam distribution after stressing devices.

A notable aspect of the contour plots of Fig. 1 in Paper VI is that the optimal reliable sizing for minimum energy for a subthreshold circuit, does not coincide with an optimal sizing for higher supply voltages. Thus a project targeting a DVS scheme for minimum energy will be forced to balance the choice of how to size standard cells, based on the expected activity of the circuit.

The analysis of Paper VI is helpful to understand how circuit complexity and RDF-induced variability affects the overall system performance when scaling up the number of transistors in a sequential subthreshold circuit. One application for which the analysis may be particularily suited lies in feasibility studies, where one wants early estimates of performance given a specific circuit complexity and a target yield. Another application for the analysis would be in standard cell optimization, e.g. to use the results to guide optimization such as target V_{DD} and gate area, or also to use some of the developed expressions to constrain and speed up the optimization process.

Chapter 5

Conclusion

5.1 Conclusion

The paper contributions of this thesis relate closely to the field of ultra low voltage or subthreshold circuit design, and seek to advance the field both through an enhanced understanding, and improved methods for the purpose of designing reliable and efficient subthreshold digital logic and memory.

Potential exploits of the inverse narrow-width effect (INWE) are highlighted in Paper I, where it is shown that arrays of minimum-sized transistors in certain cases can provide enhanced performance over wide transistors of an equivalent area. In simulation iso-area inverters in a ring oscillator achieved reduced delays (35 %–40 %) and reduced PDP (40 %–43 %). Measurement results point to the presence of such an effect although it may be weaker, thus improved models may be necessary for reliably optimizing designs to take advantage of this effect.

Multi-objective optimization (MOO) algorithms, such as employed in Papers II, III, and IV represents a powerful method for exploration of resource efficient tradeoff alternatives when implementing a circuit. Given a specific optimization problem, MOO can be used to compare circuit topologies and find the optimal implementation for a specific desired performance range. After the Pareto Front has been found, the Pareto optimal solution that best suits the application, can easily be selected for further use. Visualization, and metrics of Pareto Fronts can quickly provide performance limits of a specific implementation. MOO also has, in principle, the potential to bypass many problems related to subthreshold sizing, when the combined effects of subthreshold phenomena such as SCE, NWE, INWE, and RDF, may otherwise cause design difficulties.

The 9T SRAM cell of Paper IV was carefully optimized for resource efficient operation at a supply voltage of 300 mV. The cell uses techniques such as read-write decoupling, multiple threshold voltages, and a virtual retention supply voltage to achieve improved stability, leakage and access speed. In Paper V the 9T cell is used within a 2 kb SRAM module to demonstrate the feasibility of a design based on this cell. From 38 tested samples reliable measurements from the 9T SRAM module are provided, with V_{min} between 230 mV and 271 mV, minimum energy per operation at an average supply voltage of 321 mV, with averages of 0.57 pJ energy per access and 730 kHz for the operating frequency. These results are evidence of reliable operation, and the performance metrics are comparable to state-of-the-art published results for ULV-SRAM.

For digital circuit design in the subthreshold domain the analysis of Paper VI shows a possible path for how combined effects of RDF variability and complexity can be estimated very early in the design phase, even before W/L sizing of library gates. Through slightly simplified models the analysis allows early estimates for the minimum energy operating voltage, expected energy per gate per cycle, as well as expected operating frequency and suggested gate areas for efficient RDF mitigation. Through contour plots available trade-offs can be investigated. Such an analysis could perhaps be ideal in tasks such as performing feasibility studies to evaluate the feasibility of a specific circuit project in the subthreshold domain, or during technology selection to compare merits of different technologies.

5.2 Recommendations for Further Work

Improvements and use of the work in Paper I, in the context of circuits and systems design, will perhaps rely on foundries providing accurate models for ULV phenomena such as INWE. Since subthreshold circuit design is still considered a niche the wait may however be long. Of course for projects targeting this, statistical measurements on test devices could resolve that situation. Utilization of benefits should be fairly straightforward, applying the MST techniques to digital standard cells that require higher operating speeds and can tolerate an increase in leakage, but potentially granting savings on energy per operation.

For MOO, computational effort can be greatly reduced by finding good constraints. Specifically it would be interesting to investigate the usefulness of applying some of the results from the analysis of Paper VI to provide constraints, e.g. transistor area, for the multi-objective optimization of standard cells for a subthreshold standard cell library. It would also be interesting with further investigation into which related objectives should be preferred or combined when considering reducing the number of objective functions. After the application of a MOOmethod to a circuit intended for use in a standard cell library, when multiple resource efficient tradeoffs are available, it seems interesting to pursue a path where electronic design automation (EDA) tools could pick and choose from the available tradeoffs, in order to optimize each delay path individually. This may however require automated layouts, and likely several modifications to EDA software algorithms. Unfortunately state-of-the-art EDA software is almost entirely proprietary. A more moderate step towards this would perhaps be to realize and characterize a moderate subset of the Pareto optimal solutions to provide evidence of the efficacy of such an approach, or perhaps to do postprocessing of the gate level netlist to refine the implementation. A challenge related to using Monte-Carlo simulation in conjunction with MOO, is how to sample the distribution effectively. The problem is exacerbated by the influence of W/L scaling on the distribution. Little work has been published so far, with the aim of efficiently sampling or describing lognormal performance distributions as seen in subthreshold circuits.

To improve the 9T SRAM a thincell implementation of the cell layout may allow a strong improvement to the access energy. Since one of the main differences to the 8T SRAM cell is that the 9T cell has a differential read current, it would be prudent to first show that the sense amplifier can utilize this to improve performance, granting an overall benefit that outweighs the effects of increased area. For applications requiring extremely low V_{min} a body biasing scheme should be implemented, in the bit cell array as well as in peripheral circuits. Device properties in other fabrication technologies may be different, thus NMOS access transistors may be preferred.

Further work on the analysis of Paper VI would perhaps start at invoking a circuit simulator to test the accuracy of the model. Incorporating simulation also allows for modeling of prominent subthreshold device effects such as SCE, NWE, INWE and DIBL, although optimal individual gate W/L sizing and/or body bias tuning might be a complicating factor. Plots of worst-case delay vs. yield, and energy and operating voltage vs. yield, may be an interesting addition for visualization. An extension of the analysis could also cover the effects of multiple series transistors on the delay distribution, which can be expected to worsen the performance somewhat.

Publications

Backmatter

Bibliography

[1] C. Hu, "Future CMOS scaling and reliability," *Proceedings of the IEEE*, vol. 81, pp. 682–689, May 1993.

[2] E. A. Vittoz, *Low Power CMOS Circuits*, ch. 16. Weak Inversion for Ultimate Low-Power Logic, pp. 16–1–16–18. CRC Press, Taylor & Francis Gorup, LLC, 2006.

[3] R. Troutman, "Vlsi limitations from drain-induced barrier lowering," *Solid-State Circuits, IEEE Journal of*, vol. 14, pp. 383–391, Apr 1979.

[4] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages," *Solid-State Circuits, IEEE Journal of*, vol. 28, pp. 10 –17, jan 1993.

[5] ITRS, "The international technology roadmap for semiconductors : 2013 edition : Process integration, devices, and structures," tech. rep., ITRS, 2013.

[6] H. Ammari, N. Gomes, W. Grosky, M. Jacques, B. Maxim, and D. Yoon, *Wireless Sensor Networks: Current status and future trends.*, ch. I: Review of applications of wireless sensor networks, pp. 3–32. CRC Press, 2012.

[7] B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*. Prentice Hall, 5 ed., 2000.

[8] M. S. O. Haugland, "Multiobjective optimization of an ultra low voltage/low power standard cell library for digital logic synthesis," Master's thesis, University of Oslo, 2012.

[9] C. Piguet et al., *Low Power CMOS circuits*. CRC Press, Taylor & Francis Group, LLC, 2006.

[10] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *Solid-State Circuits, IEEE Journal of*, vol. 27, pp. 473–484, Apr 1992.

[11] Y. Tsividis, "Eric vittoz and the strong impact of weak inversion circuits," *Solid-State Circuits Society Newsletter, IEEE*, vol. 13, pp. 56–58, Summer 2008.

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

[13] H. Pao and C. Sah, "Effects of diffusion current on characteristics of metal-oxide (insulator)-semiconductor transistors," *Solid-State Electronics*, vol. 9, no. 10, pp. 927 – 937, 1966.

[14] E. A. Vittoz, *Micropower Techniques*, pp. 53–96. Prentice-Hall, Inc., 1994.

[15] C. Enz, F. Krummenacher, and E. Vittoz., "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing journal on Low-Voltage and Low-Power Design*, vol. 8, no. Special Issue July, pp. 83–114, 1995.

[16] A. Wang, B. H. Calhoun, and A. P. Chandrakasan, *Sub-threshold design for ultra low-power systems*. Springer, 2006.

[17] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," *Solid-State Circuits, IEEE Journal of*, vol. 37, pp. 1545–1554, Nov 2002.

[18] A. Wang and A. Chandrakasan, "A 180-mv subthreshold FFT processor using a minimum energy design methodology," *Solid-State Circuits, IEEE Journal of*, vol. 40, pp. 310–319, Jan 2005.

[19] D. Blaauw and B. Zhai, "Energy efficient design for subthreshold supply voltage operation," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pp. 4 pp.–32, May 2006.

[20] O. C. Akgun, J. Rodrigues, and J. Sparsø, "Energy-minimum sub-threshold self-timed circuits using current sensing completion detection," *IET Computers & Digital Techniques*, vol. 5, no. 4, pp. 342–353, 2011.

[21] G. Schrom, C. Pichler, T. Simlinger, and S. Selberherr, "On the lower bounds of CMOS supply voltage," *Solid-State Electronics*, vol. 39, no. 4, pp. 425 – 430, 1996.

[22] E. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down," *IBM Journal of Research and Development*, pp. 169–180, Mar./May 2002.

[23] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Trans. Circuits Syst. I*, vol. 59, no. 1, pp. 3–29, 2012.

[24] M. Alioto, "Understanding DC behavior of subthreshold CMOS logic through closed-form analysis," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 7, pp. 1597–1607, 2010.

[25] M. V. Dunga and et al., *BSIM4.6.0 MOSFET Model - User's Manual*. University of California, Berkeley, 2006.

[26] K. M. Cao, W. Liu, X. Jin, K. Vashanth, K. Green, J. Krick, T. Vrotsos, and C. Hu, "Modeling of pocket implanted mosfets for anomalous analog behavior," in *Electron Devices Meeting, 1999. IEDM '99. Technical Digest. International*, pp. 171–174, Dec 1999.

[27] T.-H. Kim, J. Keane, H. Eom, and C. H. Kim, "Utilizing reverse short-channel effect for optimal subthreshold circuit design," *IEEE Trans. VLSI Syst.*, vol. 15, no. 7, pp. 821–829, 2007.

[28] A. Asenov, S. Kaya, and A. Brown, "Intrinsic parameter fluctuations in decananometer mosfets introduced by gate line edge roughness," *Electron Devices, IEEE Transactions on*, vol. 50, pp. 1254–1260, May 2003.

[29] S. A. Campbell, *The science and engineering of microelectronic fabrication*. Oxford University Press, 2001.

[30] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *Electron Devices, IEEE Transactions on*, vol. 41, pp. 2216–2221, Nov 1994.

[31] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub-0.1 um mosfet's: A 3-d "atomistic" simulation study," *Electron Devices, IEEE Transactions on*, vol. 45, pp. 2505–2513, Dec 1998.

[32] V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes, and L. Camus, "NBTI degradation: From transistor to SRAM arrays," in *Reliability Physics Symposium, 2008. IRPS 2008. IEEE International*, pp. 289–300, April 2008.

[33] M. Pelgrom and M. Vertregt, "Component matching: best practices and fundamental limits.." IDESA Seminar DVDs, www.idesa-training.org.

[34] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *IEEE Transactions on Electron Devices*, vol. 45, pp. 1960–1971, Sept. 1998.

[35] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. Mcintyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Williams, and K. Zawadzki, "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 *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pp. 247–250, Dec 2007.

[36] M. Blesken, U. Rückert, D. Steenken, K. Witting, and M. Dellnitz, "Multiobjective optimization for transistor sizing of CMOS logic standard cells using set-oriented numerical techniques," in *27th Norchip Conference*, nov. 2009.

[37] O. L. D. Weck, "Multiobjective optimization: History and promise," in *Proc. 3rd China-Japan-Korea Joint Symp. Optimization Structural Mech. Syst. Invited Keynote Paper GL2-2*, 2004.

[38] J. Andersson, "A survey of multiobjective optimization in engineering design," Technical report LiTH-IKP-R-1097, Dept. Mechanical Engineering, Linköping University, 2000.

[39] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Multiobjective evolutionary algorithms: A survey of the state of the art," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 32 – 49, 2011.

[40] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[41] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength pareto evolutionary algorithm," tech. rep., 2001.

[42] M. W. Blesken, *Ein Mehrzieloptimierungsansatz zur Dimensionierung ressourceneffizienter integrierter Schaltungen*. PhD thesis, University of Paderborn, 2012.

[43] M. Fazzolari, R. Alcala, Y. Nojima, H. Ishibuchi, and F. Herrera, "A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions," *Fuzzy Systems, IEEE Transactions on*, vol. 21, pp. 45–65, Feb 2013.

[44] M. Blesken, S. Lütkemeier, and U. Rückert, "Multiobjective optimization for transistor sizing of sub-threshold CMOS logic standard cells," *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 1480 –1483, May 2010.

[45] M. H. Abu-Rahma and M. Anis, *Nanometer Variation-Tolerant SRAM*. Springer New York, 2013.

[46] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE Journal of Solid State Circuits*, vol. 22, no. 5, pp. 748–754, 1987.

[47] I. J. Chang, J.-J. Kim, S. Park, and K. Roy, "A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 650 –658, feb. 2009.

[48] B. Zimmer, S. O. Toh, H. Vo, Y. Lee, O. Thomas, K. Asanovic, and B. Nikolic, "Sram assist techniques for operation in a wide voltage range in 28-nm cmos," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 59, pp. 853–857, Dec 2012.

[49] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable SRAM cell design for the 32 nm node and beyond," in *VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on*, pp. 128–129, June 2005.

[50] L. Chang, Y. Nakamura, R. Montoye, J. Sawada, A. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS," in *VLSI Circuits, 2007 IEEE Symposium on*, pp. 252–253, June 2007.

[51] T.-H. Kim, J. Liu, and C. Kim, "A voltage scalable 0.26 V, 64 kb 8T SRAM with Vmin lowering techniques and deep sleep mode," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 1785–1795, June 2009.

[52] H. Zhu and V. Kursun, “A comprehensive comparison of data stability enhancement techniques with novel nanoscale SRAM cells under parameter fluctuations,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 61, pp. 1473–1484, May 2014.

[53] H. Berge and S. Aunet, “Benefits of decomposing wide CMOS transistors into minimum-size gates,” in *NORCHIP, 2009*, pp. 1 –4, nov. 2009.

[54] K. Ohe, S. Odanaka, K. Moriyama, T. Hori, and G. Fuse, “Narrow-width effects of shallow trench-isolated CMOS with n+ -polysilicon gate,” *IEEE Trans. Electron Devices*, vol. 36, pp. 1110–1116, June 1989.

[55] P. Sallagoity, M. Ada-Hanifi, M. Paoli, and M. Haond, “Analysis of width edge effects in advanced isolation schemes for deep submicron CMOS technologies,” *IEEE Transactions on Electron Devices*, vol. 43, pp. 1900–1906, 1996.

[56] H. K. O. Berge and S. Aunet, “Multi-objective optimization of minority-3 functions for ultra-low voltage supplies,” in *Proc. IEEE Int Circuits and Systems (ISCAS) Symp*, pp. 2313–2316, 2011.

[57] D. Hampel, K. Prost, and N. Scheingberg, “Threshold logic using complementary MOS device,” Aug. 1975.

[58] S. Aunet, “Subthreshold minority-3 gates and inverters used for 32-bit serial and parallel adders implemented in 90 nm CMOS,” in *Proceedings of NORCHIP 2009*, pp. 1 –6, nov. 2009.

[59] B. H. Calhoun, A. Wang, and A. Chandrakasan, “Device sizing for minimum energy operation in subthreshold circuits,” in *Proc. Custom Integrated Circuits Conf the IEEE 2004*, pp. 95–98, 2004.

[60] H. Berge, A. Hasanbegovic, and S. Aunet, “Muller c-elements based on minority-3 functions for ultra low voltage supplies,” in *Design and Diagnostics of Electronic Circuits Systems (DDECS), 2011 IEEE 14th International Symposium on*, pp. 195–200, April 2011.

[61] D. Muller and W. Bartky, “A theory of asynchronous circuits,” in *Proceedings of an International Symposion on the Theory of Witching*, pp. 204–243, Harvard University Press, Apr. 1959.

[62] J. Sparsø and S. Furber, *Principles of asynchronous circuit design - A systems perspective*. Kluwer Academic Publishers, 2001.

[63] K. Fant and S. Brandt, “NULL convention logic™: a complete and consistent logic for asynchronous digital circuit synthesis,” in *Application Specific Systems, Architectures and Processors, 1996. ASAP 96. Proceedings of International Conference on*, pp. 261–273, Aug 1996.

[64] I. E. Sutherland, “Micropipelines,” *Commun. ACM*, vol. 32, pp. 720–738, 1989.

[65] B. Calhoun and A. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE journal of solid-state circuits*, vol. 42, pp. 680–688, 2007.

[66] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A 0.2 v, 480 kb subthreshold SRAM with 1 k cells per bitline for ultra-low-voltage computing," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 518 –529, feb. 2008.

[67] H. K. O. Berge, M. Blesken, S. Aunet, and U. Rückert, "Design of 9T SRAM for dynamic voltage supplies by a multiobjective optimization approach," in *Proc. 17th IEEE Int Electronics, Circuits, and Systems (ICECS) Conf*, pp. 319–322, 2010.

[68] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 3163 –3173, nov. 2009.

[69] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "A variation-tolerant sub-200 mV 6-T subthreshold SRAM," *Solid-State Circuits, IEEE Journal of*, vol. 43, pp. 2338 –2348, oct. 2008.

[70] S. Lütkemeier, T. Jungeblut, H. K. O. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65 nm 32 b subthreshold processor with 9T multi-Vt SRAM and adaptive supply voltage control," *Solid-State Circuits, IEEE Journal of*, vol. PP, pp. 1 –12, Jan 2013.

[71] S. Lütkemeier, T. Jungeblut, M. Porrmann, and U. Rückert, "A 200 mV 32 b subthreshold processor with adaptive supply voltage control," in *Proc. IEEE Int. Solid-State Circuits Conf. Digest of Technical Papers (ISSCC)*, pp. 484–486, 2012.

[72] S. Lütkemeier, *Ressourceneffiziente Digitalschaltungen für den Subschwellbetrieb*. PhD thesis, University of Paderborn, 2013.

[73] M. Seok, G. Chen, S. Hanson, M. Wieckowski, D. Blaauw, and D. Sylvester, "CAS-FEST 2010: Mitigating variability in near-threshold computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, no. 1, pp. 42–49, 2011.

[74] T. Jungeblut, G. Sievers, M. Porrmann, and U. Rückert, "Design space exploration for memory subsystems of VLIW architectures," in *Networking, Architecture and Storage (NAS), 2010 IEEE Fifth International Conference on*, pp. 377 –385, july 2010.

[75] H. Berge and S. Aunet, "Yield-oriented energy and performance model for subthreshold circuits with Vth variations," in *Design and Diagnostics of Electronic Circuits Systems (DDECS), 2013 IEEE 16th International Symposium on*, pp. 193–198, April 2013.

[76] A. Wang, B. Calhoun, and A. P. Chandrakasan, *Sub-threshold Design for Ultra Low-Power Systems (Series on Integrated Circuits and Systems)*, ch. 4. Springer-Verlag New York, Inc., 2006.

[77] B. Calhoun, A. Wang, and A. Chandrakasan, “Modeling and sizing for minimum energy operation in subthreshold circuits,” *Solid-State Circuits, IEEE Journal of*, vol. 40, pp. 1778 – 1786, sept. 2005.

[78] A. Beg, “Designing array-based CMOS logic gates by using a feedback control system,” in *Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on*, pp. 935–939, Oct 2014.

[79] P. T. B. Yew, *Compact Modeling of Deep Submicron CMOS Transistor with Shallow Trench Isolation Mechanical Stress Effect*. PhD thesis, Universiti Sains Malaysia, 2008.

[80] J. Zhou, S. Jayapal, J. Stuyt, J. Huisken, and H. de Groot, “The impact of inverse narrow width effect on sub-threshold device sizing,” in *Proc. 16th Asia and South Pacific Design Automation Conf. (ASP-DAC)*, pp. 267–272, 2011.

[81] J. Zhou, S. Jayapal, B. Busze, L. Huang, and J. Stuyt, “A 40 nm inverse-narrow-width-effect-aware sub-threshold standard cell library,” pp. 441–446, 2011.

[82] V. Beiu, L. Iordaconiu, A. Beg, W. Ibrahim, and F. Kharbash, “Low power and highly reliable gates using arrays of optimally sized transistors,” in *Semiconductor Conference (CAS), 2012 International*, vol. 2, pp. 433–436, Oct 2012.

[83] K. van Berkel, “Beware the isochronic fork,” *Integr. VLSI J.*, vol. 13, pp. 103–128, June 1992.

[84] M. Sengupta, S. Saxena, L. Daldoss, G. Kramer, S. Minehane, and J. Cheng, “Application-specific worst case corners using response surfaces and statistical models,” *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, pp. 1372–1380, Sept 2005.

[85] M.-H. Tu, J.-Y. Lin, M.-C. Tsai, C.-Y. Lu, Y.-J. Lin, M.-H. Wang, H.-S. Huang, K.-D. Lee, W.-C. Shih, S.-J. Jou, and C.-T. Chuang, “A single-ended disturb-free 9T subthreshold SRAM with cross-point data-aware write word-line structure, negative bit-line, and adaptive read operation timing tracing,” *Solid-State Circuits, IEEE Journal of*, vol. 47, pp. 1469–1482, June 2012.

[86] V. Beiu, M. Tache, and F. Kharbash, “Reliability enhanced SRAM bit-cells,” in *Semiconductor Conference (CAS), 2014 International*, pp. 229–232, Oct 2014.

[87] Z. Liu and V. Kursun, “Characterization of a novel nine-transistor SRAM cell,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 16, pp. 488 –492, april 2008.

[88] H. Zhu and V. Kursun, “Symmetrical triple-threshold-voltage nine-transistor SRAM circuit with superior noise immunity and overall electrical quality,” in *SoC Design Conference (ISOCC), 2011 International*, pp. 333–336, Nov 2011.

[89] S. Tawfik and V. Kursun, "Low power and robust 7T dual-Vt SRAM circuit," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pp. 1452–1455, May 2008.

[90] H. Zhu and V. Kursun, "A comprehensive comparison of superior triple-threshold-voltage 7-transistor, 8-transistor, and 9-transistor SRAM cells," in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, pp. 2185–2188, June 2014.

[91] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mV 6T SRAM in 0.13 um CMOS," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 332–606, Feb 2007.

[92] M. Sinangil, N. Verma, and A. Chandrakasan, "A reconfigurable 65nm SRAM achieving voltage scalability from 0.25 - 1.2V and performance scalability from 20kHz - 200MHz," in *Solid-State Circuits Conference, 2008. ESSCIRC 2008. 34th European*, pp. 282–285, Sept 2008.

[93] B. Calhoun and A. Chandrakasan, "A 256kb sub-threshold SRAM in 65nm CMOS," in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pp. 2592–2601, Feb 2006.

[94] T.-H. Kim, J. Liu, and C. Kim, "A voltage scalable 0.26V, 64kb 8T SRAM with Vmin lowering techniques and deep sleep mode," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, pp. 407–410, Sept 2008.

[95] T.-H. Kim, J. Liu, J. Keane, and C. Kim, "A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, pp. 330–606, Feb 2007.

[96] M. Khayatzadeh and Y. Lian, "Average-8T differential-sensing subthreshold SRAM with bit interleaving and 1k bits per bitline," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 22, pp. 971–982, May 2014.

[97] A. Garg and T.-H. Kim, "SRAM array structures for energy efficiency enhancement," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 60, pp. 351–355, June 2013.

[98] S. Cosemans, W. Dehaene, and F. Catthoor, "A 3.6 pJ/access 480 MHz, 128 kb on-chip SRAM with 850 MHz boost mode in 90 nm CMOS with tunable sense amplifiers," *Solid-State Circuits, IEEE Journal of*, vol. 44, pp. 2065–2077, July 2009.

[99] A. Keshavarzi, S. Narendra, S. Borkar, C. Hawkins, K. Roy, and V. De, "Technology scaling behavior of optimum reverse body bias for standby leakage power reduction in CMOS IC's," in *Low Power Electronics and Design, 1999. Proceedings. 1999 International Symposium on*, pp. 252–254, Aug 1999.

[100] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. 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 *Solid-State Circuits Conference, 2002. Digest of Technical Papers. ISSCC. 2002 IEEE International*, vol. 1, pp. 422–478 vol.1, Feb 2002.

[101] S. G. Narendra, *Effect of MOSFET Threshold Voltage Variation on High-Performance Circuits*. PhD thesis, Massachusetts Institute of Technology, 2002.

[102] Y. Sinangil and A. Chandrakasan, "A 128 kbit SRAM with an embedded energy monitoring circuit and sense-amplifier offset compensation using body biasing," *Solid-State Circuits, IEEE Journal of*, vol. 49, pp. 2730–2739, Nov 2014.

[103] C. Lam and T. Le-Ngoc, "Log shifted gamma approximation to lognormal sum distributions," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1, pp. 495–499 Vol. 1, May 2005.

[104] D. Blaauw, S. Martin, T. Mudge, and K. Flautner, "Leakage current reduction in VLSI systems," *Journal of Circuits, Systems and Computers*, vol. 11, no. 6, 2002.