LUND UNIVERSITY

Design Space Exploration of Digital Circuits for Ultra-low Energy Dissipation

Sherazi, Syed Muhammad Yasser

2013

Design Space Exploration of Digital Circuits for Ultra-low Energy Dissipation

S. M. Yasser Sherazi

Lund University

Doctoral Dissertation
Digital ASICs
Lund, January 2014

S. M. Yasser Sherazi
Department of Electrical and Information Technology
Digital ASICs
Lund University
P.O. Box 118, 221 00 Lund, Sweden

Series of licentiate and doctoral dissertations
ISSN 1654-790X; No. 54
ISBN 978-91-7473-725-7

Typeset in Palatino and Helvetica using LaTeX 2ε.
Printed in Sweden by Tryckeriet i E-huset, Lund University, Lund.

Abstract

The ever expanding market of ultra portable electronic products is compelling the designer to invest major efforts in the development of small and low energy electronic devices. The driving force and benefactors of such devices are (but not limited to) e-health system, sensor network applications, security systems, environmental applications, and home automation systems. These markets have launched a massive trend towards ultra low-energy and ultra low-voltage devices. As the technology scales, the dimensions of a transistors have become extremely small, leading to reliability and process variation issues. Above all, with the ability of placing millions of gates in a small area, high current consumption have become one of the key factors in modern high-performance technologies. In portable electronics, the battery life time is a major issue, as most of the time the device is accompanied with an enclosed battery that has to last for long periods without compromise on performance. Furthermore, there are many applications where the battery lifetime sets the lifetime of the device. Therefore, research is needed to identify the techniques and the impact of them on the design operated for ultra low-energy.

The low energy dissipation requirements on a design are achievable by employing various optimization techniques. Voltage scaling is the most effective knob to reduce energy dissipation. For this reason ultra-low energy design usually translates into ultra-low voltage or subthreshold (sub-V_T) domain operation. This work presents an analysis on design space for ultra-low energy dissipation of digital circuits. The circuits are operated in the sub-V_T region with moderate throughput constraints. The drawback of operating circuits in sub-V_T is slow speed performances and reduced reliability. To combat speed degradation due to scaling of the supply voltage, the architectural design

space, needs exploration.

Techniques such as device sizing, body biasing, stacking transistors, dual threshold gates, multi threshold synthesis, pipelining, and loop unfolding, are explored and applied to the designs. The designs are synthesized in a 65 nm CMOS technology with low-power and three threshold options, both as single-V_T and as multi-V_T designs. A sub-V_T energy model is applied to characterize the designs in the sub-V_T domain. Reliability in the sub-V_T domain is analyzed by Monte-Carlo simulations. The minimum reliable operation voltage (ROV) for gates in low power 65 nm CMOS technology is found to be around 250 mV.

The applied energy model for designs to be characterized for sub-V_T domain operation is presented. The energy model encompasses single V_T implementations and multi-V_T implementations. The energy modeling is based on the 65 nm CMOS standard cells provided by the technology vendor. The energy model has been used to evaluate various techniques and constraints for a circuits operated in the sub-V_T domain.

The work describes how the energy dissipation of architectures vary w.r.t. switching activity, μ_e. The effects of pipelining together with supply voltage scaling is analyzed, which shows that they have high benefits with respect to energy dissipation. Various halfband digital (HBD) filter structures are evaluated for minimum energy dissipation in the sub-V_T domain for a throughput constrained system. All architectures, i.e., unfolded and the basic HBD filter, are implemented and simulated using 65 nm Low-Power High-Threshold (HVT) standard cells. The application of a sub-V_T energy model reveals that it is beneficial to use an unfolded implementation to achieve low energy dissipation per sample at EMV, when compared to the energy dissipated by a basic simplified HBD filter implementation.

Various available threshold options are analyzed with the help of filter structures by using 65 nm Low-Leakage High-Threshold (HVT), Standard-Threshold (SVT) and Low-Threshold (LVT) standard cells. Secondly, the design space is increased by utilization of a combination of HVT + SVT and also HVT + LVT cells. The analysis with sub-V_T energy model leads to the conclusion that a suitable design is a synergy between parallelism, and utilization of various threshold options. In this analysis the multi-V_T, implementations did not show a major advantage over single V_T implementations. A decimation filter chain consisting of 4 HBD filters is fabricated and the silicon measurements demonstrate that SVT and different architectural flavors are suitable for a ultra low energy (ULE) implementation. Silicon measurements prove functionality down to a supply at 350 mV, with a maximum clock frequency of 500 kHz, having an energy dissipation of 102 fJ/cycle.

Additionally, an alternative to SRAM macro is presented for sub-V_T opera-

tions. The memory is based on standard-cells and is referred to as SCMs. The energy per memory access as well as the maximum achievable throughput in the sub-VT domain of various SCM architectures are evaluated by means of a gate-level sub-V_T energy characterization model.

Preface

The thesis summarizes the analysis and results achieved as the result of the research work performed at the Department of Electrical and Information Technology, Lund University for Doctoral degree in Circuit Design. The thesis includes material published in the following journal or peer reviewed conference papers:

Journal Articles

- **S. Sherazi**, J. Rodrigues, O. Akgun, H. Sjöland, P. Nilsson, "Ultra low energy design exploration of digital decimation filters in 65 nm dual-V_T CMOS in the sub-V_T domain", *Microprocessors and Microsystems: Embedded Hardware Design (MICPRO), Elsevier*, vol.37/4-5, 2013.
  Contribution Research work has been performed by the first author in the guidance of the remaining authors.
- P. Meinerzhagen, **S. Sherazi**, A. Burg, J. Rodrigues, "Benchmarking of standard-cell based memories in the sub-V_T domain in 65 nm CMOS technology", *Journal of Emerging and Selected Topics in Circuits and Systems,* Vol. 1, No. 2, pp. 173-182, 2011.
  Contribution The research work has been performed jointly among the two the first and second author in the guidance of the remaining authors.
- H. Sjöland, J. B. Anderson, C. Bryant, R. Chandra, O. Edfors, A. Johansson, N. Seyed Mazloum, R. Meraji, P. Nilsson, D. Radjen, J. Rodrigues, **S. Sherazi**, V. Öwall, "A receiver architecture for devices in wireless body area networks", *Journal of Emerging and Selected Topics in Circuits and Systems,* Vol. 2, No. 1, pp. 82-95, 2012.
  Contribution The research work on the digital baseband part of the system is

performed under the supervision of the first author.

Peer reviewed Conference Papers

- **S. Sherazi**, P. Nilsson, H. Sjöland, J. Rodrigues, "A 100-fJ/cycle sub-V_T decimation filter chain in 65 nm CMOS", *IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, 2012-12-09.
  Contribution Research work has been performed by the first author in the guidance of the remaining authors.
  Contribution The research work has been performed jointly among the two the first and second author in the guidance of the last author.
- O. Andersson, **S. Sherazi**, J. Rodrigues, "Impact of switching activity on the energy minimum voltage for 65 nm sub-V_T CMOS", *NORCHIP*, 2011-11-14.
  Contribution The research work has been performed jointly among the two the first and second author in the guidance of the last author.
- **S. Sherazi**, P. Nilsson, O. Akgun, H. Sjöland, J. Rodrigues, "Design exploration of a 65 nm sub-V_T CMOS digital decimation filter chain", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2011-05-16.
  Contribution Research work has been performed by the first author in the guidance of the remaining authors.
- **S. Sherazi**, P. Nilsson, O. Akgun, H. Sjöland, J. Rodrigues, "Ultra low energy vs throughput design exploration of 65 nm sub-V_T CMOS digital filters", *NORCHIP*, 2010-11-15.
  Contribution Research work has been performed by the first author in the guidance of the remaining authors.

Additional articles have been published during the Doctoral studies, however, they are not included in this thesis.

Journal Article

- **S. Sherazi**, S. Asif, E. Backenius, M. Vesterbacka, "Reduction of substrate noise in sub clock frequency range", *IEEE Transactions on Circuits and Systems I: Regular Papers (TCAS-I)*, Vol. 57, No. 6, pp. 1287-1297, 2010.

Peer reviewed Conference Papers

- R. Meraji, **S. Sherazi**, J. B. Anderson, H. Sjöland, V. Öwall, "Analog and digital approaches for an energy efficient low complexity channel decoder", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2013-05-19.
- B. Mohammadi, **S. Sherazi**, J. Rodrigues, "Sizing of dual-V_T gates for sub-V_T circuits", *IEEE Subthreshold Microelectronics*, 2012-10-09.

- P. Meinerzhagen, O. Andersson, B. Mohammadi, **S. Sherazi**, A. Burg, J. Rodrigues, "A 500 fW/bit 14 fJ/bit-access 4kb standard-cell based sub-V_T memory in 65 nm CMOS", *ESSIRC*, 2012-09-17.
- P. Meinerzhagen, O. Andersson, **S. Sherazi**, A. Burg, J. Rodrigues, "Synthesis strategies for sub-V_T systems", *European Conference on Circuit Theory and Design, (ECCTD)*, 2011-08-29.

The research work included in this thesis is supported by the Swedish Foundation for Strategic Research (SSF).

Acknowledgments

First, I would like to express my special thanks and gratitude to my supervisors Prof. Peter Nilsson and Associate Prof. Joachim N. Rodrigues, who gave me the opportunity to perform research in one of the most relevant topics in Digital ASICs today. I am grateful for all the guidance and support that I received from Peter. Also for the good company at various conferences, I specially remember the trip to Rio de Janeiro. I am also in gratitude of Joachim, who taught me about ultra-low voltage, teaching skills, and technical writing. I will specially remember the helicopter tour we took together in Rio and the sushi experience in San Francisco.

Special thanks to Prof. Henrik Sjöland for guiding through the project of Ultra Portable Devices (UPD), funded by Swedish Foundation for Strategic Research (SSF). The requirements of the UPD project gave me the opportunity to explore the Ultra low energy design techniques. The designs created, shall be used for the project. I am also grateful and thankful to Prof. Viktor Öwall for all the constructive guidance and help throughout my PhD studies. Also for the exquisite gatherings at his home.

Second, I would also like to thank my colleges and friends here at the department. Specially Reza, who has been my office mate for the last five years. I would also like to thank Johan, Deepak, and Isael for begin friends and mentors. I would also like to thank Oskar, Babak, and Chenxin, for all the work related collaboration and fun evenings. I also am thankful to Abdulaziz, Anders, Carl, Dejan, Dimitar, Assoc. Prof. Erik L., Hemanth, Liang, Mattias, Michal, Prof. Ove E., Rohit, Rakesh, Taimoor, Waqas, and Xiaodong, for all the constructive discussions and friendly discourse in the department. I would like to thank my supervisors Peter and Joachim for proof reading entire thesis and also extend my gratitude towards my colleagues Oskar, Isael, Chenxin,

Reza, Hemanth, Waqas, Abdulaziz, Dejan, and Xiaodong for proof reading parts of the thesis.

Here I would thank Pascal and Prof. Andy Burg for a very successful collaboration on standard cell based memories. Specially, the close collaboration with Pascal was not only beatifically professionally but also personally, as we developed a good friendship. I am also grateful to my UPD project mates for all the collaboration over these five years.

Special thanks and gratitude to Pia Bruhn for managing all the administrative issues for me and for all the help through the years. Also gratitude towards Stefan, Martin, Erik J., Josef, and Robert for all the technical support.

Although I am not mentioning more names in the fear of missing out anyone, I would like to thank all my colleges within Lund University that made my five years of stay here pleasant.

In the end I would like to thank may family, specially my mother for all the sacrifices, patience, love, support, encouragement, and countless prayers. Thanks to my father, who now is in the havens, for supporting my decision of perusing this career wholeheartedly.

Thank You To All Who Helped Me Along The Journey.

S. M. Yasser Sherazi
Lund, January, 2014.

Contents

List of Figures

List of Tables

1 Introduction

Ultra-low energy circuits have gained enormous importance in the modern era. Specifically the energy dissipation of devices like hearing aids, medical implants, and remote sensors has become an important design parameter. Wearable and implantable wireless sensor networks are the backbone of future e-health system where sensors can be deployed on the body for monitoring and alerting hospitals or in-body for restoring lost internal function or to communicate with a robotic arm or leg. These sensor networks require an energy efficient, relatively high data-rate node that can collect medical data via a sensor and communicate them to base stations. The energy efficiency is most important in determining of the best suited design for an electronic device that is to be used in a wearable or implantable wireless sensor node. In order to achieve these constraints, extensive efforts are needed to design the circuitry in such a way that it consumes minimal energy. Additionally, ultra low energy dissipation is very attractive because it makes the battery last longer, which is important as it is non-trivial to change or charge one in an implant. However, the energy dissipation is bounded by the battery lifetime, i.e., high energy dissipation leads to shorter battery life.

The energy dissipation from the battery can be divided into two main parts, the dynamic and the static energy dissipation. When the circuit is in operation the energy dissipated is considered dynamic energy. When the circuit is in idle or standby mode, the energy depleted from the battery is characterized as static energy. The relation between dynamic energy and the battery voltage V_{DD} leads to the fact that reduction in the battery voltage yields quadratic reduction in energy dissipation.

The low energy dissipation requirements are achievable by employing var-

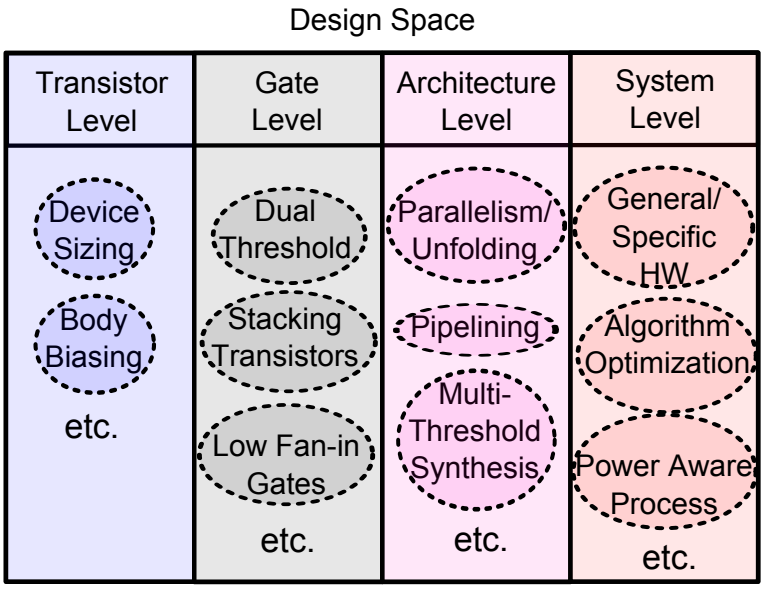


Figure 1.1.: Design space for ultra-low energy implementation

ious techniques. Voltage scaling is the most effective knob to reduce power and energy dissipation, if the timing requirements can still be met. For this reason ultra-low energy design translates into ultra-low voltage (ULV) or sub-threshold (sub-V_T) domain operation. This is one of the most effective knob to play with when reduction in energy is needed. The side effect of supply voltage reduction is an increase in the delays of the gates or the designed circuit.

There is a need of design space exploration for circuits operated in sub-V_T domain to find an optimum solution. The design space includes various level of abstractions ranging from transistor/device level to system level optimization. Figure. 1.1, shows some of the design space knobs that can be explored to find an optimized solution that fits well in the realm of ultra low energy design. Trade-offs at various levels of abstractions may differ from traditional super-threshold (super-V_T) low-power design compared to sub-V_T design. For example, device sizing w.r.t width and body biasing may not be as beneficial in terms of energy efficiency for moderate throughput requirements, when compared to circuits operated at nominal voltage. Circuits with large fan-in or stacking have larger detrimental effects on speeds in sub-V_T domain compared to super-V_T operations. Pipelining and parallelism/unfolding are

beneficial, however, extreme unfolding and pipelining may result in inefficient energy dissipation per output. Exploitation of threshold voltage options may result in energy efficient designs. The thesis includes these options and elaborates on them.

1.1. THESIS CONTRIBUTION

The thesis includes an introduction to power consumption related trends for a CMOS design, presented in Chapter 2. Insight into types of power consumption are presented, together with a brief overview of techniques that are used to reduce power. One technique that reduces all major components of the power consumption is supply voltage V_{DD} scaling. Other architectural improvements yield major advantages once employed together with V_{DD} scaling. The rest of the thesis focuses on V_T scaling and design space analysis that goes hand-in-hand with this technique.

The thesis is mainly divided into four main parts.

PART 1: SUB-V_T DOMAIN BASICS

The first part of the thesis discusses sub-V_T operation basics, this part includes Chapters 3 and 4.

Chapter 3, summarizes the basics of the sub-V_T domain. The current equations encompassing various effects of leakage for example gate induced drain leakage (GIDL) or drain induced barrier lowering (DIBL) leakage are presented. Discussion on fundamental concepts such as the ratio between on-current I_{on} and off-current I_{off} of the transistor is given. Delay degradation due to voltage scaling is shown to be exponential once the supply voltage is scaled below the threshold voltage of the adopted technology. Furthermore, reliability issues due to process variations are discussed.

Chapter 4, includes a proposed energy model used for characterization of the designs operated in sub-V_T domain. The applied model encompasses both single V_T and multi-V_T implementations. The energy modeling is based on the 65 nm CMOS standard cells provided by the technology vendor. The energy model has been used to evaluate various techniques and constraints for circuits operated in the sub-V_T domain.

PART 2: ARCHITECTURAL ANALYSIS FOR SUB-V_T DOMAIN ENERGY DISSIPATION

This part mainly focuses on the architectural analysis for circuits operating in sub-V_T domain. This part includes Chapters 5, 6, and 7.

Chapter 5 describes how the energy dissipation of architectures vary w.r.t. switching activity. Simulation results based on the sub-V_T energy model show that higher switching activity in a given design causes high energy dissipation.

Chapter 6 shows that pipelining together with supply voltage scaling have high benefits with respect to energy dissipation. Simulation results based on the sub-V_T energy model show that designs with long critical paths benefits from reduction in switching activity by the use of pipeline stages. Furthermore, it also help reduce the leakage currents. All of these reductions result in low energy dissipation in the sub-V_T domain.

In Chapter 7 four halfband digital (HBD) filter architectures are evaluated for minimum energy dissipation in the sub-V_T domain for a throughput constrained system. All architectures, i.e., unfolded and the basic HBD filter, are implemented and simulated using 65 nm Low-Leakage High-Threshold (HVT) standard cells. The application of a sub-V_T energy model reveals that it is beneficial to use an unfolded implementation to achieve low energy dissipation per sample at EMV, when compared to the energy dissipated by a basic simplified HBD filter implementation. However, there is a limit to the unfolding factor, where the energy dissipation benefits start to diminish.

PART 3: ANALYSIS ON THRESHOLD OPTIONS WITHIN A TECHNOLOGY FOR SUB-V_T DOMAIN ENERGY DISSIPATION

This part mainly focuses on threshold options within a technology for circuits operating in sub-V_T domain. This includes Chapters 8, and 9.

In Chapter 8, the effect of various available threshold options is examined by the use of HBD filter structures, which are implemented and simulated using 65 nm HVT, Standard-Threshold (SVT) and Low-Threshold (LVT) standard cells. Secondly, the design space is increased by utilization of a combination of HVT + SVT and also HVT + LVT cells. The analysis with sub-V_T energy model leads to the conclusion that different architectures are suitable for different constraints. A suitable design is a synergy between parallelism and utilization of various threshold options. However, with stringent low energy dissipation requirements combined with moderate throughput requirements, unfolded architectures synthesized with SVT cells are the most appropriate option. In this analysis, the multi-V_T, implementations do not show a significant advantages over single V_T implementations.

Chapter 9 presents a decimation filter chain, which is fabricated in 65 nm CMOS. The simulation results are validated by silicon measurements and demonstrate that low-power standard threshold logic (SVT) and different architectural flavors are suitable for a low-power implementation. Silicon mea-

surements prove functionality down to 350 mV supply, with a maximum clock frequency of 500 kHz, having an energy dissipation of 102 fJ/cycle.

PART 4: ANALYSIS ON STANDARD CELL BASED MEMORIES (SCM) IN SUB-V_T DOMAIN

Memories are an important part of many digital systems. There is a need for optimization of memory blocks that can be used for energy efficient design. The main options for embedded memories which may be operated reliably in the sub-V_T domain are: 1) specially designed SRAM macros, and 2) storage arrays built from flip-flops or latches. Standard SRAM designs require non-trivial modifications to function reliably in the sub-V_T regime. This part focuses on an alternative method for designing memories that are optimal for sub-V_T domain operation. This part contains Chapter 10, which shows that for standard-cell based ultra-low-power designs, standard-cell based memories (SCMs) are an interesting alternative to full-custom SRAM macros which must be specifically optimized to guarantee reliable operations. The main advantages of SCMs are the reduced design effort, reliable operation for the same voltage range as the associated logic, high speed (when compared to corresponding full-custom macros), and good energy efficiency for maximum-speed operation.

PART 5: FUTURE WORK

Finally, some hints towards future direction of the work related to this thesis is given in Chapter 11.

2 Power

Presently miniaturized electronic devices are getting more important in medical, sensor networks, and many other portable device applications. Engineers aim to develop ultra compact and low power circuits. The emphasis on the power consumption is also enormous in general purpose processors and other devices. The device design for low power consumption compared to the same device that is designed for low energy dissipation shall lead to very different solutions. This chapter sheds some light on this power versus energy design constraints. As maximum power consumption is bounded by both operational frequency and the amount of heat produced in the device that can be tolerated and it is related to the power density parameter [5].

Power optimization has gained emphasis in recent years. In 2011 the International technology roadmap for semiconductors (ITRS) published their paper [1], where a road map for power-aware design was given until 2025. This includes improvements in both dynamic and static power consumption. Various methods are proposed for these power optimization that include:

1. **Frequency Islands**
   The techniques exploits the spread of power by blocks designed to operate at different frequencies. Thereby, the peak power consumption and the peak current spikes are reduced. The cost of this technique is larger area with complicated engineering steps.

2. **Near-Threshold computing**
   The idea is to operate the design at around supply voltage of 400 - 500 mV, which is close to the threshold voltage of the standard devices

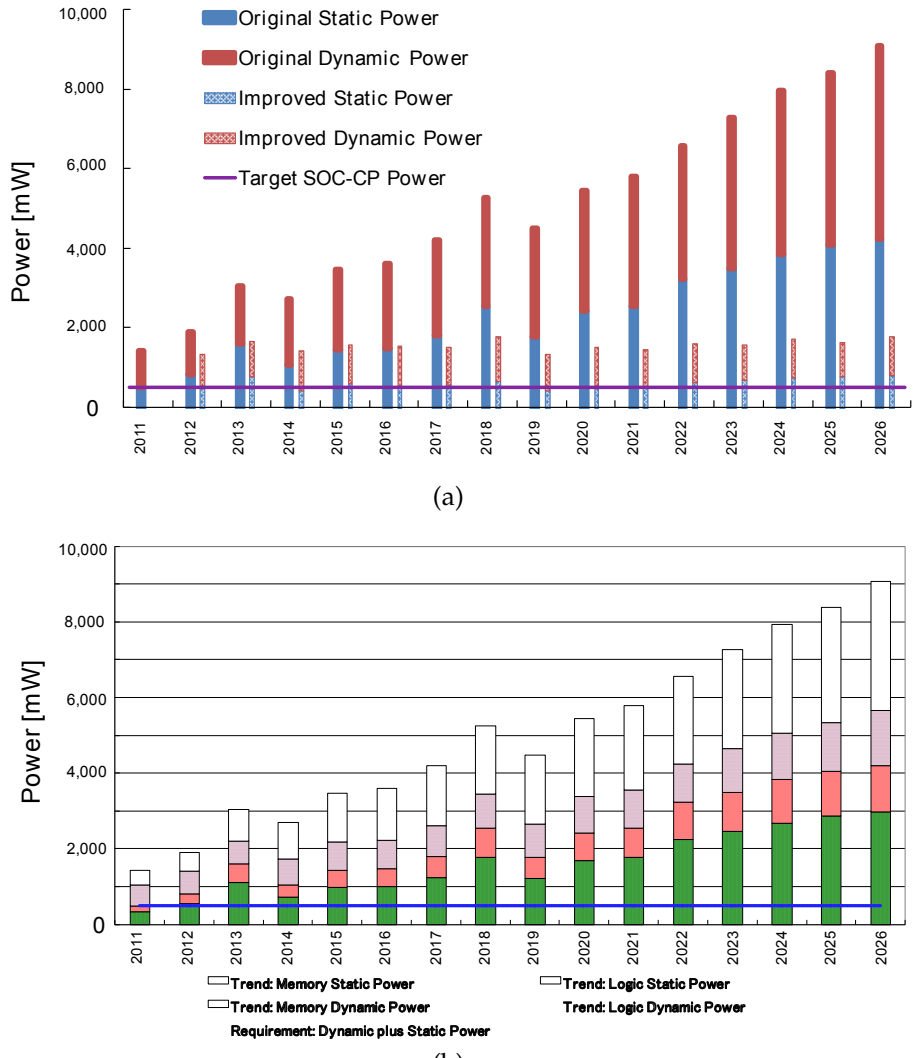


Figure 2.1.: (a) Impact of Low-Power Design Technology on SOC Consumer Portable Power Consumption (b) SOC Consumer Portable Power Consumption Trends. [1]

in 65 nm CMOS. This reduces the dynamic power in a quadratic manner. The cost is lower operating frequency, however, some level of moderate throughput can be maintained.

3. **Hardware/Software Co-partitioning**
   The co-partitioning here is based on the behavioral level analysis w.r.t power. This requires various levels of software interfaces and controls units.

4. **Heterogeneous parallel computing**
   This technique uses various types of processors in a parallel computing architecture that also help reduce the peak power. However, the idle power may increase due to higher leakage current.

5. **Power-Aware Software**
   Power consumption is used as the key parameter that defines the processes within the software that is used to run on the hardware. The technique has higher engineering complexity.

6. **Asynchronous Design**
   This techniques exploits the fact that there is no clock in the circuit, therefore, the periodic power consumption is defused. The design may result in higher area and there are no efficient automated computer aided design (CAD) tools that take the register transfer logic (RTL) to silicon.

With the application of these techniques the ITRS predicts reduction in power consumption as shown in the Figure. 2.1(a) and the Figure. 2.1(b). These figures show the trends of power consumption until 2026 and give the contribution of static and dynamic power for both logic and the memory.

2.1. POWER CONSUMPTION IN CMOS

The total power consumption for a digital circuit is given as

$$P_{\mathrm{T}} = \underbrace{\alpha C_{\mathrm{tot}} {V_{\mathrm{DD}}}^2 f_{\mathrm{clk}}}_{P_{\mathrm{dyn}}} + \underbrace{I_{\mathrm{leak}} V_{\mathrm{DD}}}_{P_{\mathrm{leak}}} + \underbrace{\alpha T_{\mathrm{sc}} I_{\mathrm{SC}} V_{\mathrm{DD}}}_{P_{\mathrm{sc}}}, \tag{2.1}$$

here P_{T}, P_{dyn}, P_{leak} and P_{sc} represent the total, the dynamic, the leakage and the short circuit power consumption, respectively. In 2.1, α is the switching activity or switching factor of the circuit, f_{clk} the clock frequency, and C_{tot} the total capacitance within the circuit. I_{leak} represents the leakage current in

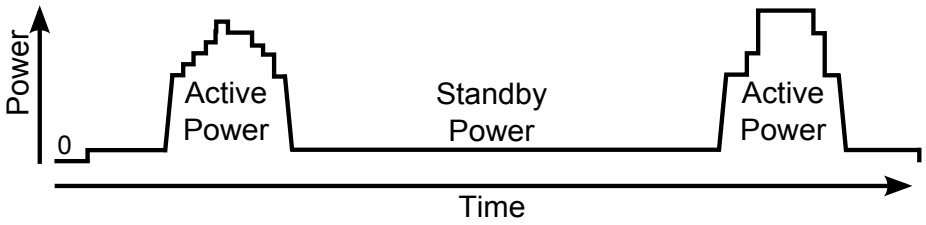


Figure 2.2.: Power profile of portable devices

static mode, i.e., when the circuit is not performing any operation. I_{SC}, is the short circuit current and T_{sc} represents the time when there is a direct path from supply voltage V_{DD} to ground. This gives the peak current consumption for this specific time period and is proportional to the dimensions of the transistor [6]. The parameter T_{sc} may be given as

$$T_{sc} = \frac{T_r + T_f}{2}, \tag{2.2}$$

here T_r and T_f, represent the rise time and fall time of the input signal, respectively. From the equation it is seen that reduction of the supply voltage causes the power consumption to be reduced in a quadratic manner for the circuits that have higher dynamic power consumption. However, for leakage or short circuit, a scaling of V_{DD} reduces power linearly.

Power consumption in any digital circuit can be divided into two main branches active power and static power. The active power consists of both dynamic and short-circuit consumption. The static power consumption compromises of leakage consumption only. The predictions from ITRS show that both the active and static power consumption have almost equal impact on the total power consumed by the devices.

2.1.1. ACTIVE POWER

Specifically the increase in the number of devices within the same area due to reduction in size of the transistor leads to an increase in the power consumption density caused by increased switching activity and also higher frequencies, i.e., higher dynamic power consumption. This means that higher power consumption lead to an increase in operational cost, with burden on the environment. To combat such high power demands the systems are designed so that they have at least two modes; one active mode, where the main processing is performed, second, is the standby mode where the system is idle. The idle mode has high impact on the reduction of the over all power profile of the design. The power profile of a design with these modes is presented in

Figure. 2.2. The power shown in the Figure. 2.2 is instantaneous power consumed by the circuit. From [7] the average power consumed in a certain time interval is given as

$$\widehat{P}_{\text{avg}} = \frac{1}{\Delta t} \int_{T_0}^{T_0+\Delta t} P_{\text{inst}}(t)dt, \tag{2.3}$$

where P_{inst}(t) is the instantaneous power consumed in the circuit, T_0 and $T_0 + \Delta\, T_0$ are start and end time of the interval that the average power is to be determined.

SHORT-CIRCUIT POWER P_{SC}

Short-circuit power consumption P_{sc} occurs during logic transitions. When the input signals changes from high to low or vise verse, the transition in the signal voltage have a finite rise and fall time. During this transition a direct path from supply voltage and ground is formed. This leads to a peak current flow and causes short-circuit power consumption. As an example, consider an inveter circuit that has a single PMOS and NMOS device as pull up network and pull down network, respectively, in a CMOS implementation. As shown in Figure. 2.3, the short-circuit current I_{SC} flow during the finite input slope, as both NMOS and PMOS are conducting during this transition. The P_{sc} is proportional to the switching activity similar to the dynamic power P_{dyn} [6].

Although, at higher supply voltages i.e. nominal V_{DD} the P_{sc} still has some effect on the total power consumption. However, the impact has reduced with scaled technologies. Furthermore, when the supply voltage is scaled down to lower voltages i.e. when $V_{\text{DD}} < (V_{\text{TN}} + |V_{\text{TP}}|)/2$, the impact of short circuit power consumption is not seen. The reason is that the devices never conduct currents simultaneously [6]. Therefore, the P_{sc} can be ignored for sub-V_{T} domain.

2.1.2. STATIC POWER

As the dimensions of the transistor are scaled down, the leakage current (I_{leak}) within the transistors increases due to thinner gate-oxides and other dimensional effects. This causes higher static power consumption. The major currents in I_{leak}, consists of channel leakage, diode leakage, and gate leakage.

CHANNEL LEAKAGE

The channel leakage current consists of subthreshold current, the drain induces barrier lowering (DIBL) leakages, as well as channel edge current.

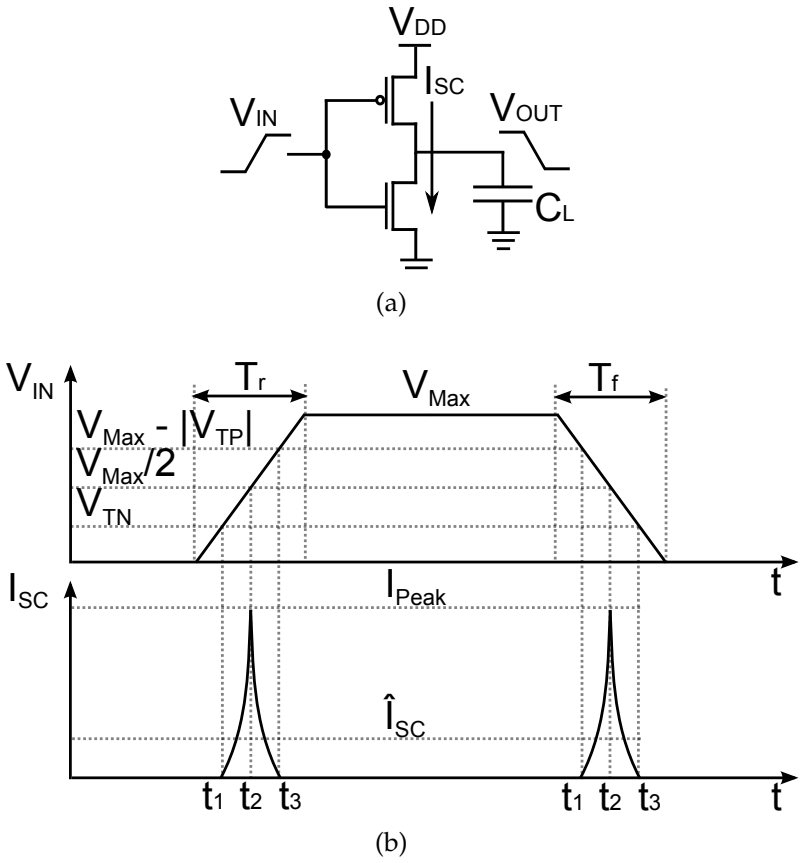


Figure 2.3.: Short-Circuit currents in CMOS Inverter

1. **Subthreshold current**
   This leakage is specially observed in short channel devices, where, the current flows even when the voltage across the gate and the source (V_{GS}) is below the threshold voltage (V_T). This leakage is higher for devices that have low V_T, i.e., closer to zero volts.

2. **DIBL current**
   In short channel transistors, the source and the drain region are physically close enough to affect each other, thereby affect the channel leakage current. The voltage at the drain increases the surface potential at the source, due to the drain potential the depletion region underneath

the channel is widen. The potential barrier is therefore lowered to a level that enables the source to inject more carriers into the channel for a given gate potential.

3. **Channel edge current**
   Physical deformities around the the gate region causes abrupt transitions. These transitions eliminate the lateral encroachment of the field around the oxide layer into the channel area, called the bird's beak [7]. That results in an increase in current near the channel edge, which is viewed as a parasitic that lowers the effective threshold.

DIODE LEAKAGE

The diode leakage compromises of two main parts, pn-junction reverse bias leakage and gate-induced drain leakage (GIDL).

1. **PN-junction reverse bias leakage**
   While in normal operation the source and the drain pn-junctions with the bulk are both reversed. This reverse bias causes leakage current that has two main components namely, reverse saturation current and generation current. The fundamental current in the pn-junction is called reverse saturation current. The generation current is caused by the thermal generation of electron-hole pairs within the region.

2. **Gate-induced drain leakage (GIDL)**
   Due to a high electric field between gate and drain, caused by the gate-drain overlap region leads to current leakage from the drain to the substrate.

GATE LEAKAGE

Nano-meter devices face various effects due to diminished gate oxide thickness. This diminished thickness leads to current leakage directly through the gate and is referred to gate oxide tunneling. There are two main gate leakage components namely, gate-to-channel direct tunneling, and source/drain extension-to-gate overlap tunneling currents.

1. **Gate-to-channel direct tunneling**
   Gate direct tunneling current is produced by the quantum-mechanical wave function of a charged carrier through the gate oxide potential barrier into the channel, which depends not only on the device structure but also on the bias conditions [8].

2. **Drain extension-to-gate overlap tunneling**
   In very short channel transistors, the portion of the gate overlap with the drain and the source becomes larger compared to the total gate length. When the gate voltage is between 0 V and the channel flat-band voltage, an accumulation of charges is formed in the poly-silicon that eventually leads to a source and drain extension-to-gate overlap tunneling current.

2.2. POWER MINIMIZATION TECHNIQUES

Various methods are employed to reduce the power consumption in a given design. These methodologies range from top level optimization, for example an algorithm or architectural optimization to low level optimization where gate or even transistors are tweaked to form low power consuming circuits. Furthermore, some of the optimization are beneficial for Static power reduction and some for minimization in the dynamic power consumption. In this section an overview of these techniques are discussed.

2.2.1. ACTIVE POWER REDUCTION TECHNIQUES

The main components of the active power consumption are the operational frequency, switching activity, capacitance and supply voltage as predicted in (2.1). Following is an overview of some of the techniques used to reduce the active power.

1. **Multiple supply voltage (V_{DD})**
   Supply voltage reduction is an effective strategy to reduce the power consumption. However, it may not be optimal. This is because the indiscriminate reduction in V_{DD} causes a delay increase in all the gates in the design. An advantageous approach is to scale V_{DD} selectively. The section can be based on the gates that fall in the following two categories [9].

   - Gates belonging to a path that complete their evaluations earlier than the rest of the circuit,
   - Gates that have to drive large capacitances and will benefit from the same delay increment.

   Furthermore, a more *Modular supply voltage* scaling approach is also applicable in the cases where circuit blocks have different speed and can be separated. As an example, consider a design with processing block and a controller. The data-path has a critical path delay of T_x and the controller block has a critical path delay of $T_x/2$. In this case the V_{DD} for the controller is lowered to a point where the critical path delay is

increased to T_x. In this case some voltage level converters are necessary to make the communication between the blocks possible. Specifically, they are needed when the module with the lower supply voltage has to drive the gates at the higher V_{DD}. Furthermore, in this technique multiple supply voltages are needed in addition to main supply voltage, they may require additional DC-DC converters.

Aggressive reduction in supply voltage (V_{DD})
The relation between dynamic power and the V_{DD} leads to the fact that supply voltage reduction yields quadratic reduction in power. Furthermore, it also helps in reduction of static power consumption. This is one of the most effective knobs to play with when reduction in power is needed. The side effect of V_{DD} include reduced reliability due to process variations and an increase in the delays of the gates or the designed circuit, thus, reduction in operational frequency. Therefore, various methods such as pipelining and (or) parallelism/interleaving are applied to compensate for the degradation in speed.

Parallelism involves adding a replica of the same hardware (e.g. an additional adder unit in an ALU) connected in parallel. The circuits can then operate at half the original intended input sample rate and are sill able to maintain the original throughput. As the reduce speed requirements on the circuits are in place, the V_{DD} can be lowered to the point where the original throughput are met. Although the capacitance is increased by a factor of two and some more due to data multiplexing circuits and additional routing. On the other hand, the clock is reduced by a factor of two, which compensates for the increase in area. However, the main reduction in power is achieved due to the reduction in V_{DD}. This technique is usually applied to the circuits that are not area constrained.

Pipelining is another approach that is often employed achieve the dynamic power reduction. As the propagation delay T_d is inversely proportional to the supply voltage reduction and is given in [6] as

$$T_d \propto \frac{V_{DD}}{V_{DD} - V_T}, \tag{2.4}$$

The T_d of a given design is reduced by the introduction of additional registers in the critical path, called pipelining. Once a design is pipelined then this circuit is capable of performances higher with respect to the original speed requirements. This higher speed performance is than traded off by the reduction of supply voltage in accordance with V_{DD} in (2.4). Hence, an overall decrease in power consumption is attained.

Other techniques at data-path or architectural level may include replacement of slower blocks with their faster counter parts. This is done to get back the performance loss due to V_{DD} reduction.

2. **Multiple clock domains**
   Application Specific integrated circuits (ASICs) now-a-days comprises of multiple blocks of functionality incorporated in one framework. Therefore, there are modules or blocks specifically designed with various throughput requirements on the same chip. Various blocks require different clock domains and supply voltages. The least critical block are therefore supplied with slower clock and lower supply voltages, to gain in power reduction. In this kind of implementations various on-chip clock generators are needed, they will cost in area and some power consumption losses. However, optimized clock domains results in a reduced active power consumption as gains are seen due to reduced clock frequencies [10].

3. **Gated clocks**
   A clock gate is one of the most common ways to reduce the power consumption. This technique is employed in the cases where part of the circuit does not need to be active all the time and the processing can be turned off with the use of clock gates. However, clock gating does not help reduce the leakage power consumption, as only the clock is turned off [9].

4. **Dynamic voltage and frequency scaling (DVFS)** DVFS is a power-management technique that employed both the voltage and frequency scaling to reduce the overall power consumption. This techniques is usually applied in processors with multiple cores. These cores can then be monitored for process activity. Based on the activity or task requirements, the operating frequency and voltage of a processor core can dynamically reduced or increased. However, as the maximum processor and memory clock frequencies are being saturated, there is a need for reduction larger static power consumption, smaller dynamic power range and better idle/sleep modes. Each of these developments limit the potential energy savings resulting from DVFS [11].

5. **Glitch reduction**
   Glitch is a false transition that may occur in combinational logic before the final result from the gate is not completely evaluated. The signal transfer variations in the inputs of a gate cause these false evaluations that are corrected once the actual input are stable at the considered gate

or set of gates. These false transitions cause erroneous charging and discharging of the load capacitances within the circuit and lead to high active power consumption. These glitches can be reduced if the circuits are designed with *balanced paths*. Some of the structures have inherent balance path properties, in case of adders the kogge-stone adder [12] has balance data path compared to an ordinary Ripple carry adder [9]. Furthermore, by introduction of pipeline also reduce the glitches with in a chain and the false signals are not allowed to propagate throughout the logic. This helps in the reduction of dynamic power consumption.

6. **Transistor sizing**
   As discussed in [9], input capacitance of a CMOS gate is directly proportional to its size and its speed. In cases where the gates achieve faster performance than set by the requirement, the gates can be resized. Here, the prime candidate for downsizing is the largest gate. The delay for that gate will increase proportionally to the downsizing of its dimensions. Therefore, it is must be performed where the largest impact is gained. However, this method is not trivial as downsizing a path also affects the delays in other paths with shared logic. It is hard to isolate and optimize a single data path. Therefore, this is mainly employed in EDA tools.

7. **Resource allocation**
   Appropriate resource allocation results in reduction of switching activity, that in return reduces the power consumption. As shown in [9], shared data-path reduces the area however, it increases the switching activity due to multiplexing and has detrimental effects on power consumption. This is due to the fact that the data is completely randomized because of multiplexing and it can lead to scenarios where all one are switched for all zeros, hence a higher switching activity is generated. Implementation of independent data-paths may lead to lower power consumption if the data is correlated [13].

8. **Word-length optimization**
   In fixed-point mathematical operator implementations the output word-length increases to maintain precision. This increment in word-length is needed so that overflows are avoided. Consider an adder, it requires N+1 output bit to avoid overflow in a two's compliment addition. Now if a direct form implementation of a M-Tap Finite-Impulse-Response (FIR) filter is considered where there are M-1 adders connected together in a series, with the first adder with N bit each input. This will lead to

a M+N adder at the end if proper word-length management is not performed. In this case a formula ($N + \lceil log_2(M) \rceil$) based on N and M is used to optimize the word-length and maintain precision [14]. Furthermore, in many cases it is highly unlikely that every arithmetical operation yield an overflow. Thus, truncation or rounding may be performed to minimize word-lengths, on the cost of less precision and more truncation errors. This results in less hardware requirement and leads to lower power consumption.

9. **Arithmetic optimization**
   Arithmetic optimization includes use of architectures that produce less switching activity in-order to calculate the mathematical results. It can also include numerical strength reduction to reduce the complexity of a given mathematical operation [13]. The basic mathematical operations are ranked in terms of their complexity, with highest to lowest given as, division, multiplication, addition/subtraction, and bit-shift. As an example consider a case with a FIR filter is designed where the coefficients are known constants. In this case the multiplication can be designed with respect to these constants and the resources are reduced to simple bit-shifts and additions, that lower the complexity rank of the multiplier. This improves the performance in terms of area, power, and speed.

2.2.2. STATIC POWER REDUCTION TECHNIQUES

The predictions from ITRS show that static power consumption has almost 50 % impact on the total power consumed by the devices. The main components of the static power consumption are the switching activity, leakage currents, delays, and supply voltage as predicted in (2.1). Following is an overview of some of the techniques used to reduce the static power.

1. **Time-Multiplexing for leakage reduction**
   As leakage is one of the major sources for static power consumption, which is directly linked to number of gate within the design. Therefore, many architectures are optimized in such a way that common resource are reused and time-multiplexed to complete an algorithms operation. In time-multiplexed design, partial computations of an operation is performed and the partial result is stored to accommodate execution of another instruction. Once the resource is available the stored partial results are then reused to continue with the operation. The results are delivered after the completion of task. Compared to a direct mapped circuit the time-multiplexed circuit need a controller and a register file

or a memory block. This overhead becomes negligible when a larger direct mapped circuit is converted to its time-multiplex counter part.

2. **Transistor stacking**
   In this technique, gates with more stacked transistors are used i.e., transistors are connected in series. The stacking of transistors leads to slight increase in the voltage at the intermediate point between the source and drain junction in a CMOS network. The increase in the -ve V_{GS} increases the threshold voltage V_T of the transistors that is not connected to ground directly. This decreases the leakage current through the transistors and thereby, reduces the static power consumption.

3. **Multiple device thresholds**
   Devices with multiple threshold can be used to trade off speed for power. In a 65 nm CMOS technology, standard cells are often available in at least 3 threshold options, characterized as high-V_T (HVT), standard-V_T (SVT), and low-V_T (LVT). The HVT devices have a high threshold voltage and due to this, the leakage currents are orders of magnitude lower than the LVT devices. Therefore, it is possible to utilize the LVT cells in the timing-critical paths, while HVT cells can be used elsewhere. This techniques mainly helps reduce the static power consumption of the design. As the parts of the design that are not critical have devices that will leak less. However, some reduction in active power is also observed. This reduction is due to reduced gate channel capacitance in the off state and a small reduction in signal swing on the internal nodes of a gate [9].

4. **Reverse body bias**
   Reverse body bias technique is used to reduce leakage current in the idle mode. The idea implied in this technique increases the V_T of the gates and thereby reduce the leakage current. This increase is achieved by applying a negative voltage to the bulk terminal of the transistor. Experiments on NMOS transistor in a 65 nm CMOS bulk technology has shown that the reverse body bias technique reduces leakage around $20\,\% \sim 30\,\%$ when the nominal supply voltage used. One of the drawbacks of this technique is the requirement of additional +ve or -ve bias voltage.

5. **Power gating**
   In idle mode the leakage power is the main source of power consumption. This leakage power is reduced with the use of power gates while the system is in idle mode. This accomplished with the employment of

large sleep transistors, that are used to cut off the supply voltage to the rest of the circuitry. These transistors are placed on the supply rail or sometimes on both supply rail and the ground rail as discussed in [9]. The transistors are controlled by a sleep signal, that is inactive during the normal operation. However, once the circuit goes in the idle mode the sleep signal is activated and this cuts the supply off from the rest of the circuitry. A finite resistance of these transistors result in additional noise within the supply rails for the circuits attached to them. Therefore, to minimize these noise fluctuations the transistors have to have very low resistances i.e., they are up-sized. However, this huge size results in an area increase. In contrast to the clock gating technique discussed earlier, power gating results in loss of the stored information. Therefore it can only be applied to designs that allow such behavior, otherwise, retention memory blocks are needed to stores the data that is required after the wake-up. These additional memory blocks will impact the benefits of the power reductions. The second option is that all the registers are connected to a non-gated supply and therefore are ready for use once the rest of the circuit wakes-up. This will also require additional power routing and will dampen the power savings.

2.3. SUMMARY

This chapter gives an overview of power consumption related trends for a CMOS design. Insight into types of power consumption are presented, together with a brief overview of techniques that are used to reduce them. The techniques involved algorithmic to low level tweaking within a gate to reduce the power consumption. However, one technique that reduces all major components of the power consumption is supply voltage scaling V_{DD}. Other architectural improvements yield major advantages once employed together with V_T scaling. In the next chapters focus on V_T scaling and design space analysis that goes hand-in-hand with this technique.

Part I

Sub-V_{T} Domain Fundamentals

This part consists of two chapters, first, an introduction to the fundamentals of the weak inversion region or the sub-threshold (sub-V_{T}) domain is given. Second, a gate-level sub-V_{T} energy characterization flow is briefly discussed. This part includes material published in the following paper.

- **S. Sherazi**, J. Rodrigues, O. Akgun, H. Sjöland, P. Nilsson, "Ultra low energy design exploration of digital decimation filters in 65 nm dual-V_{T} CMOS in the sub-V_{T} domain", *Microprocessors and Microsystems: Embedded Hardware Design (MICPRO), Elsevier*, vol.37/4-5, 2013.

3

Sub-V_T / Weak Inversion Fundamentals

This chapter deals with the fundamentals of the weak inversion region or the sub-threshold (sub-V_T) domain. Rigorous voltage supply scaling is employed to achieve low energy dissipation. This reduces the ratio between on-current (I_{on}) and off-current (I_{off}) in the transistor. Hence, the transistor operates in the sub-V_T domain or weak inversion region [15–17]. The severely degraded on/off current ratio I_{on}/I_{off} and increased sensitivity to process variations are one of the main challenges for sub-V_T circuit design [18] [19] in 65 nm CMOS technology and below.

3.1. WEAK INVERSION CONDITIONS

The transistor is considered to be in the weak inversion when the drain-to-source voltage V_{DS} is higher than zero volts, together with constraint on gate-to-source voltage V_{GS} described as [20],

$$V_A \leq V_{GS} < V_B, \tag{3.1}$$

where, V_A is the voltage at which the transition between depletion and weak inversion occur, V_B is the voltage at which the transition between weak inversion and moderate inversion occurs. These voltages are given as,

$$V_A = V_{FB} + \Phi_F + \gamma\sqrt{\Phi_F + V_{SB}}, \tag{3.2}$$

$$V_B = V_{FB} + 2\Phi_F + \gamma\sqrt{2\Phi_F + V_{SB}}, \tag{3.3}$$

where, source-to-bulk voltage is represented by V_{SB}, Φ_F represents the Fermi potential, and γ is the body factor. Here, V_{FB} is gate voltage at which the valence and conduction bands are not bent, it is referred to as flat-band voltage and is written as,

$$V_{FB} = \Phi_{GC} - \frac{Q_{ss}}{C_{ox}}, \tag{3.4}$$

here, Φ_{GC} represents the *work function* difference between the gate and the channel material. The fixed charge in the gate oxide is Q_{ss} and C_{ox} represents the gate capacitance per unit. The surface potential is equal to $\Phi_F + V_{SB}$ on the onset of weak inversion and it is equal to $2\Phi_F + V_{SB}$ on the onset of moderate inversion. Furthermore, the body factor γ is described as

$$\gamma = \frac{\sqrt{2qK_{Si}N_{sub}}}{C'_{ox}}. \tag{3.5}$$

Here, q represents the electron charge, K_{Si} is the permitivity of silicon, the substrate doping concentration is N_{sub}, and C'_{ox} is the gate capacitance per unit area.

3.2. SUB-V_T CURRENTS

In the sub-V_T domain the drain-source current I_{DS} changes exponentially with a change in the gate-source voltage V_{GS}. This is due to the fact that the carriers injected at the source end of the channel moves towards drain by diffusion. The drain-to-source current I_{DS} for a long-channel NMOS transistor operated in weak inversion is represented as

$$I_{DS} = I_S e^{\frac{V_{GS}-V_T}{nU_T}} \left[1 - e^{\frac{-V_{DS}}{U_T}}\right], \tag{3.6}$$

where n is the slope factor and is described as

$$n = 1 + \frac{C'_d}{C'_{ox}}, \tag{3.7}$$

where C'_d represents the depletion capacitance per unit area and capacitance ratio is written as

$$\frac{C'_d}{C'_{ox}} = \frac{\gamma}{2\sqrt{2\Phi_F + V_{SB}}}, \tag{3.8}$$

therefore,

$$n = 1 + \frac{\gamma}{2\sqrt{2\Phi_F + V_{SB}}}, \tag{3.9}$$

for practical use n is below 1.6. Furthermore, I_S in (3.6) is called specific current and is expressed as

$$I_S = 2n\mu C'_{ox} {U_T}^2 \frac{W}{L}, \tag{3.10}$$

where μ represents the carrier mobility, U_T is the thermal voltage also known as the Boltzmann voltage (it is 26 mV at room temperature). Here, W and L are the width and length of the transistor, respectively. V_T in (3.6) represents the threshold voltage, and depends on V_{SB} as

$$V_T = V_{T0} + (n-1)V_{SB}, \tag{3.11}$$

Here, V_{T0} is the threshold voltage defined when the V_{SB} is zero. In order to avoid parasitic bipolar effects the source junction is reversed biased or only slightly forward biased. This puts a constraint on V_{SB} that has to be larger than about -4U_T. Therefore, V_T can be increased with respect to V_{T0}. When V_{GS} is set to zero the saturation current is written as

$$\begin{aligned} I_0 &= I_S e^{\frac{-V_T}{nU_T}}, \\ &= I_S e^{\frac{-(V_{T0}+[n-1]V_{SB})}{nU_T}}, \\ &= 2n\mu C'_{ox} {U_T}^2 \tfrac{W}{L} e^{\frac{-(V_{T0}+[n-1]V_{SB})}{nU_T}}. \end{aligned} \tag{3.12}$$

The saturation current is controllable by V_{SB}, and therefore, (3.6) is reduce to

$$I_{DS} = I_0 e^{\frac{V_{GS}}{nU_T}} \left[1 - e^{\frac{-V_{DS}}{U_T}}\right]. \tag{3.13}$$

For a PMOS transistor, by changing the sign of both current and voltage the same equation becomes valid. Furthermore, it is seen from (3.12) and (3.13), that the drain-to-source current I_{DS} decreases exponentially with the increase in the threshold voltage V_T. Furthermore, variations in the V_T causes variation in the performance of the device with respect to speed. This is due to the fact that the speed is inversely related to V_T. The current also increases for positive V_{GS} and it decreases for negative V_{GS}. However, if the potential is further decreased, the current increases again as shown in [20]. This is due to leakage from drain to bulk and leakage through the gate oxide. The temperature also affects the current in this domain. This is due to the changes in carrier mobility μ, the thermal voltage U_T, and the slope factor n. Temperature also effects

the V_{T0}. The carrier mobility dependence on the temperature is expressed as

$$\mu(T) = \mu(T_r)\left(\frac{T}{T_r}\right)^{-v}, \tag{3.14}$$

where room temperature is represented by T_r, the variable v is usually between 1.2 and 2. The U_T changes linearly with the change in temperature. For the slope factor n, equation based on Fermi potential ϕ_F is used to describe the dependence on the temperature as shown below

$$\phi_F = U_T ln\left(\frac{N_{sub}}{n_i}\right), \tag{3.15}$$

where n_i represents the intrinsic carrier concentration, and is exponentially dependent on temperature [20]. Lastly, the threshold voltage V_{T0} decreases linearly with increase in the temperature, and the temperature based threshold equation is written as

$$V_{T0}(T) = V_{T0}(T_r) - c(T - T_r), \tag{3.16}$$

where c is the threshold voltage temperature coefficient which is usually 0.5 mV [20].

From these equations it is deduced that for weak inversion the current increases exponentially for higher temperatures as the slope factor increases (it is less steep) and the threshold voltage decreases. Furthermore, various leakage currents also effect the total current in the transistors as shown in Figure 3.1. Here I_g is the leakage from the gate, I_d is the current leakage from the diodes and I_c is the leakage through the channel. Some of the more interesting leakage current phenomena are discussed in the next sections.

3.2.1. DRAIN-INDUCED BARRIER LOWERING (DIBL)

In long channel transistors V_{T0} also depends on applied gate voltage, as all the depletion charge underneath the gate is originated from the MOS field effects. In this case the reverse-biased drain junction and the depletion region of the source are ignored. These effects become severe once the length of the transistors are scaled down [20]. This is due to the fact that the source and drain fields already deplete a portion of the region below the gate. Reason being that the drain is physically located too close to the source and is able to interact with the depletion region around it. This causes V_{T0} to decrease as strong inversion is achieved with lower voltages. Therefore, V_{T0} decreases with the scaling of transistor length L.

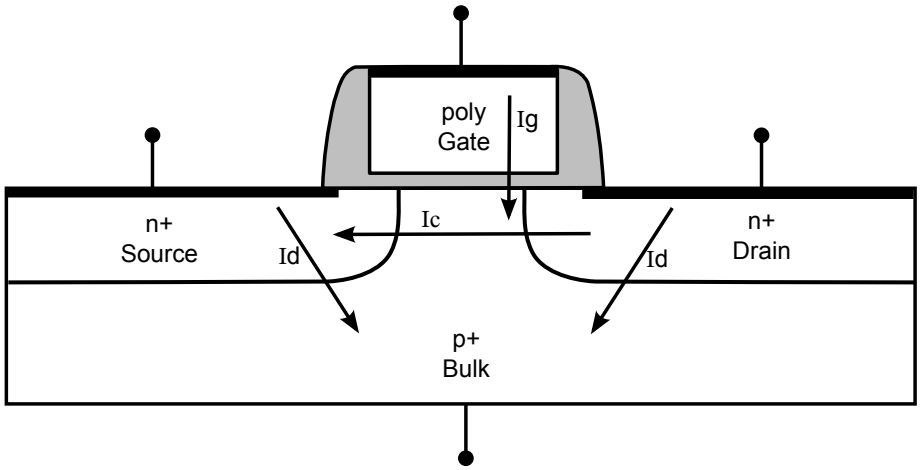


Figure 3.1.: Leakage currents in the transistor

Similarly, the same effect of lower V_{T0} is achieved by increasing the drain-source voltage. This is possible due to the fact the increased drain-source voltage causes width of the depletion region near drain-junction to increase. The potential barrier around the source is lowered and it becomes easier for the source to inject carriers into the channel for a given V_{GS}. Consequently, the threshold decrease with increased V_{DS} [21], and the effect is called Drain-induced barrier lowering (DIBL). Now, the threshold voltage is not constant, instead it is a function of the operating voltages. Furthermore, the enhanced carrier concentration in the channel leads to an increased off-state current. As given in [7], the current equation (3.13) is modified to incorporate the DIBL effect as

$$I_{DS} = I_0 e^{\frac{V_{GS}+vV_{DS}}{nU_T}} \left[1 - e^{\frac{-V_{DS}}{U_T}}\right], \tag{3.17}$$

where v is the DIBL factor.

An extreme form of DIBL may occur if the drain-source voltage V_{DS} is increase excessively. This creates a short circuit between source and drain, that leads to malfunction in a transistor. This short circuit leads to a sharp increase in the current of the transistor and the phenomenon is called "punch-through" [20]. In this state the gate loses its control over the current that flows through the channel. An upper bound on the V_{DS} is defined by the punch-through effect. The effects of DIBL are worrisome as they make the transistors prone to changes in the operational voltages. As an example in dynamic memories the sub-threshold current of the access transistor becomes a function of the voltage on the bit line and hence dependent on the data

that is to be stored or read. Therefore, DIBL becomes a data dependent noise in the dynamic memories that causes faulty operation, which renders these memories less useful.

3.2.2. REVERSE BIAS LEAKAGE

During normal operation of a MOS transistor, the Drain-Bulk and Source-Bulk pn-junction are reverse biased. The small leakage current due to this reverse bias is generated because of the two phenomenons called "reverse-saturation current" and "generation current".

The reverse-saturation current is the fundamental reverse-bias leakage current in pn-junction. On the other hand the generation current is produced by the electron-hole pair created due to heat produced in the pn-junction within the space charge region [20]. This current is represented as

$$I_{\text{reverse}-\text{bias}} = A_{\text{j}}(J_{\text{s}} + J_{\text{gen}}), \tag{3.18}$$

where A_j is the area of the pn-junction, the J_s represents the reverse-saturation current density and J_{gen} is the generation current density.

3.2.3. GATE-INDUCED DRAIN LEAKAGE (GIDL)

The high electric fields between gate and drain causes current to leak from the drain to the substrate (bulk). This phenomenon is called gate-induced drain leakage [22]. The GIDL current I_{GIDL} is given as

$$I_{\text{GIDL}} \propto A E_{\text{ox}}^{\frac{5}{2}} e^{-\frac{B}{E_{\text{ox}}}}, \tag{3.19}$$

where $A \propto E_g^{-\frac{7}{4}}$ and $B \propto E_g^{\frac{3}{2}}$. E_g is the band gap and it is very sensitive to the electric field. E_{ox} is the electric field that exists in the thin oxide. For large E_{ox} the drop in the deep-depletion layer becomes large enough to allow tunneling in the drain via near-surface traps. In that case several trap-assisted events become possible. The trap-assisted events are typically present for low electric fields, which are a strong function of temperature. The minority carriers emitted to the incipient layer are then laterally removed to the substrate, completing a path for the gate-induced drain current.

3.2.4. GATE LEAKAGE CURRENT

With the scaling of transistor length the gate oxide also becomes smaller and thinner. The advantage of having an ultra-thin gate oxide is a reduction in short channel effects, which enhances the speed performance of the transistor. However, the disadvantage is direct gate-leakage current as shown in

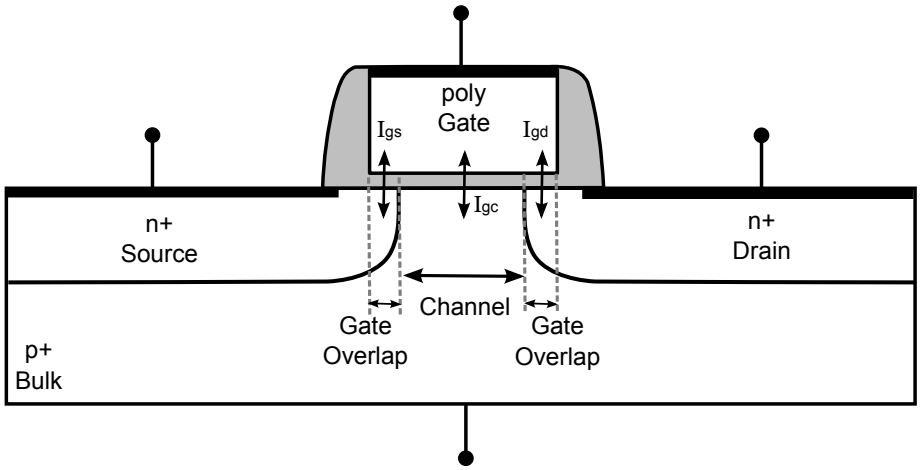


Figure 3.2.: Leakage currents in the transistor

Figure 3.2. Furthermore, as the electric field E_{ox} increases, the tunneling current through the gate oxide I_{gc} will increase exponentially. This phenomena is called Fowler-Nordheim tunneling [23]. This means that for circuits that require charge-conservation or charge-bootstrapping, including the sample and hold circuits, a significant performance degradation is expected for gate oxide thickness $t_{ox} < 1.5\,nm$. This is also true even for low voltage operations [8]. The currents through source and the drain extensions (SDE) are also called the gate overlap tunneling current represented by I_{gs} and I_{gd}, respectively, which also become dominant for gate voltages between the channel flat-band voltage and the SDE flat-band voltage.

3.3. PERFORMANCE IN SUB-V_T

The performance in sub-V_T is associated with the on-current I_{on} and from (3.17), the current I_{on} when V_{GS}=V_{DS}=V_{DD} is given by [2]

$$I_{on} = I_0 e^{\frac{V_{DD}-V_{T0}+\upsilon V_{DD}}{nU_T}} \left[1 - e^{\frac{-V_{DD}}{U_T}}\right]. \tag{3.20}$$

For simplification purposes the above equation is rewritten based on the assumptions that the whole on-current I_{on} of fully saturated transistor, driven by the supply voltage V_{DD} flows through the capacitor C and given as

$$I_{on} \approx I_0 e^{\frac{V_{DD}}{nU_T}}. \tag{3.21}$$

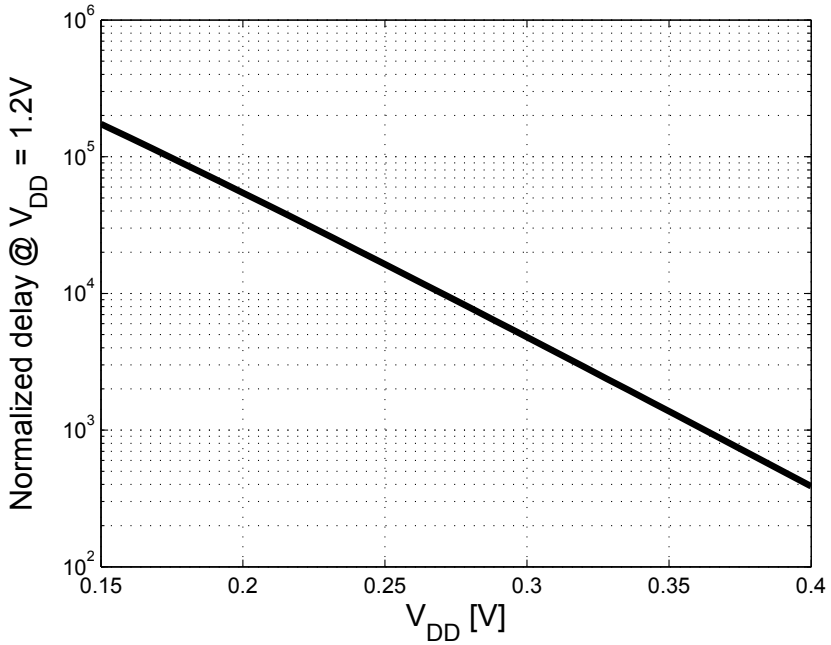


Figure 3.3.: Delay of circuit normalized to value at V_{DD} = 1.2

Here the assumption is that the supply voltage V_{DD} is at least 4 times of the thermal voltage U_T. From this equation it is evident that with a scaled supply voltage V_{DD}, the current decreases exponentially. Thereby, the delay will increase in a similar fashion. This is expressed as

$$T_d = \frac{C}{I_{on}} \frac{V_{DD}}{2},$$

$$T_d = \frac{CV_{DD}}{I_{on} e^{\frac{V_{DD}}{nU_T}}}. \tag{3.22}$$

The critical path of a circuit normalized at nominal supply voltage (here, V_{DD} = 1.2V) is plotted versus the scaled sub-threshold supply voltage V_{DD} in 65 nm CMOS technology is shown in Figure 3.3. The y-axis is in log scale. This shows that the delay of the said circuit decreases exponentially with respect to the scaled supply voltage. This is significantly different from traditional circuits that are operated in strong inversion region.

3.3.1. EFFECT OF THE CAPACITANCE IN SUB-V_T OPERATION

Aggressive supply voltage scaling down to the sub-V_T regime also affects the capacitance formed within the transistor. However, the effects are not severe. In [24] a simplistic analytically model for intrinsic capacitance related to the transistors channel between two terminals is described. The small-signal transistor capacitance C is defined by the charge Q flowing into a terminal i, caused by the changes in the voltage V of another node j, is given by

$$C_{ij} = \frac{\partial Q_i}{\partial V_j}. \tag{3.23}$$

For CMOS logic, the input sees the self-capacitance from the gate, and the self-capacitance is then written as

$$C_{gg} = \frac{\partial Q_g}{\partial V_g}. \tag{3.24}$$

In order to calculate the C_{gg} at the gate, the bias voltages are to be specified, such as V_{GS} is set equal to input voltage V_{in}, the V_{DS} is set equal to the output voltage V_{out}, and V_{SB} is equal to zero, i.e., there is no body bias. The V_{out} changes with a delay w.r.t. V_{in}. This due to the fact that the transient of V_{out} is dependent on the gate propagation delay w.r.t. the switching of the input capacitance. Therefore, for the C_{gg} evaluation, the V_{out} is assumed to be constant and equal to the value before the input switching [24]. Consequently, for an NMOS transistor, the input self-capacitance C_{gg} at a rising input transition is evaluated with V_{DS} = V_{DD} and for a falling input transition it is evaluated with V_{DS} = 0. Furthermore, for simplicity purposes of the analysis the author in [24] has used the relation of proportionality between C_{gg} and the channel capacitance $W \cdot L \cdot C_{ox}$. The W and L are the effective width and length of the channel and C_{ox} is the effective gate capacitance based on the effective gate oxide thickness of the transistor. Therefore, the effective C_{gg} is written as

$$C_{gg} = WLC_{ox} f(V_{in}), \tag{3.25}$$

where the function f(V_{in}) shows the dependence of the input self-capacitance on the input voltage. For the sub-V_T operation the f(V_{in}) function is approximated and is evaluated based on the observation that the charges at the gate, and the bulk depletion, are set by the gate voltage. Furthermore, the bulk depletion charge Q_{dep} normalized to $W \cdot L \cdot C_{ox}$ is given as

$$\frac{Q_{dep}}{WLC_{ox}} = \frac{K_{10X}^2}{2}\left[-1 + \sqrt{1 + 4\left(\frac{V_{GS} - V_{fb} - V_{SB}}{K_{10X}^2}\right)}\right], \tag{3.26}$$

where K_{10X} and the flat band voltage V_{fb} are both BSIM parameters that model the effect of non-homogeneous channel doping on the threshold voltage and the gate-bulk flatband voltage [24]. In order to get the function f(V_{in}) the differentiation of (3.23) is performed and used in (3.26) for $V_{GS} = V_{in}$ and with the body bias equal to zero, i.e., $V_{SB} = 0$ and the limit of the function f(V_{in}) is when $V_{in} \longrightarrow 0$. The result is given as

$$\begin{aligned} f(V_{in}) &= \lim_{V_{in}\to 0} \frac{\partial Q_G / WLC_{ox}}{\partial V_G}, \\ &= \frac{1}{\sqrt{1-4V_{fb}/K_{10X}^2}}. \end{aligned} \tag{3.27}$$

For a 65 nm technology, it was reported in [2] that the gate capacitance in sub-V_T is smaller than the above threshold and the reduction was found to be around 20 %. In a 65 nm technology, this junction capacitance increases around 30 %. This increase and decrease in the two capacitance lead to a neutral effect for most of the practical operations [2] [24]. However, the effect due to changes in the capacitance are observable in SRAMs as the gate and junction capacitance are also effected by the bit line accumulative capacitance.

3.3.2. I_{ON}/I_{OFF} IN SUB-V_T OPERATION

When the transistor is operated in the sub-V_T domain the ratio of the I_{on}/I_{off} current decreases and this adversely effects the performance of the device. From (3.21), the off-state current is given as an approximation

$$I_{off} \approx I_0, \tag{3.28}$$

and therefore, the on-off ratio is given as another approximation

$$I_{on}/I_{off} \approx e^{\frac{V_{DD}}{nU_T}}. \tag{3.29}$$

This shows that the ratio depends exponentially on the supply voltage V_{DD} and secondly on the slope factor n of the technology. For a 65 nm CMOS technology, n is typically between 1.3 - 1.5 [2]. Typically the ratio degrades by a factor of 10-15 times per 100 mV in the sub-V_T domain, this is shown in the Figure 3.4. The I_{on}/I_{off} ratio degrades by a huge factor in sub-V_T domain compared to above threshold domain. The impact of this degradation in the ratio means that the off-state current of the transistor has become significant compared to that of the on-state current of the transistor. This indicates an enormous impact of leakage current on overall power consumption or energy dissipation.

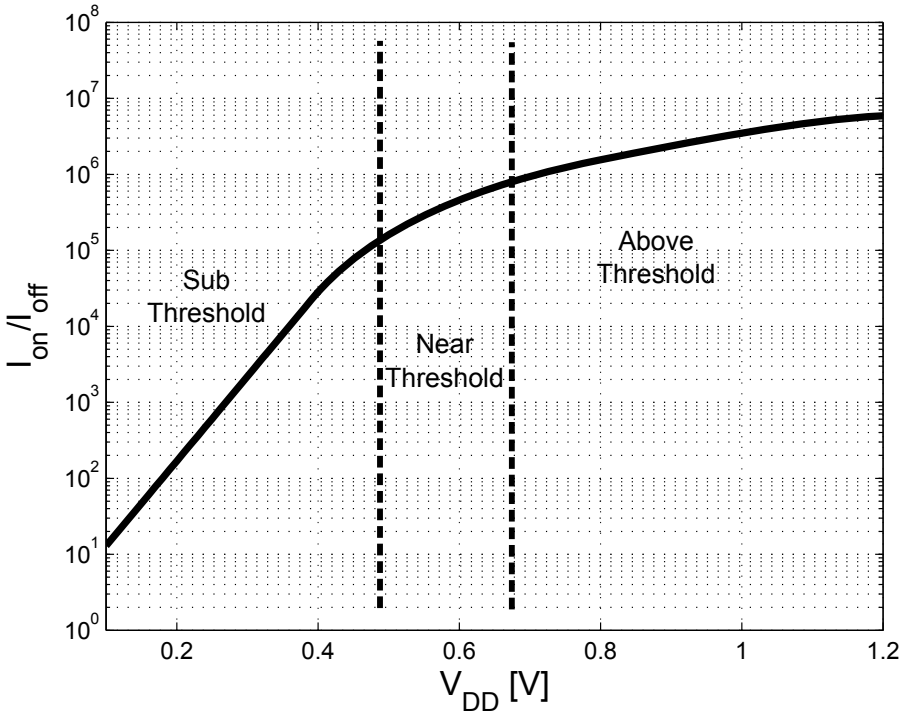


Figure 3.4.: I_{on}/I_{off} versus Supply voltage V_{DD} [2]

Secondly, in the case where multiple transistors of same dimensions, connected to a single node, suffer from degradation of robustness. As discussed in [2], consider a gate where there are m transistors connected in parallel with a node X. Assume that *m-1* transistors are off and only one transistor conducts current through itself. In this case the correct operation requires the on-state current I_{on} of the transistor to dominate the over all off-state currents I_{off} of all the *m-1* transistors, so that the high and low level of over all current is distinguishable. Furthermore, when the gate is to be operated in the sub-V_T domain, care has to be taken so that the number of m transistors connected to a common node must be kept one or two orders of magnitude below I_{on}/I_{off} ratio. Therefore, when the gates are to be operated in the sub-V_T domain, they are to be redesigned so that the count for the transistors decreases exponentially per gate. This imposes a constraints on practical circuits that have high fan-in or memories.

Table 3.1.: Parameters for 65 nm CMOS low power devices [4]

Device	**n**	V_{T0}	λ_{DS}	λ_{SB}
NMOS	1.39	0.598	$9e^{-2}$	$9.9e^{-2}$
PMOS	1.27	0.532	$8e^{-2}$	$1.1e^{-1}$

DEVICE STRENGTH

In the device the threshold voltage also depends on the drain-source voltage through the drain induced barrier lowering (DIBL) effect and the bulk-source voltage through the body effect, and is written as

$$V_T = V_{T0} - \lambda_{DS} V_{DS} - \lambda_{SB} V_{SB}, \tag{3.30}$$

where is λ_{DS} is the DIBL coefficient and λ_{SB} is the body effect coefficient. The author in [4] has reported the values of these parameters for a 65 nm CMOS high threshold low power option, which is given in Table 3.1. Furthermore, from these parameters, the strength of the device within the sub-V_T domain is also formulated in [2] and is given as

$$\beta = I_S \frac{W}{L} e^{-(V_{T0} - \lambda_{SB} V_{SB})/nU_T}. \tag{3.31}$$

This shows that the *W/L* ratio of the transistor can determine the strength of the device. In a 65 nm CMOS technology there are various intrinsic threshold options that may be used to get a specific strength of the transistor. Furthermore, application of a body bias through bulk voltage dynamically may also play a role in the strength of the transistor.

3.3.3. REVERSE BODY BIAS (RBB)

Different techniques are used to reduce the leakage current in the idle mode and one of them is the reverse body bias. The idea implied in this technique is to increase the threshold voltage of the gates and thereby reduce the leakage currents or I_{off}. Figure 3.5, shows the I_{off} or leakage current estimated for the NMOS devices when the bias voltage is swept from 0 to 1V. It is observed that with the help of the RBB sweep an optimum reverse body bias voltage is found for various supply voltages within super-threshold or above-V_T domain. However, as soon as the supply voltage V_{DD} is scaled down to the sub-V_T regime, it becomes less trivial to find an optimum bias voltage V_{bias}. In

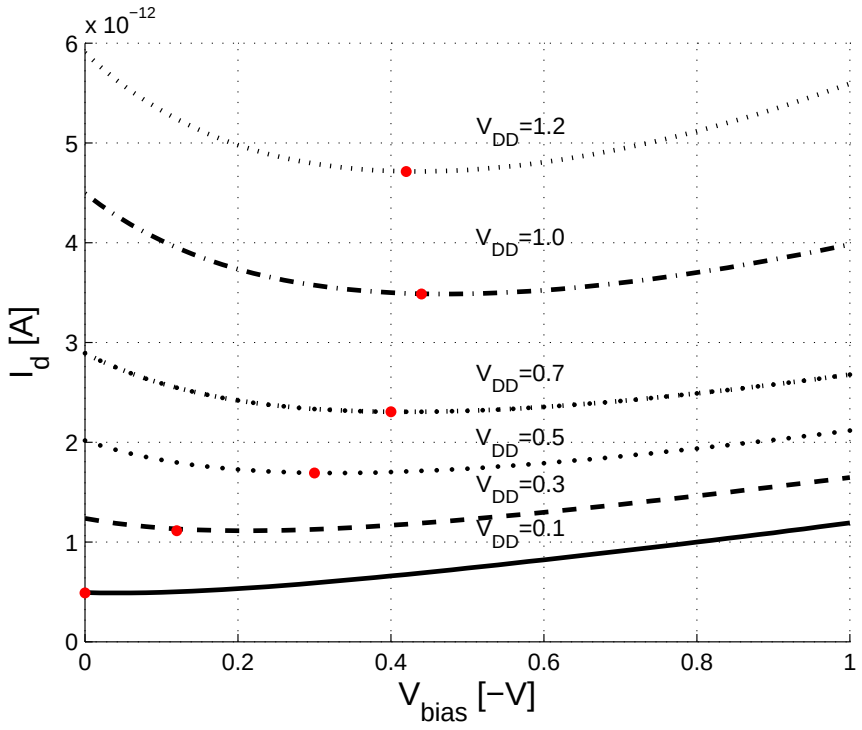


Figure 3.5.: NMOS leakage at various supply voltages V_{DD} versus V_{bias}

the last case of a V_{DD} of 0.1 V no benefit is achieved by the application of RBB. The dot on the plots indicates the lowest leakage current with respect to the V_{bias}. For V_{DD} of 1.2 V the lowest leakage is observed at around $V_{bias} = 0.4$ V and for supply voltage 0.1 V a lowest leakage point by increased V_{bias} could not be observed. Similar leakage behavior is observed for a PMOS device.

The scenario when the supply voltage is scaled down to zero volts and only V_{bias} is swept, the leakage current at the internal p-n junction diodes is then isolated. In this case various properties of the device such as the threshold voltage V_T, substrate current I_{bulk}, the channel current I_d, and the power, are controlled with the biasing of the bulk voltage. The effect of RBB on these properties is shown in the Figure 3.6. It is observed that the V_T increases and both drain and bulk currents go down, then the diode currents take over. Thereby, the overall current goes up exponentially when the reverse body bias voltage is increased further than -0.5 V for an NMOS. Therefore, the power

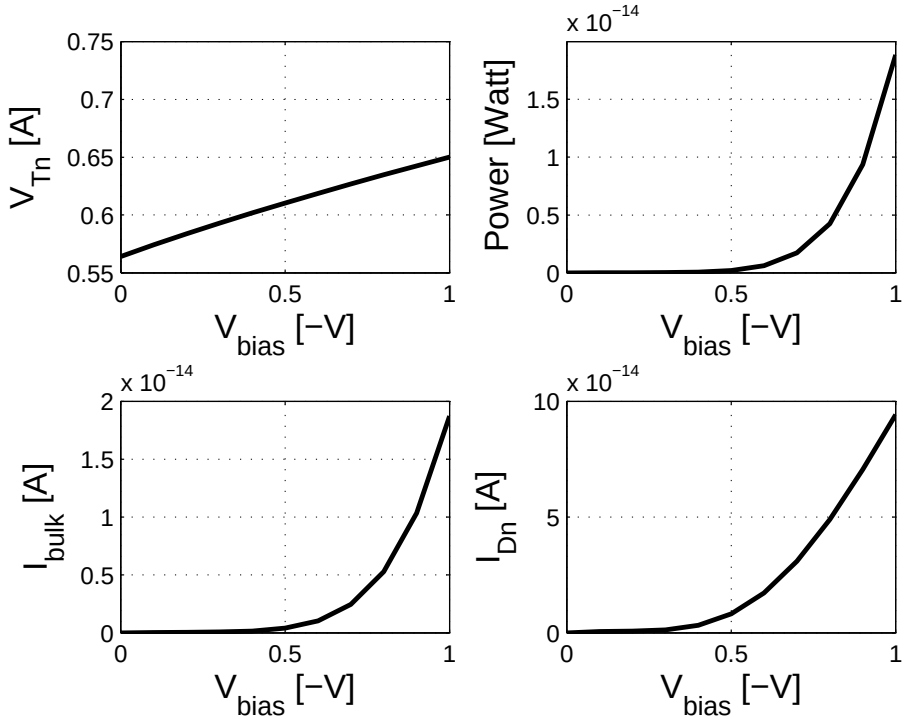


Figure 3.6.: V_T, Power, I_{bulk} and I_D of NMOS at V_{DD} = 0V

also increases due to this increase in currents. Furthermore, the threshold increases with increase of V_{bias}.

From these experiments on NMOS transistor, it becomes apparent that the RBB technique reduces leakage around 20 % ∼ 30 % when the supply voltage is above threshold. However, when the device is operated in sub-V_T domain RBB does not provide any considerable benefit. Furthermore, the adoption of a RBB voltage to increase the threshold voltage leads to a decrease in robustness and strength of the device [2].

3.3.4. NMOS/PMOS BALANCE IN SUB-V_T REGIME

When the supply voltage is scaled to the sub-V_T regime the strength of NMOS and PMOS degrades, correspondingly. Furthermore, this causes an imbalance between the two with respect to noise margin and the rise/fall transition time [2]. Secondly, The output voltage levels also degrade due to the imbalance of strength between the NMOS and PMOS and this leads to an increase

in the leakage power consumption of the subsequent logic gate. At ultra-low voltages, the NMOS/PMOS imbalance is typically much higher, thereby degrading the noise margin [25]. The imbalance factor is describe in [2] and is given by

$$IF = max\left(\frac{\beta_p}{\beta_n}\frac{\beta_n}{\beta_p}\right) \geq 1. \tag{3.32}$$

From (3.32), the imbalance factor is seen as the ratio of the strength between the stronger and the weaker transistor. Furthermore, from the equation it is evident that the strength ratio is irrespective of whether the stronger one is the PMOS or the NMOS. The strength β is dependent on the technology and it may be either greater or less than 1 when compared to super-V_T regime. The intrinsic threshold voltage of PMOS/NMOS are dependent on the doping process, therefore the intrinsic threshold may vary significantly for the devices when compared to each other [25]. From (3.31), it is known that the strength is sensitive to the intrinsic threshold of the device. Thereby, a slight difference in the intrinsic threshold voltage between PMOS and NMOS may lead to a large difference in the strength of the two devices.

In [2] the author describes that the NMOS and PMOS transistors, when operated in the sub-V_T region, suffer from a high imbalance, that means that the IF factor is much greater than 1. In order to match the strengths of the two devices, a considerable increase (more precisely, by a factor IF) in the strength of the weaker transistor, is required. As an example in [2] the specific case of the 65 nm CMOS technology is discussed, the imbalance factor IF is stated to be around 7 i.e., IF $\approx$ 7, when the devices are operated in the sub-V_T domain. It is stated that the NMOS strength is larger than PMOS by the same factor. Furthermore, it is stated that this sub-V_T domain imbalance factor is much greater than that IF of above threshold, which is found to be only 1.8. Therefore, in order to get the perfect balance among the two devices the PMOS has to be strengthen by IF $\approx$ 7.

The increase in the strength of PMOS may be achieved by the following steps as described in [2]. First, an increase in strength is obtained by application of Forward body bias (FBB) on PMOS and strategically no body bias is applied on NMOS. This is achieved when the bulk terminal of both devices are connected to the ground. With this step the IF decreases to half (IF $\approx$ 3.5) of the initial value of IF $\approx$ 7. Here, the RBB as discussed in Sec. 3.3.3, is not applied to NMOS because of the fact that it would require the generation of voltages that are below ground. That would also require additional boosting circuits like charge pumps that leads to higher design effort, which are typically impractical in ultra low power chips with tight constraints on the energy cost. Second, the IF may further be reduced by the use of re-sizing the PMOS

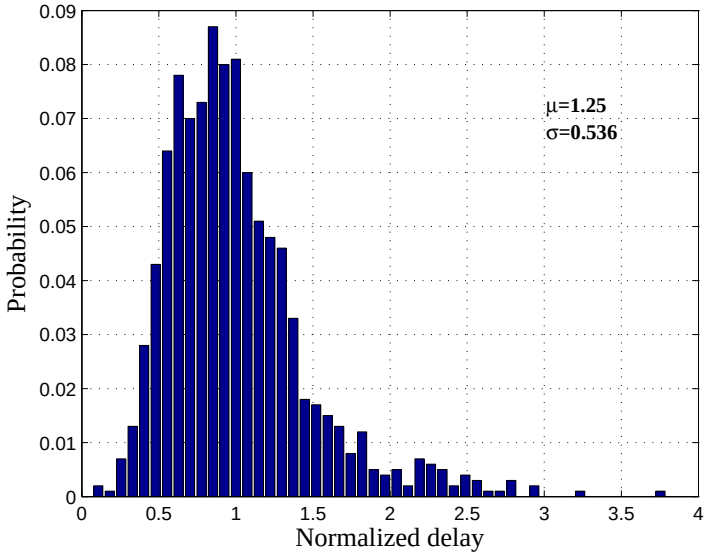


Figure 3.7.: 1000 point Monte-Carlo delay simulation of an Inverter @250 mV.

and thereby increase the strength of the device. However, this increase in size lead to larger capacitance and higher energy dissipation.

3.3.5. PROCESS VARIATIONS

The exponential dependency of the sub-V_T currents on process parameters like threshold voltage (V_T), doping, and slope factor, makes the transistor performance and functionality extremely vulnerable to process variations [27]. Thus, the transistor's performance in terms of delay and reliability is considerably degraded compared to super-V_T operation [28–30]. This reduces the maximum attainable throughput and adds extra energy overhead to the design. To illustrate performance degradation due to process variation, 1000 (point) Monte-Carlo based simulations are performed on an inverter circuit. The delay variation is analyzed on a minimum sized inverter. The cell selected in this case has minimum dimensions for its transistors. Figure 3.7 show the delay variation normalized to the mean delay (μ), due to process variations and mismatches @250 mV supply voltage. The delay variation at this low voltage is high and can deviate by a factor of ~ 4 in the worst case. This is considerable large when compared to variation in nominal voltage,

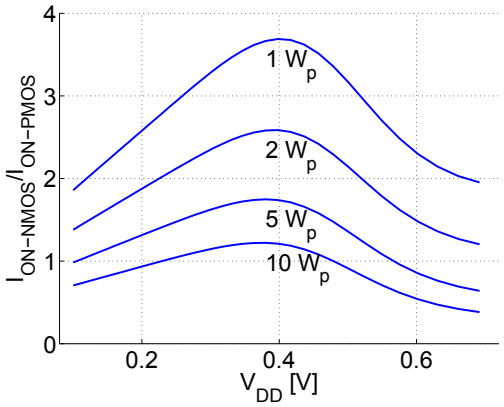


Figure 3.8.: The ratio of active currents of HVT-NMOS and HVT-PMOS in sub-V_T. V_T of transistors ∼700 mV. W_P is the min size allowed in the technology. The NMOS transistor has the minimum width.

that is within 20 %. This shows that the energy dissipation will also very correspondingly.

Full-custom cells (FCL) are often used for the realization of sub-V_T optimized circuits [31] [32]. The FCL may have up-sized transistors or additional transistors to combat of process variations in the sub-V_T regime. Up-sizing transistors improves the timing in the sense that it may equalized rise/fall time, and increases noise margins at the cost of higher area and energy [31]. In modern sub-micron CMOS technologies different threshold options are available, which gives designers the opportunity to address the leakage energy by employing gates consisting of high threshold transistors, whereas if high speed performance is required, gates with low threshold transistors are used [33]. However, this method is mainly employed on gate level. The advantages of using different threshold options on lower level, i.e., inside gates are explored in [34]. Where the authors show that the transistor strength balancing is one of the techniques that effect positively on design's performance and reliability. The driving balance of a circuit depends on different process parameters, i.e., the primary process parameter V_T and secondary parameters drain induced barrier lowering (DIBL) and slope factor. The traditional method to equalize the imbalance is transistor sizing. This is done by a relatively low-size ratios of PMOS and NMOS in the super-V_T regime. However, the transistor size-ratios become very large in the sub-V_T domain, see Fig. 3.8.

The peak current ratio between PMOS and NMOS is found in the sub-V_T regime. Furthermore, it is observed that by upsizing the PMOS transistor by $10\times$, a strength balancing is still not achieved.

The balanced strength improves the gate's stability and robustness, as the switching threshold voltage (V_m) moves to its ideal value (V_{DD} /2) and thereby increases the noise-margins (NM). Unbalanced switching threshold and low NMs are among the main sources of functionality and stability failures in sub-V_T regime. Therefore, designing the gates with maximum possible NM ($NM_{Low} = NM_{High}$) is of vital importance.

To speed-up the performance bottlenecks in gates and balance the driving strength of pull-up and pull-down networks (PUN and PDN). The authors in [34] employ a technique referred to as dual-V_T gates (DVTG). Where selected transistors are replaced by their lower-V_T equivalent. The readers are encouraged to read more on DVTG in [34].

3.4. SUMMARY

This chapter summarizes the fundamental of sub-V_T regime operation. The current equations encompassing various effects of leakage for example GIDL or DIBL are presented. Discussion on fundamental concepts such as the ratio between on-current I_{on} and off-current I_{off} in the transistor is given. When the supply voltage is scaled down to ultra-low voltages, the delay of the gate also degrades. This degradation of delay is shown to be exponential once the supply voltage is scaled below the threshold voltage of the adopted technology.

The issue of strength imbalance of PMOS/NMOS becomes an important parameter when the design is considered for the sub-V_T operation, as it directly influences the robustness and leakage power consumption of the circuit. In order to improve the IF and reduce process variations, techniques such as FBB, sizing, or DVTG may be employed as effective nobs to play with. However, the device is operated in the sub-V_T domain RBB does not provide any considerable benefit, if applied as a standalone technique for leakage reduction. Furthermore, the adoption of a RBB to increase the threshold voltage leads to the decrease of robustness and strength of the device.

4

Sub-V_{T} Energy Profiling

Energy dissipation of a circuit is of high importance in sensor node and medical implantable devices. As the operations in these devices depend on the energy that can be provided by the battery encased with them. The energy cost is one of the most important factors in determination of best suited design for an electronic device that is used in medical implants. As designs may have a different amount of power consumption but still, it can lead to same amount of energy dissipation.

When only the power consumption of a circuit is considered, it is known that higher computational performance require higher power consumption. Higher power consumption can be traded of by the time to gain a reduction in it. Consider an example of design that performs a certain task in a time period T, which is operated at nominal supply voltage V_{DD}. The design can be operated at a frequency f with the operation ending at time T. This gives a certain power consumption P during that period. Now, the same design is operated at twice the frequency $2f$ and the task ends in half the time $T/2$, in this case, the power consumed will be twice as high compared to the previous scenario. However, when energy is considered, both cases will have the same amount of energy dissipation as shown in Figure 4.1. The energies *E1* and *E2* represent the energy dissipated in the two cases, when the operating frequency is $2f$ and f, respectively. The y-axis depict the consumed power and the x-axis show the time spent to complete an operation.

In this case it may be observed that as the circuit is able to be operated at a high clock frequency $2f$, at the nominal supply voltage. Therefore, the V_{DD} may then be lowered to the point where the circuit when operated at clock frequency f gives zero slack. Therefore, to really achieve a gain in the energy

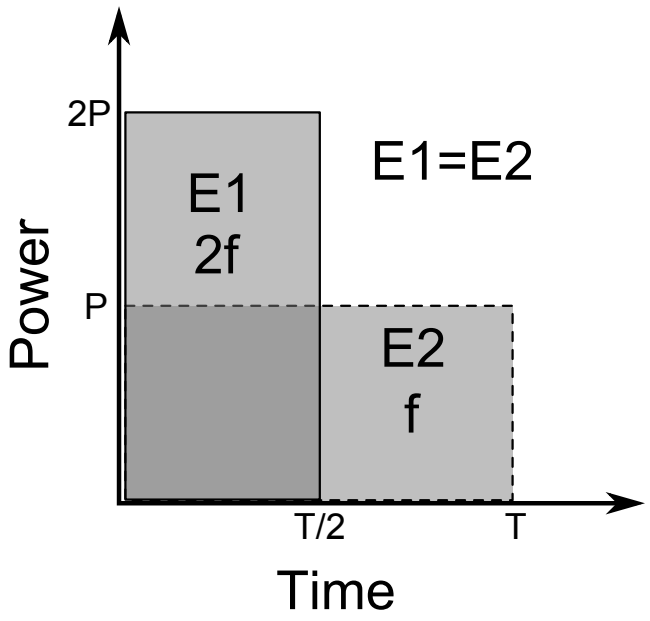


Figure 4.1.: Power and energy for an operation

dissipation, the V_{DD} is lowered in the case of a low input clock frequency. This will give a significant saving for the energy dissipation. Therefore, it is vital to look at the energy cost rather than the power costs in-order to optimize the designs for energy efficiency. From earlier discussions in Chapter 2 and 3, sub-V_T operations give the lowest energy dissipation. The next section discusses sub-V_T energy modeling for designs.

4.1. SUB-V_T MODELING

In order to exhaustively analyze the energy dissipation and the critical path delay of a given design with a certain architecture, a gate-level sub-V_T characterization flow is applied [35]. The benefit of such a flow is that it characterizes the circuit with respect to the sub-V_T regime. This characterization is necessary as the energy minimum operating point (E_{min}) lies somewhere in the sub-V_T domain, as shown in the Figure 4.2. This shows that the dynamic energy (E_{dyn}) scales down quadratically with the scaling of the supply voltage V_{DD}. On the other hand leakage energy increase exponentially at lower voltages. This is due to the fact that the gates become very slow and leakage dominates throughout the circuit. Therefore, there is a sweet spot where

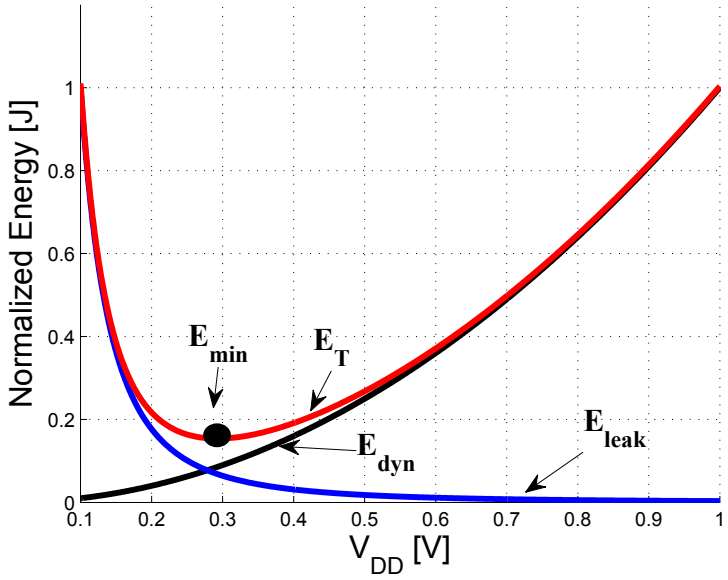


Figure 4.2.: Energy dissipation in circuit

the contribution of dynamic and leakage energy results in a local minimum total energy dissipation (E_T). This local minimum is described as the energy minimum voltage (EMV) point.

The sub-V_T characterization model is proposed by Akgun, et al. in [35], and is described in Section 4.1.1, which is expanded on for multi-threshold gates in Section 4.1.2.

4.1.1. SUB-V_T CHARACTERIZATION MODEL

The total energy dissipation E_T of static CMOS circuits operated in the sub-V_T regime is modeled as

$$E_T = \underbrace{\alpha C_{tot} {V_{DD}}^2}_{E_{dyn}} + \underbrace{I_{leak} V_{DD} T_{clk}}_{E_{leak}} + \underbrace{I_{peak} t_{sc} V_{DD}}_{E_{sc}}, \tag{4.1}$$

where E_{dyn}, E_{leak}, and E_{sc} are the average energy dissipation due to switching activity, the energy dissipation resulting from integrating the leakage power over one clock cycle T_{clk}, and the energy dissipation due to short circuit currents, respectively. The energy dissipation E_{sc} has been shown to be negligible in the sub-V_T regime [19]. The switching current causing the energy dissipa-

tion E_{dyn} results from sub-threshold currents [36], i.e., from the drain currents of MOS transistors whose gate-to-source voltage V_{GS} is equal to or lower than the threshold voltage V_T ($V_{GS} \leq V_T$). Whenever the sub-threshold current is not used to switch a circuit node, it contributes to E_{leak} together with all other types of leakage currents.

For a given clock period T_{clk}, (4.1) may be rewritten as

$$E_T = \mu_e C_{inv} k_{cap} {V_{DD}}^2 + k_{leak} I_0 V_{DD} T_{clk}, \tag{4.2}$$

where I_0 and C_{inv} are the average leakage current and the input capacitance of a single inverter, respectively. Furthermore, k_{leak} and k_{cap} are the average leakage and the capacitance of the circuit, respectively, both normalized to a single inverter. Moreover, μ_e is the circuit's average switching activity.

In the sub-V_T domain, it is beneficial to operate at the maximum achievable frequency to reach minimum energy dissipation per operation. In the following, (4.2) is therefore rewritten for the case where the clock period T_{clk} is equal to the critical path delay (T_{clk} denotes the critical path delay in the remainder of this section). The critical path delay itself may be written as

$$T_{clk} = k_{crit} T_{sw\_inv}, \tag{4.3}$$

where k_{crit} is the critical path delay of the circuit normalized to the inverter delay T_{sw_inv}. In [19], the delay T_{sw_inv} of an inverter operating in the sub-V_T regime is given by

$$T_{sw\_inv} = \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{4.4}$$

where n and U_t denote the slope factor and the thermal voltage, respectively. By introducing (4.4) into (4.3), the the critical path delay is now given by

$$T_{clk} = k_{crit} \frac{C_{inv} V_{DD}}{I_0 e^{V_{DD}/(nU_t)}}, \tag{4.5}$$

where, the reciprocal of (4.5) defines the maximum frequency at which the circuit may be operated for a given supply voltage V_{DD}.

Finally, the total energy dissipation E_T assuming operation at the maximum frequency is found by introducing (4.5) into (4.2), which yields

$$E_T = C_{inv} {V_{DD}}^2 \left[\mu_e k_{cap} + k_{crit} k_{leak} e^{-V_{DD}/(nU_t)} \right]. \tag{4.6}$$

The key parameters, which this sub-V_T characterization model relies on, are extracted from a fully placed, routed, and back-annotated netlists, with gate-level power simulations. For the architectural analysis the following chapters, (4.6) has been used. For more details, the reader is referred to [35].

4.1.2. MODELLING OF MULTI-V_T IMPLEMENTATIONS

The original energy model [35] was further developed to be able to handle multi-V_T implementations. In the original energy model, the $k_{factors}$ are calculated based on an inverter with a given threshold. However, in the multi-V_T case these $k_{factors}$ needs are calculated on the bases of the inverters of both threshold options. This method has been presented in [37].

The scaling factor for the capacitance is separated into two factors namely $k_{cap,1}$ and $k_{cap,2}$. The total capacitance for each threshold option within the circuit is also separated, which is given by $C_{Total,1}$ and $C_{Total,2}$. Where, $k_{cap,1}$ and $k_{cap,2}$ are capacitance scaling factors for the two threshold options. The coefficient $C_{Total,n}$ and $k_{caps,n}$ are given by

$$k_{cap,1} = \frac{C_{Total,1}}{C_{inv,1}}, \tag{4.7}$$

and

$$k_{cap,2} = \frac{C_{Total,2}}{C_{inv,2}}. \tag{4.8}$$

The coefficient k'_{cap} defines the total capacitance scaling factor of the circuit and is given by

$$k'_{cap} = k_{cap,1} + k_{cap,2}. \tag{4.9}$$

The effective inverter capacitance is given by

$$C'_{inv} = C_{inv,1} \times C_{r,1} + C_{inv,2} \times C_{r,2}, \tag{4.10}$$

where C'_{inv} is the effective inverter capacitance for the design implemented in multi-V_T. However, here the base capacitance C'_{inv} is calculated with respect to ratios of the capacitance of two inverter cells with different threshold options. The $C_{inv,1}$ and $C_{inv,2}$, represent the capacitance of a single inverter for the two threshold options, respectively. The factors $C_{r,1}$ and $C_{r,2}$ are their respective ratios in the circuit, and the ratios are specified as

$$C_{r,1} = \frac{C_{Total,1}}{C_{Total}}, \tag{4.11}$$

and

$$C_{r,2} = \frac{C_{Total,2}}{C_{Total}}, \tag{4.12}$$

where $C_{Total,1}$ and $C_{Total,2}$ are the respective capacitances of the two threshold options and C_{Total} is the total capacitance of the circuit. Similarly, the leakage

factor for the circuit is calculated for the threshold options separately and then combined to give the total leakage scaling factor, as

$$k_{\text{leak},1} = \frac{L_{\text{Total},1}}{L_{\text{inv},1}}, \tag{4.13}$$

and

$$k_{\text{leak},2} = \frac{L_{\text{Total},2}}{L_{\text{inv},2}}. \tag{4.14}$$

In (4.13) and (4.14), factors $k_{\text{leak},1}$ and $k_{\text{leak},2}$ are leakage scaling factors for the two threshold options, where $L_{\text{Total},1}$ and $L_{\text{Total},2}$ are the total leakage in the circuit for the respective options. The factors $L_{\text{inv},1}$ and $L_{\text{inv},2}$ are the average leakage current of the inverters. The coefficient k'_{leak} defines the total scaling factor of the circuit's leakage current and is specified as

$$k'_{\text{leak}} = k_{\text{leak},1} + k_{\text{leak},2}. \tag{4.15}$$

The effective inverter leakage current is specified as

$$L'_{\text{inv}} = L_{\text{inv},1} \times L_{\text{r},1} + L_{\text{inv},2} \times L_{\text{r},2}, \tag{4.16}$$

where L'_{inv} is the effective inverter leakage current. However, here the base current leakage L'_{inv} is calculated with respect to the ratios of the leakage current of two inverter cells with different threshold options. The factors $L_{\text{r},1}$ and $L_{\text{r},2}$ are their respective ratios within the circuit. These leakage current ratios are specified as

$$L_{\text{r},1} = \frac{L_{\text{Total},1}}{L_{\text{Total}}}, \tag{4.17}$$

and

$$L_{\text{r},2} = \frac{L_{\text{Total},2}}{L_{\text{Total}}}, \tag{4.18}$$

where $L_{\text{Total},1}$ and $L_{\text{Total},2}$ are the respective leakage currents and L_{Total} specifies the total leakage.

The critical path in multi-V_T implementations contain cells with two threshold options. Therefore, the timing factors are also calculated separately. Namely, $k_{\text{crit},1}$ and $k_{\text{crit},2}$ are the scaling factor for the critical path delay. The factors $T_{\text{crit},1}$ and $T_{\text{crit},2}$ represents the delay on the critical path by the corresponding cells, respectively. Therefore, the scaling factors are specified as

$$k_{\text{crit},1} = \frac{T_{\text{crit},1}}{T_{\text{inv},1}}, \tag{4.19}$$

and

$$k_{\text{crit},2} = \frac{T_{\text{crit},2}}{T_{\text{inv},2}}, \tag{4.20}$$

where $T_{\text{inv},1}$ and $T_{\text{inv},2}$ represent the average inverter delay for the threshold options. The coefficient k'_{crit} defines the total scaling factor for critical path delay of the circuit, which is specified as

$$k'_{\text{crit}} = k_{\text{crit},1} + k_{\text{crit},2}. \tag{4.21}$$

The currents due to the cells in the critical path will set the maximal speed limit of the circuit. Therefore, the ratios of this leakage current in the critical paths is specified as

$$TL_{\text{r},1} = \frac{TL_{\text{total},1}}{TL_{\text{total}}}, \tag{4.22}$$

and

$$TL_{\text{r},2} = \frac{TL_{\text{total},2}}{TL_{\text{total}}}, \tag{4.23}$$

where $TL_{\text{total},1}$ and $TL_{\text{total},2}$ represents the sum of the leakage currents that corresponds to the different threshold levels in the critical path. The factor TL_{total} represents the total leakage of the cells within the critical path. These ratios are used to calculate the effective off state current I_0, given as

$$I'_0 = I_{0,1} \times TL_{0,1} + I_{0,2} \times TL_{0,2}, \tag{4.24}$$

where, the coefficients $I_{0,1}$ and $I_{0,2}$ are the average leakage currents of a single inverter when the gate to source voltage is equal to zero for the two selected threshold options, respectively [19]. Finally, (4.2) and (4.6) are re-written as

$$E_{\text{T}} = \mu_{\text{e}} C'_{\text{inv}} k'_{\text{cap}} {V_{\text{DD}}}^2 + k'_{\text{leak}} I'_0 V_{\text{DD}} T_{\text{clk}}, \tag{4.25}$$

and

$$E_{\text{T}} = C'_{\text{inv}} {V_{\text{DD}}}^2 \left[\mu_{\text{e}} k'_{\text{cap}} + k'_{\text{crit}} k'_{\text{leak}} \text{e}^{-V_{\text{DD}}/(nU_{\text{t}})} \right]. \tag{4.26}$$

These two equations are used for the characterization of the circuits.

4.2. ENERGY MODEL FLOW

In this section, the flow developed for the energy model is described. Figure. 4.3, shows the flow chart of the sub-V_{T} energy model. The first step is to create a hardware description of the design that is to be tested or analyzed.

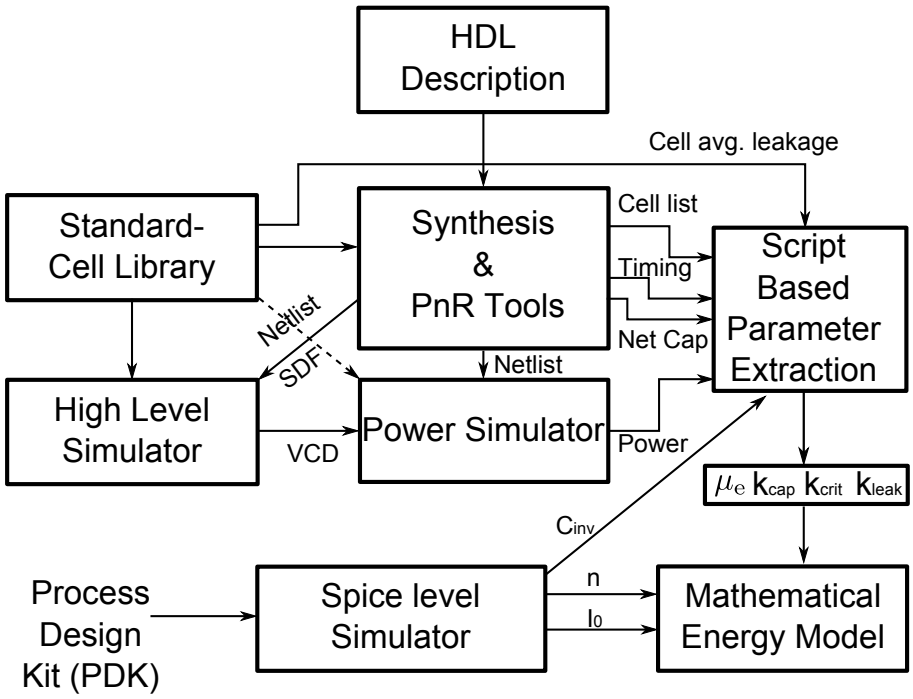


Figure 4.3.: Sub-V_T energy model flow

The circuit description may be performed with any hardware description language. The next step is the synthesis of such a description based on the standard cell libraries, usually provided by a vendor. In this case the synthesis is performed with the help of the Design Compiler. The synthesized netlist is then placed and routed with the help of Digital Implementation System tools. The placed and routed (PnR) netlist is then generated and again read into Design Compiler to generate reports of timing, net capacitance, and the list of cells used in the design. The netlist is also used to generate the back-annotated toggle information with the help of a simulator. In this case, High Level Simulators is used to generate toggle information that is stored in a "value change dump" (VCD) file. The simulator requires the netlist, the delay information stored in a "standard delay format" (SDF) file, which is generated by the PnR and Design compiler, and the standard cell library information. The back-annotated toggle information is supplied to a power estimation tool, together with the netlist, to generate the power profile of the design. The power is calculated based on the nominal voltage of the used technology. These reports

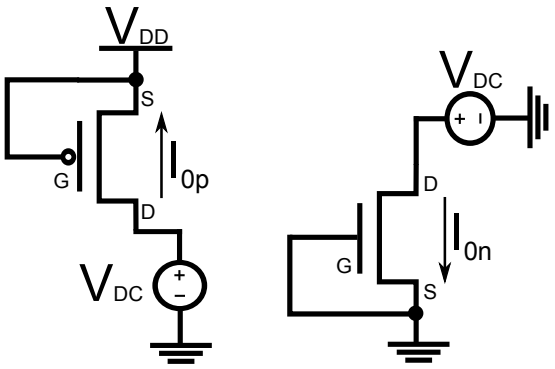
are then used to extract the prime parameters required in the sub-V_T energy mathematical model. In-house developed scripts are used to extract these parameter. The scripts have inputs from the reports generated by the Synthesis and PnR tools together with the leakage current information of the cells from the standard-cell library.

In addition to the prime parameter generated from the high level placed and routed netlist, transistor level simulation are performed to get I_0. The Figure 4.4(a), shows a simulation setup for both the PMOS and NMOS transistor. The average leakage current for the transistors are represented as I_{0p} and I_{0n}, respectively. Here, V_{DC} is the drain voltage that is swept from 0 to near V_T. The gate voltages are set to the supply voltage V_{DD} and ground Gnd for the PMOS and NMOS, respectively. Furthermore, there is no body bias for either of the transistors. Therefore, the transistors are off and the leakage current is only based on the drain voltage sweep. The overall average leakage for an inverter is given as

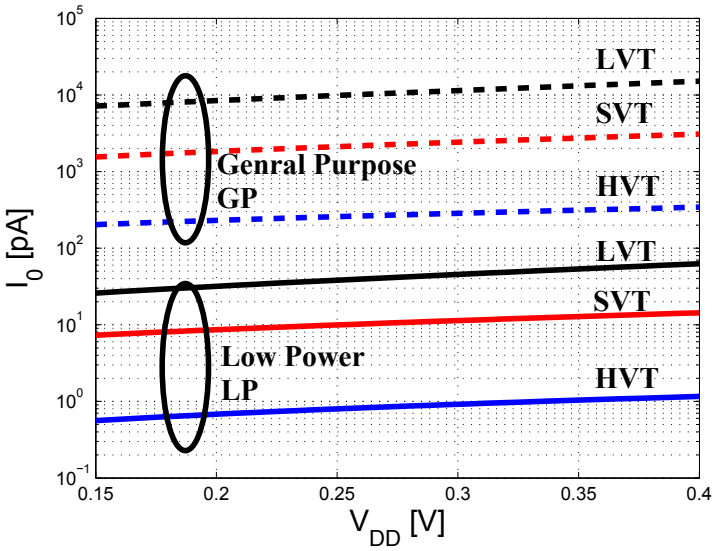
$$I_0 = \frac{|I_{0p}| + I_{0n}}{2}. \tag{4.27}$$

The current I_0, is an important parameter that is used to analyze frequency constrained architectures. In Figure 4.4(b), the current I_0 of an inverter generated with Spectre simulation is plotted versus the V_{DD}. The inverter in this simulation is modeled after the minimum sized standard-cell model, provided by the vendor. Furthermore, the inverter is simulated for both the low power (LP) and general purpose (GP) technology option. In each library the designer has the option to choose from three different thresholds of transistors, called high (HVT), standard (SVT), and low (LVT). As seen in the Figure 4.4(b), the leakage current reduces with the reduction of V_{DD}. The LP-HVT inverter setup has the lowest leakage profile and the GP-LVT has the highest current leakage profile. In the case of V_{DD} = 300 mV, the average leakage current drained for the inverter in LP-HVT setup is ∼1 pA. The GP-LVT based inverter consumes ∼10 nA, which is $\sim 10^4$ times higher compared to the former inverter. From (4.5), a higher I_0 will result in faster circuits in the sub-V_T domain. However, the overall energy dissipation profile will also increase. The inverters based on the LP library have lower leakage compared to the GP library based inverters.

The range of leakage current profile available in the 65 nm CMOS technology increases the design space for which the circuits have to be analyzed. The leakage difference within the LP library is also very high. In the case of V_{DD} = 300 mV, the average leakage current drained for an inverter, in the LP-HVT setup is ∼1 pA. The LP-LVT based inverter drains ∼50 pA, which is 50 times higher compared to the former inverter. Similarly, in the GP library, the case



(a)



(b)

Figure 4.4.: (a) Spectre simulation setup for I_0 for both PMOS and NMOS Devices (b) Average I_0 for a min. sized inverter constructed in various flavors of threshold options in 65 nm CMOS technology

where V_{DD} = 300 mV, the average leakage current drained for inverter in GP-HVT setup is ∼300 pA. The GP-LVT based inverter drains ∼10 nA. In the case of the GP-LVT inverter drains leakage current ∼33 times higher compared to the GP-HVT inverter. Therefore, there is a large leakage current difference within the library options. This results in a larger design space with respect to speed and energy dissipation.

4.3. RELIABILITY ANALYSIS

Beside the desire to operate at the energy-minimum, one of the limiting factors with respect to voltage scaling in the sub-V_T domain is the reliability of the circuit. Reliability issues arise mainly from within-die process variations and are aggravated in deep sub-micron technologies. Consequently, ensuring robust operation in the sub-V_T regime has been one of the most important concerns in the design of full-custom sub-V_T circuits.

In [38], the accuracy of the sub-V_T characterization model is verified by comparison with HSPICE transient simulations. It is found that the sub-V_T model predicts the energy dissipation with less than 3.8 % error for all considered ISCAS85 benchmark circuits. Furthermore, the accuracy of the model is validated by various measurements that are presented in [39], [35], [40]. It is shown that the measured energy is in the near vicinity of the simulated energy dissipation. The mean of the absolute modeling error is calculated to 5.2 %, with a standard deviation of 6.6 %. Moreover, it is also shown that the predicted maximum frequency at a given V_{DD} matches well with the measured maximum frequency of the implemented ASIC.

4.4. SUMMARY

This chapter introduces the energy model used for characterization of designs operated in sub-V_T domain operation. The presented model encompasses single V_T implementations and multi-V_T implementations. The energy modeling is based on the 65 nm CMOS standard cells provided by the technology vendor. The flow of the model is also described. It includes all the steps from high level circuit modeling, synthesis, and simulations. The energy model flow is achieved by the utilization of standard tools and in-house specialized scripts. It is described that extensive Spectre or HSPICE simulations of the designed circuit are not needed to get the initial estimations of the energy dissipation of a design operated in the sub-V_T domain. Although, basic leakage currents and sub-V_T current slope simulations for an inverter are needed, which are used in the energy model.

The leakage currents in the off-state of an inverter is presented for the six threshold options for standard cells available in the 65 nm CMOS technology. The variation in the leakage currents of these threshold options show that the speed and energy dissipation can vary by a margin, within the design space.

Part II

Architectural Analysis for Sub-V_{T} Operation

This part consists of a chapter that provide an analysis on the effect of switching activity within a circuit that is operated in the sub-V_{T} region. Furthermore, it includes chapters that discuss the effectiveness of techniques such pipelining and unfolding, when applied to circuits that are to be operated in the sub-V_{T} domain. This part includes material published in the following papers.

- O. Andersson, **S. Sherazi**, J. Rodrigues, "Impact of switching activity on the energy minimum voltage for 65 nm Sub-VV_{T} CMOS", *NORCHIP*, 2011-11-14.
- **S. Sherazi**, P. Nilsson, O. Akgun, H. Sjöland, J. Rodrigues, "Ultra low energy vs throughput design exploration of 65 nm sub-V_{T} CMOS digital filters", *NORCHIP*, 2010-11-15.

The material in this chapter originates from the article and is mutually used by the authors

5

Switching Activity Analysis on Energy Dissipation in Sub-V_{T}

Switching activity within a circuit plays an important role in defining the energy dissipation profile. As reliable statistics of switching activity is important for system integration. This chapter deals with the effects of the average switching activity in a design, in particular how the energy minimum voltage (EMV) moves. Extensive analyses of the energy optimization in the sub-V_{T} region is discussed in [17] [41] [42]. However, switching activity has, received little attention as a factor for energy dissipation and optimization of a design. The work in this chapter has been published in [43].

The remaining of the chapter is structured as follow. In Sec. 5.1 four architectures of growing complexity are used to study the effects of switching activity regarding the energy dissipation. In Sec. 5.2, the energy dissipation results based on the energy model is explained in chapter 4, attained from the four designs, are shown and compared, finally, concluded in Sec. 5.3.

5.1. TEST DESIGNS

In this section, the architectures used for this experiment are briefly discussed. Four architectures with increasing complexity and gate count are considered for this work. First, a multiplier shown in Figure 5.1(a), and second, an add-multiplier (ADD-MULT) architectures shown in Figure 5.1(b). Third, a larger add-multiplier (AMB) design is shown in Figure 5.1(c), and the fourth architecture is a multiplier-accumulator (MUL-ACC) design shown in Figure 5.1(d).

For all evaluated architectures the wordlength is set to 16 bits. Additionally, the multipliers in the architectures are chosen to be implemented as parallel

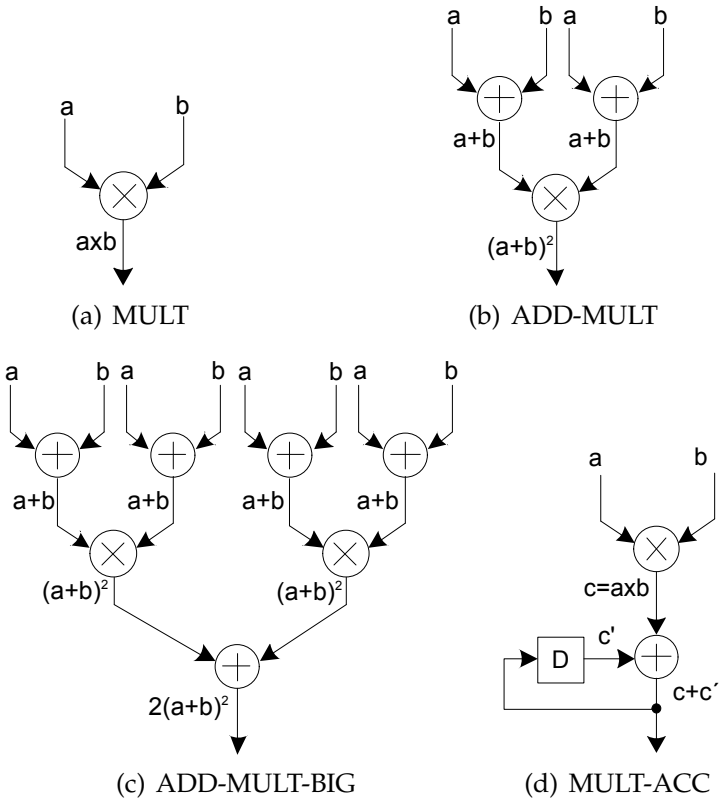


Figure 5.1.: Evaluated architectures.

Booth multipliers [44] [45]. One of the motivation with this selection of multiplier is the utilization of different standard cells (specifically 24 cells) in the synthesis of MULT, in order to get a wider range of analysis on the standard cells in the library.

Secondly, Booth's algorithm is vastly used to implement a parallel multiplier in the digital ASICs. One of the benefits of Booth's algorithm based multipliers is that they handle two's complement data with a high level of precision. In the case of two's complement data inputs, the architecture gives the correct product if the sign bit is included in the calculation. The Booth architecture uses partial products as the basic blocks and they are conventionally added, one at a time, in an array of adders. Then the result is attained in a final carry propagate add stage [44] [45].

Table 5.1.: Input Stimuli.

Test	Input a	Input b	ρ
1	Random (uniform)	Random (uniform)	0.001
2	Rect. pulse	Rect. pulse (identical)	1
3	$\sin xt$	$\sin(xt + \frac{\pi}{4})$	0.707

In the ADD-MULT design the Booth multiplier architecture and two ripple carry adders are synthesized. In this architectures, 19 different standard cells are used in the synthesis. A ripple carry adder (RCA) is one of the most straightforward implementation of an adder [44].

In the AMB architecture, the Booth multipliers and ripple-carry adders are synthesized. In this architectures, 21 different standard cells are used in the synthesis.

Finally, a multiplier-accumulator (MUL-ACC) is implemented. The purpose of this design is to observe the switching activity in a design with a feed-back loop. In this architecture 24 different standard cells are used in the synthesis.

The chosen architectures are seen as a representable collection for various mathematical operations implemented in digital ASICs. These operations are very often realized in digital signal processing (DSP) implementations that have pure combinatorial adders, multipliers, and sequential feed-back loops in their data paths.

5.2. SIMULATION RESULTS

A thorough investigation on the effect of switching activity (μ_e) is carried out by application of various input stimuli to the selected architectures. Moreover, further analysis was carried out by the use of a forced μ_e, to cover μ_e that was not achieved with the set of used input stimuli. Multiple types of input stimuli (random data, rectangular pulses and sinusoids) with different parameters are investigated. A selection of stimuli with different correlation coefficients (ρ) are presented in Table 5.1. These test cases are chosen as they cover typical input data, processed by these architectures in a larger design.

The designs are synthesized and simulated with back-annotated gate-level netlists and toggle information. The designs are recorded in Value Change Dump (VCD) files. The power simulations are carried out based on the toggle information [46]. The acquired data is used as input for the sub-V_T model, from which the EMV, switching activity, μ_e and other parameters are calcu-

Table 5.2.: Parameters for architectures.

Design	k_{cap}	k_{crit}	k_{leak}	Area [NAND2 eq.]
MULT	4523	263	2757	1798
ADD-MULT	5719	250	4646	2106
AMB	11916	287	9677	4380
MULT-ACC	5828	235	5221	2283

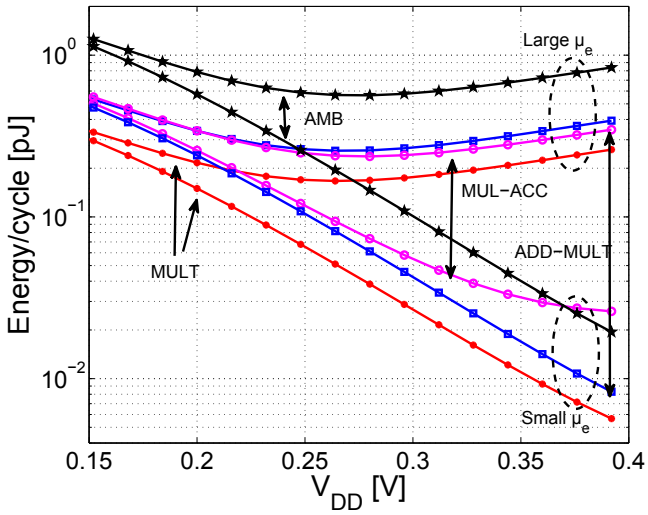
lated. The parameters, k_{cap}, k_{leak}, k_{crit}, and area are populated in Table 5.2. Where area is normalized to a two-input NAND gate. As seen in the table k_{cap} increases proportionally with the area of the designs. Similarly, k_{leak} increases proportionally with an increase in the area. The k_{crit} remains in a close range due to the chosen architectures for the implemented adders and multipliers. Furthermore, this also shows that the multiplier dominates the critical path, which is the limiting factor for speed.

5.2.1. SWITCHING ACTIVITY

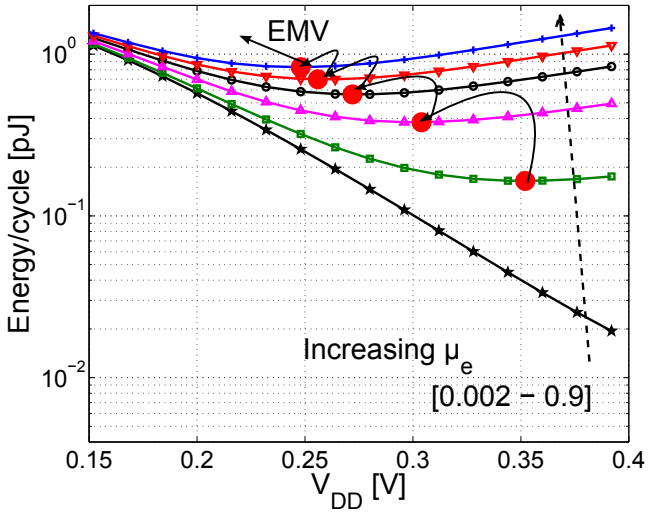
The three test cases that generate the most, least and moderate amount of switching are *Test 1*, *Test 2* and *Test 3*, respectively. The test case *Test 1*, *Test 2*, and *Test 3* correspond to input stimuli of random input, square wave and a sinusoidal wave, respectively. The energy curves for *Test 1* and *Test 2* w.r.t. supply voltage (V_{DD}) for the investigated designs are plotted in Figure 5.2(a). It is observed as expected that the largest architecture, AMB, dissipates most energy and the multiplier the least for *Test 1*. A similar trend appears for *Test 2* except for higher voltages, whereas MULT-ACC dissipates more energy than AMB. Although the switching activity for both the designs is very low, however, MULT-ACC has an order of magnitude higher μ_e compared to AMB.

The switching activity (μ_e) and EMV are populated in Table 5.3 for all three test cases. As expected μ_e is highest for random data. Sinusoids have lower μ_e and an increase of EMV. The μ_e generated by the rectangular pulse is generally below 0.03 and no EMV is found within the [0.25 0.4] V interval. Moreover, there is a slight increase of μ_e for the ADD-MULT architectures compared to MULT-ACC for *Test 1* and *Test 2*, as more nodes in this implementation switch.

Additionally, for AMB, μ_e and the EMV are very similar to the ADD-MULT design. Rectangular pulses generate very low μ_e even for this design, and the existance of EMV in the sub-V_T region was not observed. Many designs may experience low μ_e due to the nature of the design, as an example the memories based on standard cells (SCM) is described in [47]. Here, a low μ_e leads to



(a) *Test 1* (large μ_e) and *Test 2* (small μ_e).



(b) Sweep over μ_e for AMB.

Figure 5.2.: Sub-V_T energy profiles for different architectures w.r.t. μ_e.

Table 5.3.: Characteristics of architectures w.r.t test cases.

Design		Test 1	Test 2	Test 3
MULT	μ_e	0.424	0.003	0.317
	EMV [V]	0.264	> 0.4	0.280
ADD-MULT	μ_e	0.506	0.002	0.423
	EMV [V]	0.272	> 0.4	0.280
AMB	μ_e	0.517	0.002	0.434
	EMV [V]	0.262	> 0.4	0.280
MULT-ACC	μ_e	0.435	0.024	0.360
	EMV [V]	0.280	> 0.4	0.288

similar energy profile without an EMV within the [0.25 0.4] V interval.

In digital designs, feedback paths may exist and the effect on switching activity may vary due to the feedback. For a general analysis the multiply accumulate (MULT-ACC) design is used. A typical behaviour is observed for random data and sinusoids, where μ_e is similar to the former three architectures. However, for the case of rectangular pulses μ_e is higher by one order of magnitude. This is due to a register, which increases the switching in the design as there exists a clock path that switches periodically, and thereby, increases the overall μ_e.

5.2.2. ENERGY MINIMUM VOLTAGE

For an energy constrained design it is vital to identify the optimal conditions for operation where the energy dissipation is at its minimum, with the best possible throughput and with respect to the supply voltage V_{DD}. Thereby, a supply voltage that gives minimum energy with maximum throughput is defined as the energy minimum voltage (EMV).

With scaled supply voltages the total energy dissipation, as seen in (4.1), of a design is dominated by E_{leak} compared to E_{dyn}. However, as the voltage increases, E_{dyn} increases exponentially and becomes dominant. As seen in (4.2), μ_e is directly proportional to E_{dyn}. Therefore, μ_e and EMV are in close relation. A decrease of μ_e leads to a shift of EMV towards higher voltages, likewise an increase of μ_e shifts EMV towards lower voltages. This behaviour is visible for *Test 1* (a random input stimuli) and *Test 2* (a Rect. pulse input stimuli), with high and low μ_e, respectively.

In order to completely understand the relationship between μ_e and EMV, additional μ_e values where forced. Figure 5.2(b) shows energy dissipation of AMB (Add-Mult-Big test design) for various μ_e, where dots (•) indicate the

Table 5.4.: Characteristics of AMB for forced values of μ_e.

Forced cases	μ_e	f_{EMV} [kHz]	EMV [V]	σ (EMV)
1	0.1	98.7	0.352	+0.29
2	0.3	29.7	0.304	+0.12
3	0.7	9.0	0.256	−0.06
4	0.9	7.4	0.248	−0.09

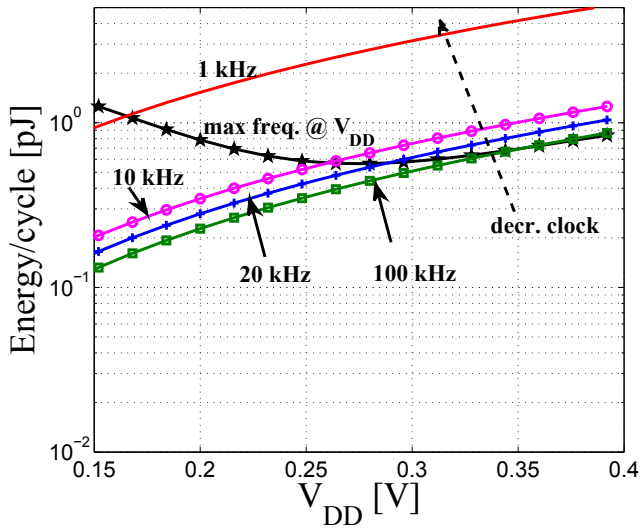
shift of EMV w.r.t. the change in μ_e. Table 5.4 shows the EMV w.r.t. to μ_e, the deviation (σ) of EMV from *Test 1* for AMB and the frequency at EMV (f_{EMV}). Secondly, the shift of EMV towards higher voltages is more pronounced with a decrease in μ_e. With an increased μ_e the EMV shifts towards lower voltages, however, this shift is not as pronounced as for the former situation. With a higher μ_e, E_{dyn} becomes the dominant factor at a lower voltage. That in turn increases the overall energy profile of the design, as seen in Figure 5.2(b). This means that an architecture with $\mu_e = 0.9$ has much higher energy dissipation than the same circuit with $\mu_e = 0.1$.

As an example for a $\mu_e = 0.1$, the EMV occurs at 0.352 V, dissipates 0.16 pJ per clock cycle and operates at f_{EMV} of 98.7 kHz. On the other hand with a $\mu_e = 0.9$ the EMV has shifted to 0.248 V, the energy dissipation has increased to 0.83 pJ, which approximates to an increase of five times, at f_{EMV} of 7.4 kHz. In this example EMV shifts 94 mV towards lower V_{DD} that results in an exponential decrease of frequency. Dramatic changes in the clock frequency are observed by slight variation in EMV.

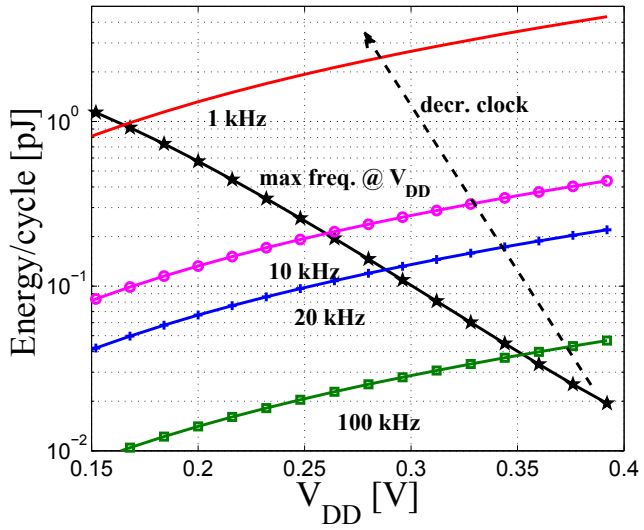
5.2.3. THROUGHPUT ANALYSIS

Circuits that have to be optimized for extreme low energy dissipation often have relaxed requirement on processing speed, which therefore are operated with a relaxed constraint on the clock frequency. Therefore, it is necessary to observe the behaviour of a design with respect to energy dissipation when operated at a fixed clock frequency. For this analysis AMB is chosen as it dissipates the most amount of energy. Four different clocks constraints are considered, 1 kHz (solid line), 10 kHz (○), 20 kHz (+) and 100 kHz (□), together with the maximum operational speed (⋆), for *Test 1* and *Test 2*, shown in Figure 5.3(a) and 5.3(b), respectively. Where, the max speed is governed by the critical path speed constrained by V_{DD}.

From Figure 5.3(a) it is observed that with an extremely low constraint clock frequency of 1 kHz the optimum energy point is achieved at a very low voltage of 160 mV. However, this voltage is far below minimum reliable supply



(a) *Test 1.*



(b) *Test 2.*

Figure 5.3.: Energy profile of AMB with constrained clock frequencies.

voltage of 250 mV [42]. Therefore, the circuit has to be operated at 250 mV or above for reliability issues. As an example, consider a supply voltage of 300 mV as a constraint on the system. For *Test 1* with this clock (1 kHz) and voltage (300 mV) constraint, the design dissipates 5 times more energy than the design operated with optimum clock speed. Similarly, in *Test 2*, where μ_e is extremely low, the energy loss is 30 times the optimum achievable scenario. The losses observed in this case arise due to the increased slack-time of the design. This increase in slack-time is due to the increased supply voltage. As the supply voltage increases the gates will operate much faster and therefore, the actual operation time decreases. Which leads to a longer idle time and therefore, the gates leak longer after the evaluation is performed. Increased leakage, i.e., E_{leak} in (4.1), leads to an overall higher energy profile.

By an increase in clock frequency by 10 to 20 times an operation point very close to the energy optimum point is achieved for both cases, near 300 mV. The loss in energy dissipation is worst for *Test 2* at a speed of 10 kHz, where the loss is three times the optimum energy dissipation. On the other hand by an increase of the frequency to 20 kHz the energy dissipation loss is reduced to less than 50 % for both tests. Furthermore, if the requirement on the clock frequency is 100 kHz for 300 mV the failure rate of the design would be very high. As seen in the figures, in order to operate at 100 kHz, a higher V_{DD} is required. Therefore, the choice of the operational frequency and operational supply voltages must be analyzed extensively before implementation of the design is carried out.

5.3. SUMMARY

The chapter focuses on how the energy dissipation of architectures vary w.r.t. the switching activity, μ_e. The simulation results based on the sub-V_T energy model, show that a higher μ_e in a design causes high energy dissipation that in turn moves the energy minimum voltage point (EMV) to lower voltages. Consecutively, for lower μ_e the overall energy dissipation decreases and the EMV shift to higher voltages. Therefore, the same design may have a different energy profile, due to different μ_e. Secondly, the shift of EMV towards higher voltages is more pronounced with a decrease in μ_e. With an increased μ_e the EMV shifts towards lower voltages. However, this shift is not as substantial as for the former situation. Thirdly, from the analysis of the simulation results it is observed that if the chosen designs is not operated at the maximum operable frequency for a given supply voltage V_{DD}, leads to loss in energy dissipation. However, by correct selection of the operational clock frequency the energy dissipation is reduced by orders of magnitude. Finally, the overall

analysis shows that it is crucial to have knowledge of input data w.r.t. μ_e. The knowledge of μ_e may lead to significant design considerations, i.e. supply voltage and throughput.

6

Efficiency of Pipelining in Sub-V_{T} Operation

Pipelining is considered as an effective technique to increase speed of a design. The increased speed can be traded-off by reduction in supply voltage to gain in low energy dissipation. This chapter discusses the effectiveness of pipelining when the circuits are subjected to ultra-low voltage situations. In particular how the energy dissipation changes w.r.t the number of pipeline stages with in an architecture.

The remaining of the chapter is structured as follow. In Sec. 6.1 four architectures of growing complexity are used to study the effects of the pipeline stages on the energy dissipation. In Sec. 6.2 the energy dissipation results based on the energy model explained in Chapter 4, attained from the two designs, are shown and compared. Finally, a summary is presented Sec. 6.3.

6.1. TEST DESIGNS

In this section, the architectures used for this experiment are briefly discussed. Two architectures of increasing complexity and gate count are considered. First, an Addition-Multiplication-Addition (AMA) is shown in Figure 6.1(a), and second, a Multiplication-Tree (MT) architectures shown in Figure 6.1(b). The inputs and output are bounded by the stage of flip-flops in both designs.

The AMA design has a 16-bit input wordlength. The 16-bit output from the first stage adder is supplied to 16-bit multipliers. The multiplier gives a 32-bit output, that is given to the 32-bit adders. The final result is of 32-bit wordlength. The inputs to the design and the outputs are all registered. The Figure 6.1(a), shows the pipelining applied to AMA architecture. The

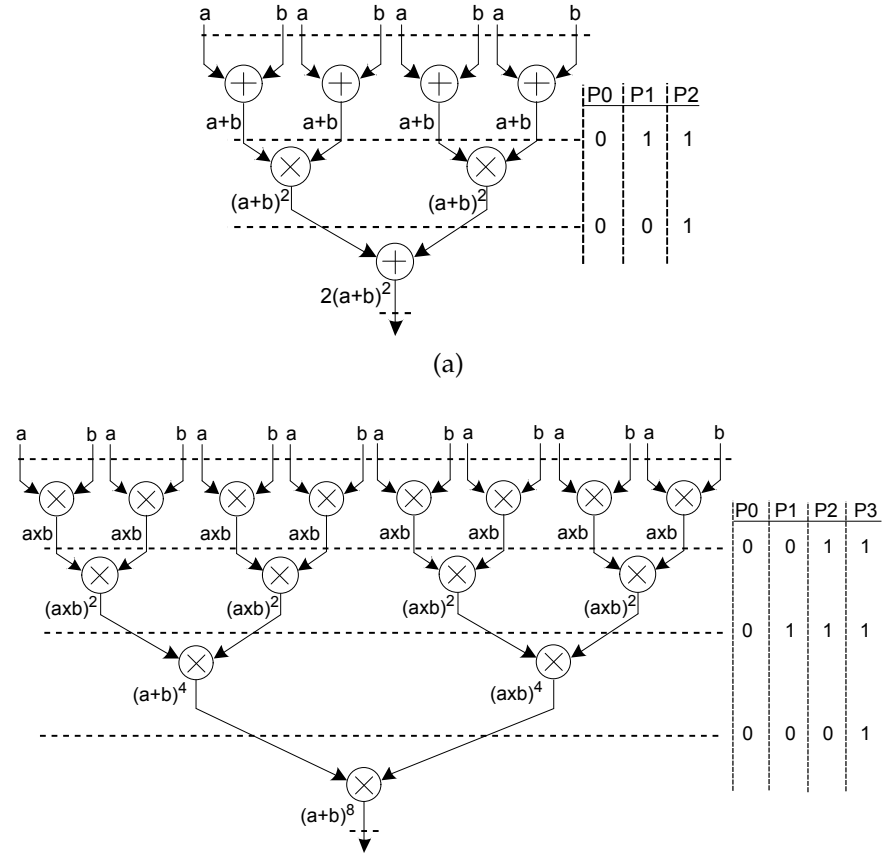


Figure 6.1.: Evaluated architectures. a) AMA. b) MT.

dotted line shows the stage where the pipeline is applied, the P0, P1, P2 show if the pipeline is present or not. P0 has zero pipeline stages, P1 has one and the location of the pipeline is indicated in the figure by '1'. P2 has two pipeline stages as indicated by by '1's in the figure corresponding to the pipeline stages. In addition to these three architectural options a forth architectural case is studied where a third pipeline stage is employed at the output of the multiplier. Re-timing is used to balance/pipeline the multiplier unit, this is named as P3.

In the second design of the multiplier tree (MT), the input to the first stage

of the multipliers is of 8-bit wordlength. The results of the first stage multiplier are of 16-bit wordlength that is then passed on to the second stage of multipliers. The results from these multipliers are of 32-bit wordlength, however, the outputs are truncated to 16-bits that are passed on to the third stage of multipliers. Similarly, the fourth stage multiplier also receives a 16-bit input and generates an output of 32-bit wordlength. The Figure 6.1(b), shows the pipeline stage P0 has zero pipeline stages, P1 has a single pipeline stage that is placed after second stage multipliers. P2 has two pipeline stages that are placed after the first stage multipliers and second stage multipliers. P3 has pipeline stage after the 1st, 2nd, and 3rd stage multiplier, therefore, in P3 all the multiplier outputs are pipelined. In addition to these four architectural cases, a fifth case is evaluated that has an additional pipeline stage is applied at the output of the multiplier and then re-timing is used to balance/pipeline the multiplier unit, which is called P4.

The chosen architectures are seen as a representable collection for various computational heavy mathematical operations implemented in digital ASICs. These mathematical operations are often seen in digital signal processing (DSP) implementations that have purely combinatorial adders and multipliers, e.g., FFTs etc.

6.1.1. SYNTHESIS

Both designs are synthesized based on the standard cell library provided by the vendor. The library used in the case study is a Low-power Standard-threshold (LP-SVT) library. In both AMA and MT designs the multiplier is implemented as a parallel booth multiplier architecture. The adders in AMA are based on ripple carry adder (RCA) structures. The synthesis of these designs is performed for minimum area and maximum speed. In case of the AMA design P3 and MT design P4, a re-time option is used to redistribute the registers within the design to balance and reduce the critical path.

6.2. SUB-V_T SIMULATION RESULTS

A thorough investigation on the effect of various stages of pipelining is carried out by application of random input stimuli to all the architectures. The designs are synthesized and simulated with back-annotated gate-level netlist and toggle information of the design is recorded in Value Change Dump (VCD) file. A power simulation is then carried out based on the toggle information. The sub-V_T energy model is applied to the designs with the extracted parameters as discussed in Chapter 4. The important parameters such as switching activity μ_e, k_{leak}, k_{cap}, and k_{crit} are given in the figures for both

Table 6.1.: Cells and Area for AMA.

pipe	No. of Cells			Area μm^2		
	Total	Combinational	Flip-Flop	Total	Combinational	Flip-Flop
P0	2748	2588	160	12914	11228	1685
P1	2778	2554	224	12252	9885	2367
P2	2839	2551	288	12470	9400	3070
P3	3200	2848	352	14569	10852	3716

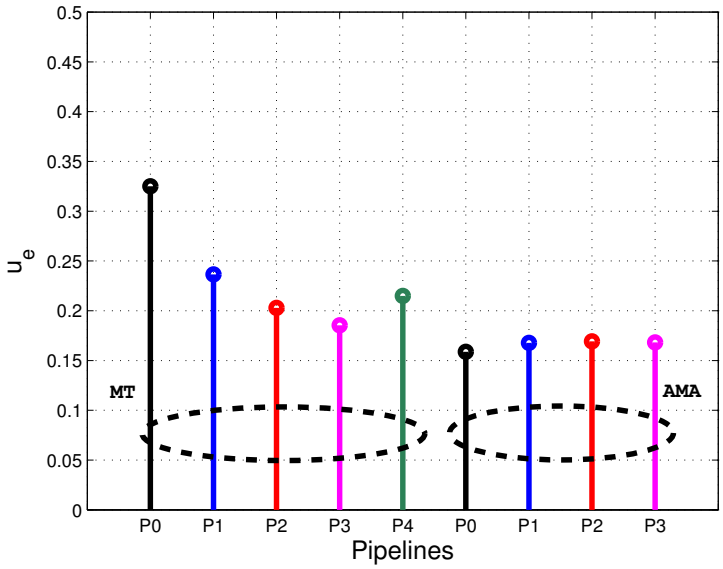


Figure 6.2.: Switching activity for the two designs corresponding to the number of pipeline stages

designs and their pipeline configurations.

6.2.1. ADDITION-MULTIPLICATION-ADDITION (AMA)

Four design options for the AMA architectures are analyzed for sub-V_T operation. Table 6.1, contains the breakdown of the cells within the designs, for the combinational gates and the flip-flops. As seen the number of flip-flops increase corresponding to the increase in the pipeline stage. The variation in the number of combinational is due to the variation in inverter and buffers

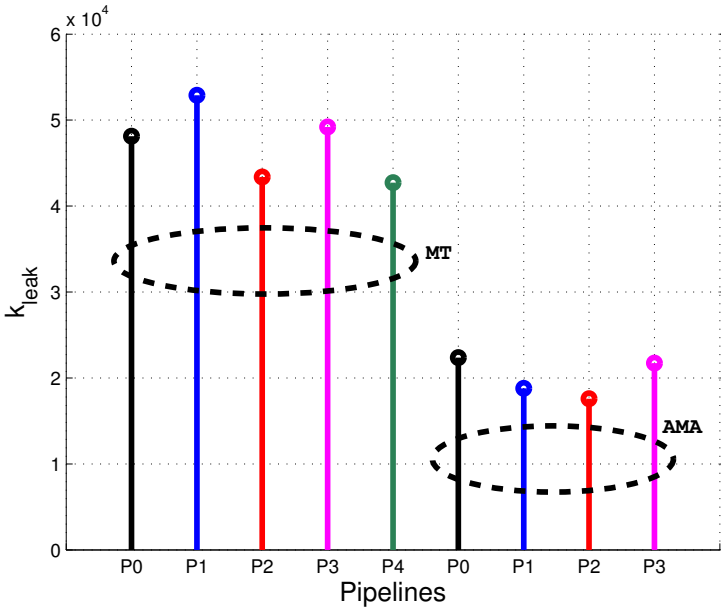


Figure 6.3.: k_{leak} for the two designs corresponding to the pipelines

cells needed to remove the hold violations. The area occupied by the flip-flop for P0 is only 13 % and in P1 with single stage pipeline the share of the flip-flop area increases to 19 %. In P2 and P3 cases, the area share for the flip-flop increases to 24 % and 25 %, respectively.

Figure 6.2, shows the switching activity (μ_e) within the four options of the AMA architecture, based on a random input stimuli. The results show that a there is not a big variation among the μ_e(s) of the designs. This indicates that the energy minimum point (EMV) for the designs shown do not vary by much, due to the change of μ_e. Figure 6.3, shows the leakage current within the design normalized to an inverter, k_{leak}. In this case the AMA design is synthesized without any pipeline stage exhibits higher leakage due to higher area. This is because the synthesizer tries to increase the speed of the design with the use of larger cells and it ends up with higher area. The area cost is reduced for both P1 and P2, even though more cells are used. However, in this case, the synthesizer is not tempted to use cells with large drive strengths and speed due to the pipeline stages. Similarly, the capacitance within these designs show the same characteristics as shown in Figure 6.4, that represents the k_{cap}. This indicates that the design with high leakage and capacitance

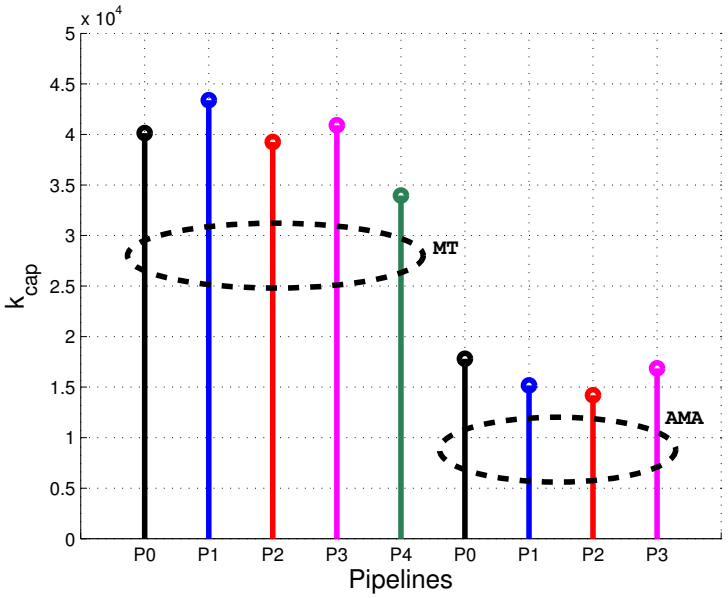


Figure 6.4.: k_{cap} for the two designs corresponding to the pipelines

will have higher energy dissipation profiles. The last parameter that effect the energy profile of a design is k_{crit}, which give critical path for the design, normalized to an inverter. Here, it is seen that with the pipeline stages in both P1 and P2 the critical path is shorten compared to P0. However, in the fourth case, with the addition of an additional pipeline stage the overall critical path is not reduced more than that of P2.

The energy dissipation w.r.t supply voltage (V_{DD}) for the pipeline options of AMA design are shown in Figure 6.6. Here, the AMA design with zero pipeline (P0) has the highest energy dissipation per cycle compared to any other pipelined design option. The P2 design exhibits the lowest energy dissipation profile. In this case, the low energy profile is due to the higher speed that is traded-off with the supply voltage reduction. This can be seen more clearly when the energy dissipation is plotted against the clock frequency as shown in Figure 6.7. As larger cells are used in the P0, the area of P0 is 3 % bigger than the P2 implementation, this contributes to higher energy dissipation in P0, especially at very low voltages. In the case of P3, the critical path is not reduced, however, there is an area penalty, that causes this design option to perform worse than P2 w.r.t energy dissipation due to high leakage.

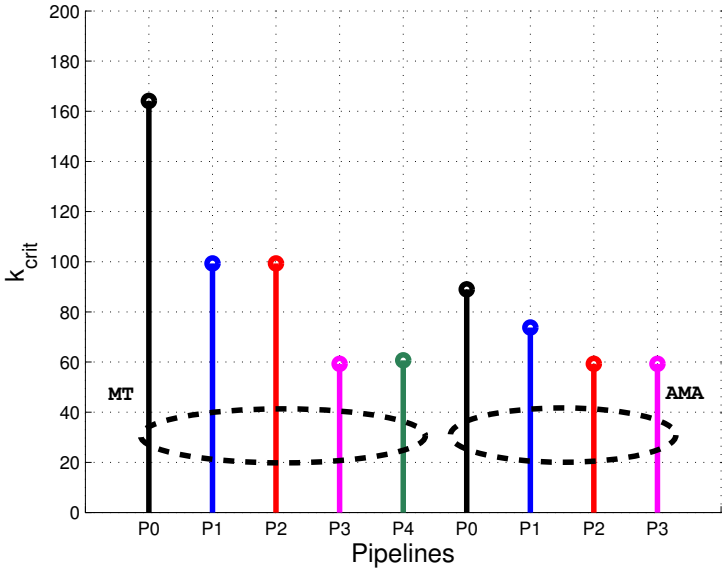


Figure 6.5.: k_{crit} for the two designs corresponding to the pipelines

At 300 mV P0 dissipates around 40 % more energy compared to P2, and at 400 mV the difference is of only 15 %. This shows that the benefits comes from V_{DD} scaling. In other words, a low logic depth allows reduced leakage at the expense of a larger dynamic energy. The dynamic energy is reduced by further V_{DD} reduction. Hence the combination of pipelining and V_{DD} reductions helps in an overall reduction of the energy per cycle for deigns operated in sub-V_{DD} [2] [30] [48].

6.2.2. MULTIPLICATION-TREE (MT)

Five design options for the MT architectures are analyzed for sub-V_T operation. Table 6.2, contains the breakdown of cells within the designs, for the combinational gates and the flip-flops used in the implementation. As seen the number of flip-flops increases corresponding to the increase in the pipeline stages. The variation in the number of combinational gates is again due to the variation in number of inverters and buffer cells used to remove the hold violations. The area occupied by the flip-flops for P0 is only 5 %, in P1 with single stage pipeline the share of the flip-flop area increases to 7 %. In P2, P3, and P4 cases, the area share for the flip-flop increases to 11 %, 12 %, and 19 %, respectively. This shows that the effect of pipelining should be more

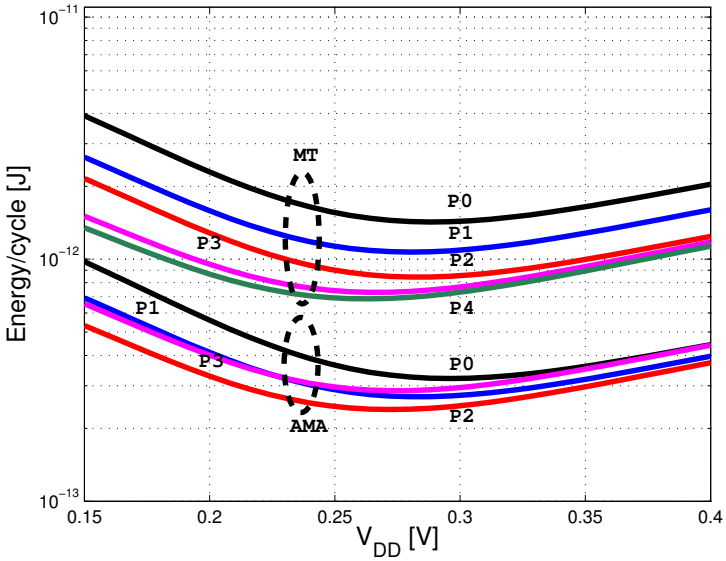


Figure 6.6.: Energy per cycle vs V_{DD} @max freq. for the two designs corresponding to the pipelines

Table 6.2.: Cells and Area for MT.

pipe	No. of Cells			Area μm^2		
	Total	Combinational	Flip-Flop	Total	Combinational	Flip-Flop
P0	6149	5989	160	30749	29050	1699
P1	6951	6727	224	33626	31258	2368
P2	7004	6652	352	33027	29335	3691
P3	6923	6539	384	33707	29678	4028
P4	5939	5427	512	28257	22824	5433

prominent in the cases with a higher flip-flop percentage.

Figure 6.2, shows the switching activity (μ_e) within the five options of the MT architecture, based on a random input stimuli. The results show that there is a slight variation among the μ_e(s) of the designs. This indicates that the energy minimum point (EMV) for the designs varies slightly due to the μ_e. Figure 6.3, shows the leakage current within the design normalized to an inverter, k_{leak}. In this case the MT design synthesized with one pipeline stage,

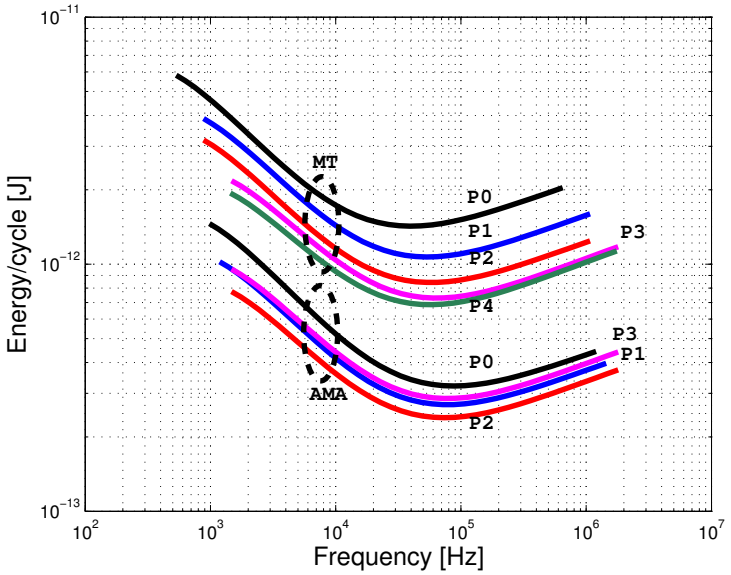


Figure 6.7.: Energy per cycle vs max Frequency for the two designs corresponding to the pipelines

represented as P1, exhibits higher leakage due to higher area, where the combinational gates have 93 % of the area share. This is because the synthesizer tries to increase the speed of the design with the use of larger combinational cells, that results in a higher leakage current. The area cost is reduced for P2, even though more cells are used in P2. However, in this case, the synthesizer does not use cells with large driving strengths and speed due to the two pipeline stages. In case of P3 the total area increases although the number of cells are less than in P2, this shows that the cell used in the synthesis are large with higher drive strength. This leads to higher current leakage as is seen in the Figure 6.3, that k_{leak} of P3 is higher compared to P2. In the fifth case where an additional pipeline is used that is redistributed to minimized the the critical path has lower area and also lowest cell count, which is why the k_{leak} is low.

Similarly, the capacitance with in these designs show the same characteristics as shown in Figure 6.4, that represents the k_{cap}. This indicates that the designs with high leakage and capacitance will have a higher energy dissipation profiles. The last parameter that effect the energy profile of a design is k_{crit}, which indicates the critical path for the design, normalized to an inverter. Here, it is seen that with the pipelines, in both P1 and P2, the critical path is

shorten compared to P0. In the case of P4 and P5, with the addition of an additional pipeline stage, the critical path is reduced compared to P1 and P2.

The energy dissipation, w.r.t supply voltage (V_{DD}), for the pipeline options of MT design are shown in Figure 6.6. Here, as expected the MT design with zero pipeline (P0) has the highest energy dissipation per cycle compared to any other pipelined design option. The P3 and P4 designs exhibit the lowest energy dissipation profile. Although, in the case of P4 the advantage of additional pipeline stages do not yield a higher gain. However, in these two cases the low energy profile is due to the higher speed that is traded-off with the supply voltage reduction. This can be seen more clearly when the energy dissipation is plotted against the clock frequency as shown in Figure 6.7. In the case of P0, all the k-parameter are high compared to the counterpart design option and the μ_e is also higher, this leads to a higher energy profile even though the area is relatively small compared to other counterparts. In case of P2, although the k_{crit} is the same as that of P1. However, because of the two pipeline stages the switching activity is lower. Furthermore, k_{leak} and k_{cap} for P2 is lower than the k- parameters of P1 indicating lower leakage and parasitics that leads to a lower energy dissipation profile.

At 300 mV the P0 dissipates around 88 % higher energy compared to P3, and at 400 mV the difference is only 53 %. This shows that the benefits comes from V_{DD} scaling. In other words a low logic depth allows reduce leakage at the expense of a larger dynamic energy. The dynamic energy is reduced by further V_{DD} reduction. Hence the combination of pipelining and V_{DD} reductions becomes very effective in overall reduction of the energy per cycle for deigns operated in sub-V_T.

6.2.3. DISCUSSION

The adoption of pipelined designs with reduced supply voltage V_{DD} increase the energy efficiency per operation. The pipelined architectures have reduced dynamic energy dissipation near EMV. However, care has to be taken when deciding the number of pipeline stages. As a rule of thumb pipeline stages less than three, results in efficient designs. As higher number of registers within a given design that do not reduce the critical path by some margin lead to higher energy dissipation. As shown in the case of AMA design, the addition of flip-flops in P3 resulted in higher energy dissipation. Similarly, in MT design when the addition pipeline stage was forced in P4, the benefits in energy dissipation were next to none.

As discussed in [2] one of the major drawbacks of heavy pipelined designs that are operated in sub-V_T domain suffer from high process variations. This has a strong impact on critical path and therefore a strong impact on

static energy dissipation. Therefore, joint employment of deep pipelining and ultra-low supply voltage may lead to significantly degrade robustness of the design. Furthermore, it may also add significant static energy dissipation overhead due to process variations. This means that in order to fully exploit this technique the designers have to reduce the process variations and keep them within acceptable limits through appropriate techniques. In [2] it is also shown that clocking schemes that allow time borrowing to average out delay variations among adjacent propagation paths may reduce critical time variations. Time borrowing allows the critical data-paths to inherently borrow time in the current cycle from the next cycle, therefore, the clock frequency can be increased for the design [49]. Time borrowing can be exploited with the use of latch based storage elements with a two phase clocking mechanism.

In the pipeline system the logic depth is kept low to avoid false switching. Therefore, if in static timing analysis hold violations are detected, the only mechanism to fix them is addition of buffers, that in turn increase the logic depth and increase the area. Therefore, the benefits of low energy dissipation will diminish. Therefore, flip-flop topologies that are intrinsically robust against hold variations must be employed [2].

6.3. SUMMARY

In this chapter it is shown that a reasonable number of pipeline stages together with supply voltage scaling have benefits with respect to energy dissipation. The simulation results based on the sub-V_T energy model, show that designs with high combinational logic gates; the pipeline stages reduce the switching activity μ_e, furthermore, there is reduction is leakage currents. All of these reductions result in lower energy dissipation in sub-V_T domain. In addition to these benefits it is also discussed that the benefits appear at low voltages, therefore, these designs are susceptible to process variations. Therefore, designers have to use robust flip-flop or resort to time-borrowing techniques in order to avoid functional failures.

7

Unfolded Architectures in Sub-V_{T}

Unfolding is a transformation technique used to increases the number of calculable iterations within a cycle for a given iterative design, thus increasing its speed. The factor by which a design is unfolded is called the unfolding factor [50]. Unfolding is considered an effective technique to increase the speed of a design. Inputs are concurrently applied to replicated blocks of the hardware so that concurrent output data is computed in a single clock cycle. This means that a higher throughput is achieved with the employment of this technique. The increased speed can be traded of by reduction in supply voltage to gain low energy dissipation. In other words the dynamic energy can be reduced. However, the overall static leakage will increase due to the area overhead. This chapter discusses the effectiveness of unfolding when the circuits are subjected to ultra low voltage scenarios. In particular how the energy dissipation changes w.r.t. unfolding factor for an architecture. The work in this chapter was published in [37] [51] [52].

As a case-study the effectiveness of unfolding technique is tested on the digital baseband part of a receiver system that is used in system reported in [53]. This receiver less than 1 mW and 1 ÎijW power consumption in active and standby mode constraints, respectively. Furthermore, the receiver is capable to handle data rates up to 250 kbits/s, and realization on a single chip with an area of 1 mm2 in 65 nm CMOS. A block diagram shows the receiver system in Figure 7.1, containing a RF front-end (2.5GHz), an analog-to-digital converter, a digital baseband for demodulation and control, and finally, a decoder that processes the received data packets. The first task of the digital baseband circuit is to re-sample data from 4 Msamples/s to 250 ksamples/s. A chain of decimation filters by 2 is applied for achieving the re-sample data

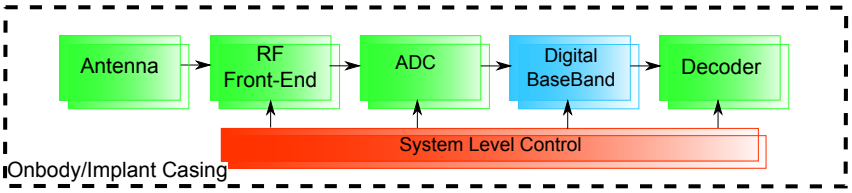


Figure 7.1.: Receiver system

rates. To achieve lower energy dissipation, supply voltage scaling techniques is rigorously applied, hence making the circuits run in the subthreshold (sub-V_T) domain [19]. Consequently, the circuits need to be optimized in terms of energy dissipation and throughput for sub-V_T operation.

The remaining of the chapter is structured as follows. In Sec. 7.1 a 12-bit architecture of a Half Band Digital (HBD) filter that is implemented as direct mapped and its various unfolded structures is discussed. In Sec. 7.2 the HBD filters energy dissipation results, based on the energy model explained in Chapter 4, are shown and discussed. Finally, a summary is presented Sec. 7.3.

7.1. HALF-BAND FILTER

Half-band filters are widely used in multi-rate signal processing applications when interpolating/decimating by a factor of two. Half-band filters are implemented efficiently in poly-phase form, because approximately half of its coefficients are equal to zero. Half-band filters are characteristics by, the maximum pass-band magnitude ripple σ_1, the maximum stop-band magnitude ripple σ_2 ripples, and the equidistant pass-band-edge f_p and stop-band-edge f_s frequencies from the half-band frequency $\pi/2$ [54]. Figure 7.2, shows the magnitude/gain response of a FIR half-band filters of an order of 60.

A half-band IIR filter can have fewer multipliers than the FIR filter for the same sharp cut-off specification. Elliptic IIR filters are the most efficient [54]. To overcome phase non-linearity one can use optimization to design an IIR filter with an approximately linear phase response [55] or apply the double filtering technique with the Powell and Chau modification for real-time processing [56] [57].

7.1.1. FILTER ARCHITECTURES

Minimum energy dissipation for a circuit operated at medium to high throughput puts stringent constraints on the design of the said circuit. Therefore, it is important to explore and analyze the architectures that best fulfill the re-

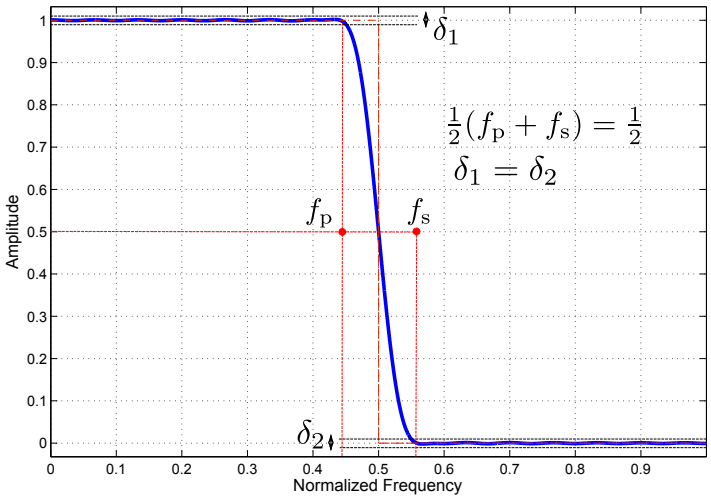


Figure 7.2.: Magnitude response of a FIR based Half Band Filter

quirements. This section presents half-band IIR Digital (HBD) filters and the architectural differences in the basic and unfolded versions.

A third order Wave Digital Filter (WDF) filter is used as the base of this methodology [14] [58] and has been presented in [52]. Figure 7.3, shows the architecture of the 3rd order filter. The filter is an example, which consists of one 1st order section to the right and one 2nd order section to the left of the input/output signals. Figure 7.3, shows that the architecture has 3 registers, shown by the R blocks, 3 multipliers, 10 adders, and a shift, are needed. The architecture in Figure 7.3 is described by [52],

$$y_i = \frac{1}{2}\left[k_i + a_0(k_i - x_i) - c_i - a_i(c_i - x_i)\right], \tag{7.1}$$

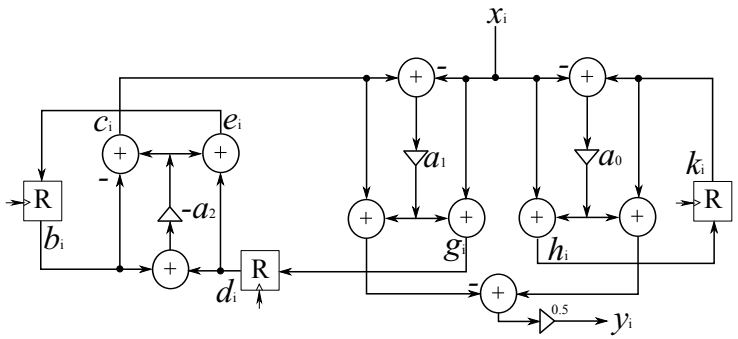


Figure 7.3.: Architecture of an IIR 3rd order Half-Band Filter

where the literal b_i, c_i, d_i, e_i, g_i and h_i are given as

$$\begin{aligned} c_i &= -b_i + a_2(b_i + d_i), \\ e_i &= d_i + a_2(b_i + d_i), \\ g_i &= x_i + a_1(c_i - x_i), \\ h_i &= x_i + a_0(k_i - x_i), \\ b_i &= e_{i-1}, \\ d_i &= g_{i-1}, \\ k_i &= h_{i-1}. \end{aligned} \tag{7.2}$$

In (7.1), the two left terms are the output from the 1st order section and the right two terms comes from the 2nd order section. WDF has a property of amplification, in this case the signal is scaled up by 2. Therefore, there is a multiplication factor, of 0.5 in (7.1), to compensate for that. The coefficients a_0, a_1, and a_2 attained by simulations and are specified as 0.37510, 0.57812510, and 0.32812510, respectively.

Hardware reduction is achieved by modifying the filter coefficients. Trivial coefficients like $a_0 = 0$, $a_1 = 0.5$, and $a_2 = 0$, are used for convenience. The behavior of this trivial filter is similar to the filter described by 3rd order HBD WDF. However, this new filter deviates from the cut-off frequency and stop-band attenuation characteristic of the larger filter as shown in Figure 7.4. Exchanging the coefficient values in (7.1) to the trivial coefficients yields (7.3)

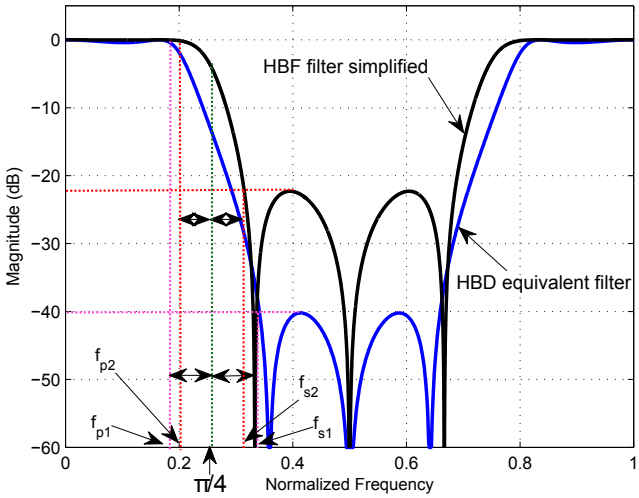


Figure 7.4.: Magnitude response of 3rd-order IIR Half-Band Filter and a simplified 3rd-order IIR Half-Band Filter.

and (7.4).

$$y_i = \frac{1}{2}\left[g_{i-2} + \frac{1}{2}(g_{i-2} + x_i) + x_{i-1})\right], \tag{7.3}$$

where

$$g_i = x_i - \frac{1}{2}(g_{i-2} + x_i). \tag{7.4}$$

This optimization yields a smaller filter that has 3 registers, 4 adders, and 2 shifts. However, the trivial filter can be further simplified without a change in the numerical result. Equations (7.3) and (7.4) can also be expressed as shown in (7.5) and (7.6). The architecture of this optimized filter is shown in Figure 7.5.

$$y_i = \frac{1}{2}\left[2g_{i-2} + g_i + x_{i-1}\right], \tag{7.5}$$

where

$$g_i = x_i - \frac{1}{2}(g_{i-2} + x_i) = -\frac{1}{2}g_{i-2} + \frac{1}{2}x_i. \tag{7.6}$$

The optimized third order filter structure is then evaluated for minimum energy dissipation, presented in [51]. The filter structure for the parallel implementation, see Figure 7.5, is a parallel third-order bi-reciprocal lattice wave

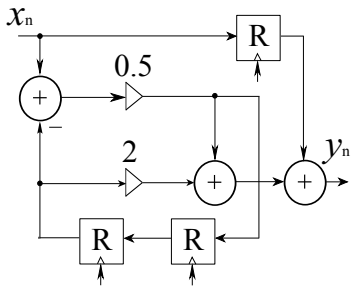


Figure 7.5.: Single equivalent HBD Filter. (*Org*)

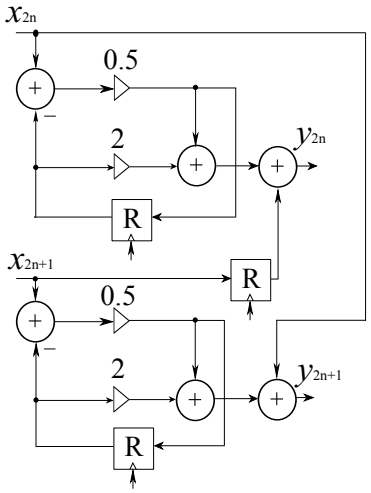


Figure 7.6.: Unfolded by 2 Architectures of the equivalent HBD filter. (*Uf-2*)

digital filter, [59], considered as suitable as decimator or interpolator, for sample rate conversions with a factor of two. The benefit of using this type of filter is that all filtering may be performed with low arithmetic complexity, therefore, yielding both low energy dissipation and low chip area [60]. The transfer function of the proposed filter is,

$$H(z) = \frac{1 + 2z^{-1} + 2z^{-2} + z^{-3}}{2 + z^{-2}}, \tag{7.7}$$

All the filter coefficients are 1/2 or 2, and thus implemented by simple shifting, thereby saving in area and energy dissipation. An initial analysis indicates that the required throughput would not be achieved by a single sample

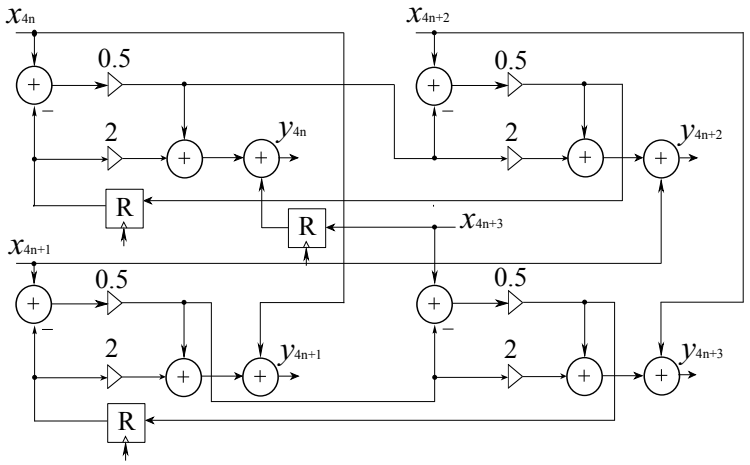


Figure 7.7.: Unfolded by 4 Architectures of the equivalent HBD filter. (*Uf-4*)

implementation of this filter. Therefore, unfolding was applied. Unfolding is a transformation technique that calculate *j* samples per clock cycle, where *j* is the unfolding factor. Unfolding has a property of preserving the number of delays in a Direct Form Graph (DFG) [50]. The basic HBD filter architecture was unfolded to get three more structures, i.e., unfolded by 2 (*Uf-2*), unfolded by 4 (*Uf-4*) and, unfolded by 8 (*Uf-8*). In all unfolded architectures the number of registers remains unchanged, whereas the adders scale proportional to the unfolding factor. Figure 7.6, shows the *Uf-2* version of the filter. Furthermore, the critical path of this circuit is equal to the original HBD filter structure. Figure 7.7 shows an architecture that is unfolded by a factor of 4. The number of adders has increased according to the unfolding factor. The critical path has increased, since two of the feedback paths do not contain a register. Similarly, Figure 7.8, shows the architecture of the *Uf-8* HBD. The adders have increased by a factor of 8, compared to the original HBD structure. The critical path increases, since six of the feedback paths do not contain any register. However, there are more samples processed per clock cycle in the unfolded structures, which wins with respect to throughput over a limited increase in the critical path [61].

7.1.2. HARDWARE MAPPING

All the cells used for implementation are from a low-leakage high-threshold (LL-HVT) standard cell library. Tight synthesis constraints were set to get

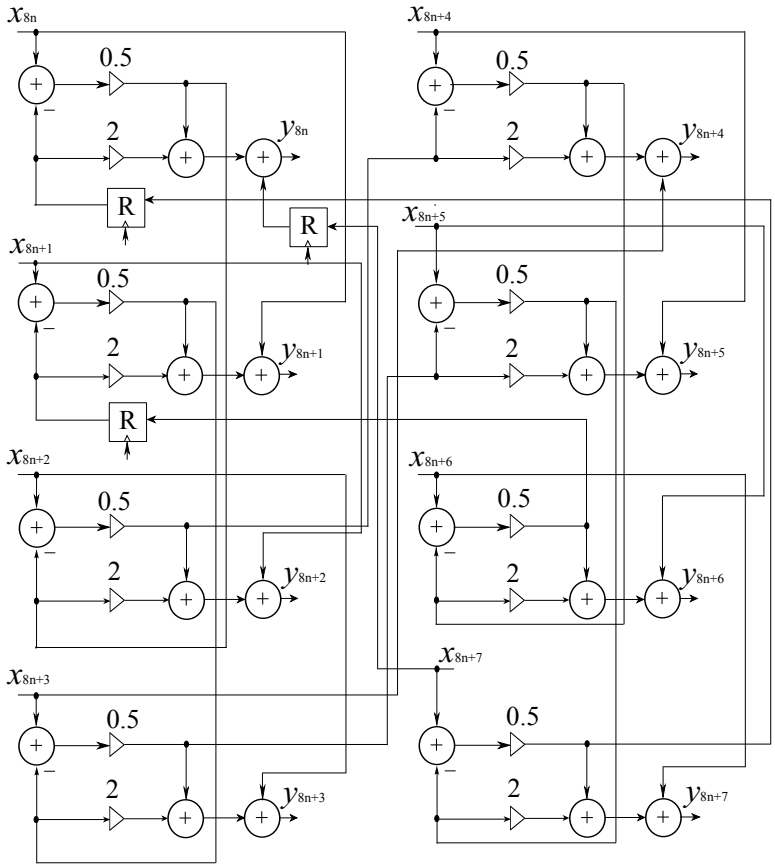


Figure 7.8.: Unfolded by 8 Architectures of the equivalent HBD filter. (*Uf-8*)

minimum area and a short critical path. The parameters for the energy model were retrieved by gate-level simulations with back annotated toggle and timing information, which includes glitches. The parameters obtained were applied to the energy model to characterize the designs in the sub-V_T domain.

7.2. SIMULATION RESULT

In this section the architectures of the filter are evaluated with respect to energy and throughput. The parameters required for the energy model discussed in Chapter 4, Section 4.1.1, are extracted during synthesis. The energy

Table 7.1.: Extracted Parameter for the Synthesized Implementations

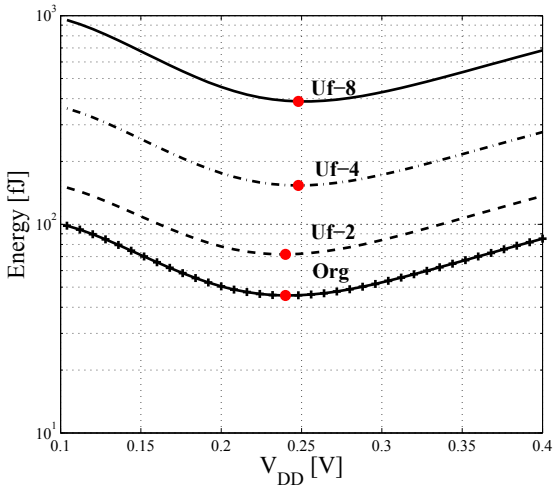
Arch.	$\mathbf{k}_{leak}$	$\mathbf{k}_{cap}$	$\mathbf{k}_{crit}$	μ_e	**Area**	$\mathbf{t}_p$**[nsec]**
Org	1113.6	835.4	127.4	0.727	1124	2.84
Uf-2	1695.5	1375.7	127.4	0.708	1836	2.84
Uf-4	3172.5	2797.9	164.2	0.703	3275	3.66
Uf-8	5924.5	5422.3	232.2	0.890	6170	5.22

Table 7.2.: Characterization of the Implementations at EMV

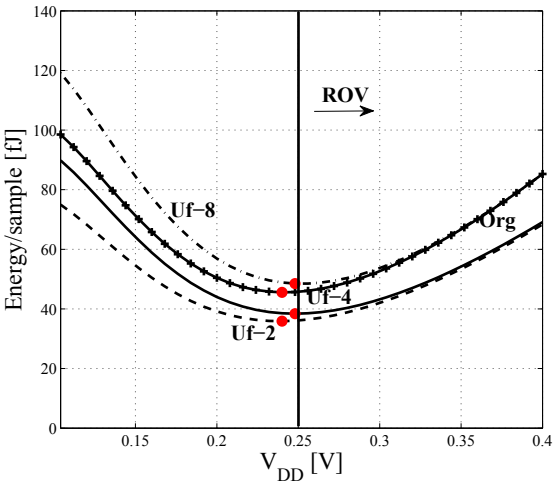
Arch.	**EMV** [mV]	**Freq.** [kHz]	**Throughput** [ksamples/s]	**E/Cyc** [fJ]	**E/smp** [fJ]
Org	241	23.6	23.6	45	45
Uf-2	238	23.6	47.2	71	35
Uf-4	247	22.0	88.0	150	38
Uf-8	251	15.4	123.4	380	48

simulations are presented in Table 7.1. The values for k_{leak} follow the area cost, indicating proportional leakage with respect to the area. The k parameters for the unfolded implementations are not proportional to the unfolding factor j since the number of internal registers remain unchanged from the basic implementation, although there is an increase in the number of input and output registers.

Energy dissipation is calculated under the assumption that the designs operate at critical path speed, which gives an Energy Minimum Voltage (EMV) point [39]. The threshold voltage for this low-power high-threshold (*LP-HVT*) device is around 630 mV. The designs' energy characteristics, over a scaled supply voltage V_{DD} per clock cycle is presented in Figure 7.9(a). The basic HBD filter implementation denoted by *(Org)* dissipates the minimum amount of energy per clock cycle when compared to the other three implementations. This due to the fact that the leakage for this circuit is less than that of the other circuits due to less area. The energy minima (per clock cycle) of $\sim$ 46 fJ for *Org* implementation is achieved $\sim$ 240 mV (indicated by the dot (•)), which is lower than EMV of any other architecture, which confirms that lesser area



(a)



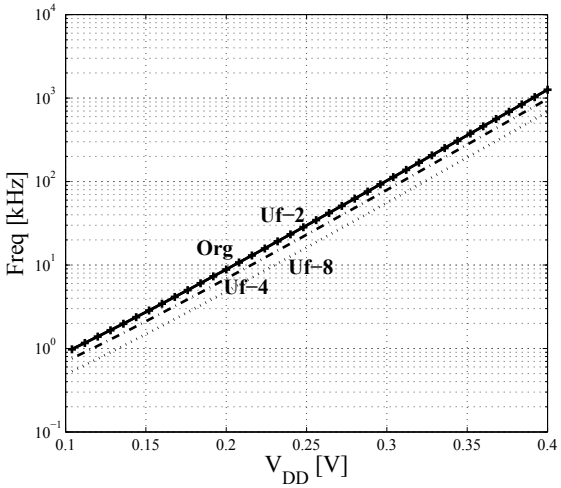
(b)

Figure 7.9.: Simulation Plots of HBD filter architectures, (a) Energy vs V_{DD} per clock cycle, (b) Energy vs V_{DD} per sample.

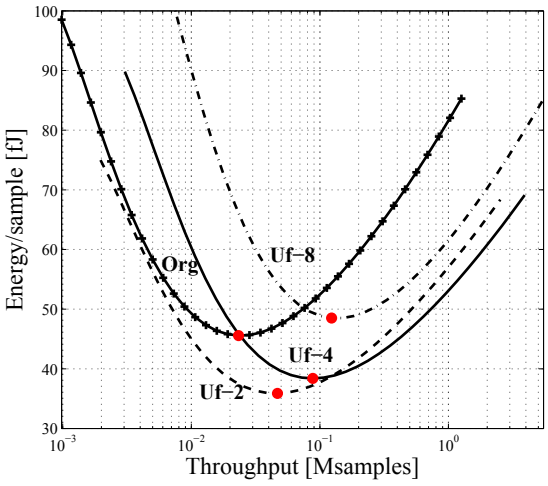
contributes to less energy per clock cycle. However, it is crucial to investigate the energy spent on the processing of each sample of data, and the apparent benefit of using *Org* structure is lost when the energy per sample is considered. Figure 7.9(b), shows the energy dissipation per sample for different structures, and the unfolded structures show higher energy efficiency compared to *Org*. The unfolded circuits perform twice, four and eight times as much operations per clock cycle, therefore, the overall energy per sample for these circuits is reduced when compared to a single sample implementation, however, with a limit. Figure 7.9(b), shows that the most efficient architecture is *Uf-2* as it dissipates ~ 36 fJ per sample which is ~ 45 % less than the energy dissipated by the *Org* structure. Here, it is observed that the *Uf-8* architecture is less energy efficient than *Org*, and is almost equal to *Org*, near the threshold voltages. The reason for this behavior is that the *Uf-8* has higher switching activity μ_e. The maximum frequency attainable with respect to V_{DD} is shown in Fig 7.10(a), the maximum frequency for both *Org* and *Uf-2*, is always higher than their counterparts due to a shorter critical path, and the *Uf-8* has the slowest maximum speed because of longer critical path, see Table 7.1. Fig 7.10(b), shows the energy dissipation of all the structures with respect to throughput.

Table 7.2, presents the characteristics of all the presented architectures at EMV. It also shows the maximum frequencies attainable, the corresponding throughputs, energy dissipated per clock cycle, as well as per sample. These simulations show that we benefit from unfolding technique, both in energy per sample and in throughput.

A chain of four HBD filters is needed to reduce the high frequency data at 4 Msamples/s from the ADC to the actual data rate of 250 ksamples/s. This decimation chain is to be used in a system present in [53]. The first HBD filter need to process the input data stream with the rate of 2 Msamples/s. This throughput requirement is only fulfilled by using *Uf-8* HBD near 390 mV, as shown in Table 7.3 and Figure 7.10(b). The throughput requirement of data with the rate of 1 Msamples/s for the second HBD is fulfilled by using any three of the unfolded structure, *Uf-8*, *Uf-4* and *Uf-2*. The throughput requirement of data with the rate of 500 ksamples/s for third HBD is fulfilled by all four structures as shown in Table 7.3 and Figure 7.10(b). The throughput requirement of data with the rate of 250 ksamples/s for last HBD is again fulfilled by all structures. In Figure 7.9(b), the *Uf-2* structure appears to be the most energy efficient circuit. However, when stringent throughput requirements are in-place, the *Uf-4* structure proves to be the best option as shown in Figure 7.10(b) and Table 7.3. This analysis shows that its crucial to identify the most suitable architectures for the given throughput and energy requirements. Furthermore, in [17] it is argued that low-leakage low-threshold cells



(a)



(b)

Figure 7.10.: Sub-V_T characterization of HBD filter architectures, (a) Frequency vs V_{DD}, (b) Energy vs Throughput

Table 7.3.: Performances of the Implementations at Required Throughputs

Throughput	Circuits	V_{DD} [mV]	E/Cyc [fJ]	E/smp [fJ]
2 Msamples/s	Uf-8	390	656	**82.2**
1 Msamples/s	Uf-8	368	586	73.3
	Uf-4	376	246	**61.5**
	Uf-2	400	136	68.3
500 ksamples/s	Uf-8	344	525	65.2
	Uf-4	352	226	**54.7**
	Uf-2	368	116	58.4
	Org	400	85.2	85.2
250 ksamples/s	Uf-8	300	434	55.0
	Uf-4	320	188	**47.0**
	Uf-2	344	126	51.8
	Org	368	72.9	72.9

are more beneficial at higher throughput rates in sub-V_T domain, which needs to be further investigated for these filter implementations.

In [19] it was shown in that the supply voltage of sub-V_T circuits may be reduced down to 50 mV. However, in practical terms at such low voltage values, functional failures frequently occur due to the process variations. It was found in [62] that the supply voltage value which realizes operation with less than a 0.001 failure rate for a 65 nm LP-HVT process is 250 mV and this value is taken as the minimum reliable operating voltage (ROV), indicated in the Figure 7.9(b) by a line at 250 mV. However, the operation at 250 mV still suffer from variations due to process variations as discussed in Chapter 3.

7.3. SUMMARY

In this chapter four HBD filter structures were developed and evaluated for minimum energy dissipation in the sub-V_T domain for a throughput constrained system. All architectures, i.e., the unfolded by 2, 4, 8 and the basic HBD filter, are implemented and simulated using 65 nm LP-HVT standard cells. The application of a sub-V_T energy model reveals that it is beneficial to use unfolded implementation to achieve low energy dissipation per sample at EMV, when compared to the energy dissipated by a basic simplified HBD

filter implementation. However, there is a limit to the unfolding factor, where the energy dissipation benefits start to diminish.

Part III

Sub-V_{T} Analysis on Threshold Options

This part consists of a chapter that provide an analysis on energy dissipation w.r.t. throughput with the utilization of various threshold options available in 65 nm CMOS, for a circuit that is operated in the sub-V_{T} region. Second, it includes a chapters that discusses silicon measurements of a design operated in sub-V_{T} domain. This part includes material published in the following papers.

- **S. Sherazi**, J. Rodrigues, O. Akgun, H. Sjöland, P. Nilsson, "Ultra low energy design exploration of digital decimation filters in 65 nm dual-V_{T} CMOS in the sub-V_{T} domain", *Microprocessors and Microsystems: Embedded Hardware Design (MICPRO), Elsevier*, vol.37/4-5, 2013.
- **S. Sherazi**, P. Nilsson, H. Sjöland, J. Rodrigues, "A 100-fJ/cycle sub-V_{T} decimation filter chain in 65 nm CMOS", *IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, 2012-12-09.
- **S. Sherazi**, P. Nilsson, O. Akgun, H. Sjöland, J. Rodrigues, "Design exploration of a 65 nm sub-V_{T} CMOS digital decimation filter chain", *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2011-05-16.
- H. Sjöland, J. B. Anderson, C. Bryant, R. Chandra, O. Edfors, A. Johansson, N. Seyed Mazloum, R. Meraji, P. Nilsson, D. Radjen, J. Rodrigues, **S. Sherazi**, V. Öwall, "A receiver architecture for devices in wireless body area networks", *Journal of Emerging and Selected Topics in Circuits and Systems*, Vol. 2, No. 1, pp. 82-95, 2012.

The material in this chapter originates from the article and is mutually used by the authors

8

Threshold Options within a Technology for Sub-V_T Domain Energy Dissipation

Threshold voltage of a device has a significant effect on the speed and energy dissipation of that device when operated in sub-V_T domain. In this chapter an analysis on energy dissipation of digital half-band filters presented in Chapter 7, considered for the subthreshold (sub-V_T) domain operations with throughput and supply voltage constraints are evaluated for implementations based on various threshold options. This analysis is performed in order to evaluate the design space within the frame of threshold voltage and moderate throughput. The work in this chapter has been published in [37] [40] and is filters are part of the digital baseband in a receiver that is used in system reported in [53].

A 12-bit simplified half-band filter is implemented along with various unfolded structures. The application target is to construct a decimation filter chain that is applied after a sigma delta ADC to re-sample data from 2 Msamples/s to 125 ksamples/s. Therefore, a chain of four decimation filters, that decimates by a factor of two at each stage, needs to be applied. The designs are synthesized in a 65 nm low-leakage CMOS technology with various threshold voltages.

A sub-V_T energy model presented in Chapter 4 is applied to characterize the designs in the sub-V_T domain. The Half-band Digital (HBD) filter that is implemented as direct mapped design and its various unfolded structures, is discussed in Chapter 7. The remaining of the chapter contains the Hardware Mapping information for three standalone threshold options in Sec. 8.1. In Sec. 8.1.1 the energy dissipation results based on the energy model explained in Chapter 4, attained from the HBD filters for these three threshold options are shown and discussed. Furthermore, the Hardware Mapping information

Table 8.1.: Extracted Parameter for the Synthesized Implementations

Arch.	Cells	k_{leak}	k_{cap}	k_{crit}	μ_e	Area
Org	HVT	1113	835	127	0.727	1124
	SVT	1181	803	112	0.723	1163
	LVT	1619	875	101	0.671	1208
Uf-2	HVT	1695	1375	127	0.708	1836
	SVT	1971	1553	116	0.620	1871
	LVT	4485	1434	105	0.720	2069
Uf-4	HVT	3172	2798	164	0.703	3275
	SVT	3199	2709	150	0.710	3390
	LVT	8524	2721	133	0.760	3750

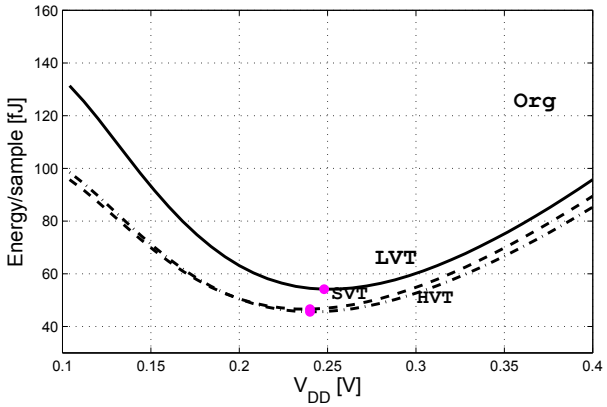


Figure 8.1.: Energy vs V_{DD} per sample simulation plots of simplified HBD filter (*Org*) architectures

for multi-mode threshold options are discussed in Sec. 8.2. A comparisons on the results is discussed in Sec. 8.3. Finally, a summary is presented in Sec. 8.4.

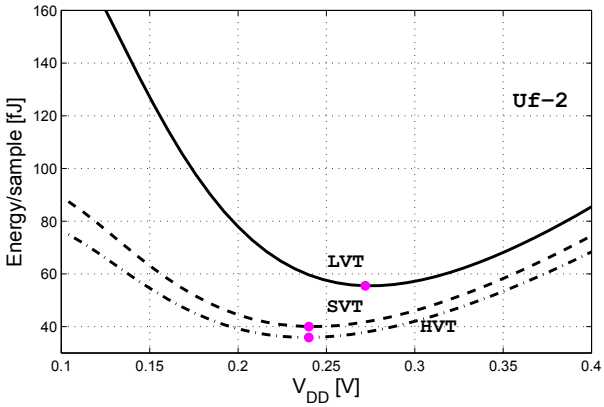


Figure 8.2.: Energy vs V_{DD} per sample simulation plots of unfolded by 2 HBD filter (*Uf-2*) architectures

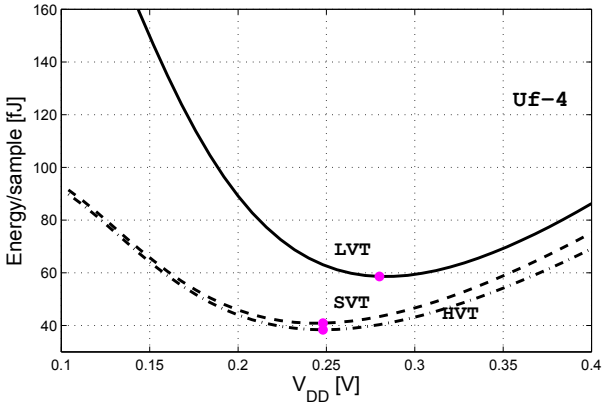


Figure 8.3.: Energy vs V_{DD} per sample simulation plots of unfolded by 4 HBD filter (*Uf-4*) architectures

8.1. HARDWARE MAPPING FOR THREE STANDALONE THRESHOLD OPTIONS

Each architecture was synthesized with Low Leakage (LL) libraries with different threshold voltage options. The first synthesis is performed using high-threshold (HVT) cells, second, using standard-threshold (SVT) cells and last, using low-threshold (LVT) cells. Tight synthesis constraints were set to achieve

Table 8.2.: Characterization of the Implementations at EMV

Arch.	**Cells**	**EMV** [mV]	**Freq.** [kHz]	**Throughput** [ksamples/s]	**E/smp** [fJ]
Org	HVT	241	23	23	45
	SVT	237	398	398	46
	LVT	251	3710	3710	55
Uf-2	HVT	238	23	46	**35**
	SVT	242	383	767	**40**
	LVT	271	6600	13200	**55**
Uf-4	HVT	247	22	88	38
	SVT	241	297	1180	41
	LVT	280	6500	26000	59

minimum area, minimum leakage, and a short critical path. The parameters for the energy model are retrieved by gate-level simulations with back annotated toggle and timing information, based on random input stimuli.

8.1.1. SIMULATION RESULT FOR THE THREE THRESHOLD OPTIONS

In this section the filter architectures are evaluated with respect to energy, throughput, and supply voltage constraints. The parameters required for the energy model [38] are presented in Table 8.1. The values for k_{leak} follow the area cost, indicating proportional leakage with respect to area for both the HVT and SVT implementations. However, k_{leak} values are higher for LVT implementation. A reason for this increase is high fanout buffers that have large current leakage, are used to increase the driving capacity and speed. Therefore, LVT implementation is faster than both HVT and SVT implementations, as expected.

The energy dissipation is calculated under the assumption that the designs operate at critical path speed. Minimizing the energy per clock cycle with respect to supply voltage gives the so called Energy Minimum Voltage (EMV) point [39]. The designs' energy characteristics, over a scaled supply voltage V_{DD} per sample are presented in Figures 8.1,8.2,8.3. Figure 8.1, shows the energy dissipated by gate-level implementations of *Org* for the various threshold voltage options, indicated as LVT, SVT, and HVT. Similarly, Figure 8.2 and Figure 8.3, show the energy dissipation curves for the *Uf-2* and *Uf-4* architectures. The dot on the curves indicates EMV for each architecture

Table 8.3.: Performances at Required Throughputs

Throughput samples/s	**Arch.**	**Vdd** [mV]	**E/smp[fJ]** @EMV	**E/smp [fJ]** @260 mV
2 M	***Uf-4***	**260**	**42**	**42**
	Uf-2	280	43	-
	Org	300	56	-
1 M	*Uf-4*	240	41	52
	Uf-2	**250**	**40**	**44**
	Org	280	51	-
500 K	*Uf-4*	200	46	72
	Uf-2	**220**	**41**	57
	Org	240	47	**53**
250 K	*Uf-4*	170	53	112
	Uf-2	**200**	**46**	81
	Org	220	47	**66**

and threshold voltage type. In all cases the minimum energy is achieved by HVT implementations. *Uf-2* appears to be the architecture that dissipates least energy per sample.

Table 8.2, presents the EMVs of each gate-level implementation, the maximum clock frequency at EMV, the corresponding throughput in samples per second, and the energy dissipated per sample. The *Uf-2* architectures dissipates least energy per sample at EMV. The simulations show that the LVT implementation is able to operate at a much higher frequency at EMV compared to their counterparts. The reason for this behavior is higher currents in the cells, both drive currents and leakage currents. Increased leakage, and drive current, pushes frequency higher to reduce the energy per cycle. Similarly, the SVT and HVT implementations have frequencies corresponding to their cell currents. The simulations show that the maximum clock frequency increases exponentially with increasing supply voltage i.e., the current increases exponentially in the cells, which leads to a further analysis on energy dissipation versus throughput.

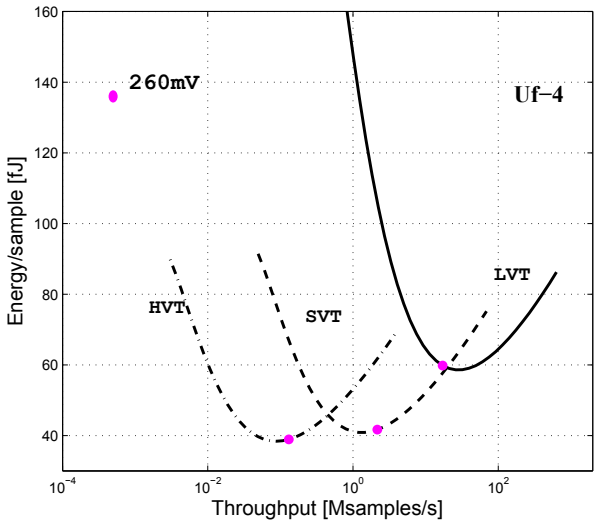


Figure 8.4.: Energy vs Throughput simulation plots of unfolded by 4 HBD filter (*Uf-4*) architectures

THROUGHPUT CONSTRAINTS

Figure 8.4, shows the energy vs throughput plot of the *Uf-4* and it is shown that SVT implementation is the most suitable choice of implementation for a throughput requirement within the range of 2 to 20 Msamples/s. The Figs. 8.5 and 8.6, show the energy dissipation vs throughput curves for the *Uf-2* and *Org* architectures. The SVT implementation of the *Uf-2* architecture is suitable for the throughput range of 250 ksamples/s to 2 Msamples/s.

The throughput constraints for the system are of 2 and 1 Msamples/s for the first two decimation filters and for the last two 500 and 250 ksamples/s. These requirements are fulfilled with least energy dissipation by different architectures using SVT cells. Therefore, further analysis is based on SVT implementations only. Table 8.3, presents the energy dissipation per sample for the required throughputs at corresponding supply voltages for different architectures for SVT implementations. The first filter with a throughput requirement of 2 Msamples/s is fulfilled by an *Uf-4* filter architecture as the most suitable option. Whereas, second, third and forth filters with throughput requirements of 1 Msamples/s, 500 and 250 ksamples/s are best achieved by the *Uf-2* filter architecture, the optimal values are shown in the bold in Table 8.3.

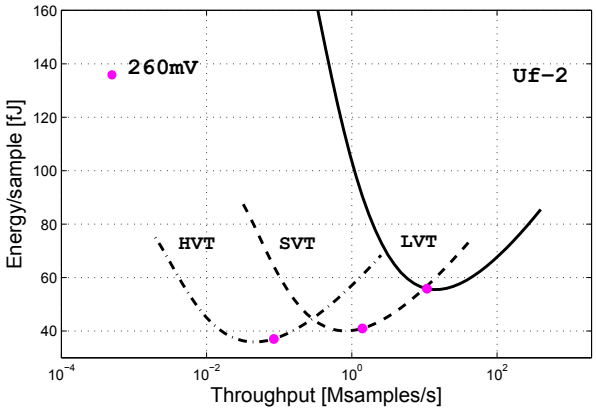


Figure 8.5.: Energy vs Throughput simulation plots of unfolded by 2 HBD filter (*Uf-2*) architectures

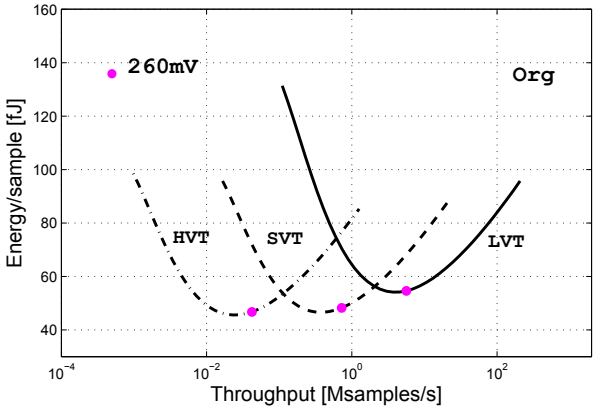


Figure 8.6.: Energy vs Throughput simulation plots of simplified HBD filter (*Org*) architectures

SUPPLY VOLTAGE AND THROUGHPUT CONSTRAINTS

In [62], it is found that the supply voltage value, which realizes operation with less than 0.001 failure rate for a 65 nm process is 250 mV and this value is taken as the minimum reliable operating voltage (ROV). The simulations show that the required throughput for the first and second decimation filters are fulfilled using *Uf-4* and *Uf-2* at 260 mV, and 250 mV, respectively. Having

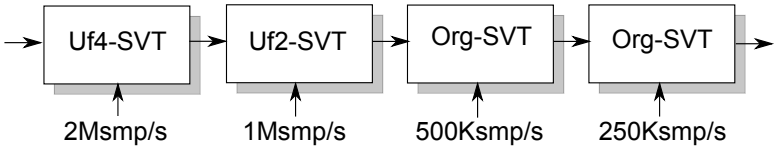


Figure 8.7.: Suitable Filter Chain.

multiple power domains increases the cost with respect to area and energy dissipation, which, is not desired. A single supply voltage of 260 mV is introduced as another constraint on the system. The selection of this voltage is based on the analysis that the first filter with a higher throughput constraint is fulfilled by using *Uf-4* at 260 mV. Therefore, 260 mV is selected as a supply voltage constraint and all the filters will operate at 260 mV. The assumption that the data is provided to the filter at critical path speed is not valid anymore. Therefore, the equation (4.2) for clock constrained systems is used to find the energy dissipation [62]. The last column in Table 8.3, shows energy dissipation per sample at 260 mV, for the three architectures at the required throughputs. Using a single power domain will have an impact on the criteria of selection of the suitable filter structures that are least energy dissipating.

The energy dissipation of the first filter remains unchanged, as *Uf-4* is operated at critical path speed. The energy dissipation of second filter increases, as the implementation is clocked slower than the critical path delay and therefore, there is an increase in leakage energy. However, *Uf-2* is still the most suitable filter architecture at these throughput and supply voltage constraints. The most suitable architecture for throughput requirements of both 500 and 250 ksamples/s is *Org*, shown in Table 8.3. The *Org* filter has the least area, therefore, once the implementations are not operating with critical path speed, has an advantage of dissipating less energy because of lesser leakage currents. Hence, with all the requirements in place, all filters in the chain will have SVT implementations, with the first filter being *Uf-4*, the second *Uf-2*, and the last two filters will be the *Org* filter architecture, as shown in Figure 8.7. The total energy dissipation per output sample for the filter chain is 205 fJ.

8.2. HARDWARE IMPLEMENTATION AND SYNTHESIS FOR MULTI-THRESHOLD OPTIONS

Each architecture is synthesized with Low Leakage (LL) libraries with different threshold voltage options. The synthesis is performed using solely HVT

and SVT cells, as well as HVT and SVT mixed, represented as (H+S). Furthermore, LVT synthesis was conducted in order to get an analysis over the entire design space and is represented as (H+L). Tight synthesis constraints are set to achieve minimum area, minimum leakage, and a short critical path. The parameters for the energy model are retrieved by gate-level simulations with back-annotated toggle and timing information, based on random input stimuli.

For multi-V_T synthesis, first the designs were synthesized with only HVT cells. Afterwards, the SVT cell library is instantiated, and timing constraints were tightened. A new synthesis was performed to get a multi-V_T implementation of H+S. As an illustrative example, lets consider the case of the *Org* filter. This filter is synthesized with HVT cells that results in an implementation with 196 cells. The critical path contains 22 HVT cells and has a delay of 2.8 ns at nominal V_{DD}. With the SVT library instantiated and constraints tightened, a new synthesis is performed. This results in a multi-V_T implementation that contains a total of 132 HVT, and 55 SVT cells. The leakage current contributed by SVT cells corresponds to the 84% of the total leakage current of the circuit. The critical path contains 10 HVT, and 14 SVT cells. The delay of the critical path is reduced to 1.5 ns at nominal V_{DD}. The effects of the characteristics of the cells are modeled based on the energy model described in Chapter 4, Section 4.1.2 and the simulation results are presented in Sec 8.3 for all the architectures. Similar experiments are also performed by the synthesis of HVT cells together with LVT cells.

8.3. SIMULATION RESULTS FOR THE MULTI-THRESHOLD OPTIONS

In this section the architectures are evaluated with respect to energy versus V_{DD}, energy versus throughput, and for required throughput, as well as energy at a fixed V_{DD}. The parameters required for the energy model [42], are extracted during synthesis and power simulations, as discussed in 4.1.1, and presented in Table 8.4. The values for k_{leak} follow the area cost, indicating proportional leakage with respect to area for both HVT and SVT implementations.

As shown in [40] SVT implementations have higher leakage compared to HVT, and therefore a higher k_{leak} factor. A reason for this increase is a larger leakage current, which is used to increase the driving capacity and speed. Therefore, the SVT implementation is faster than both HVT implementations, as expected. The characteristics of the HVT, and SVT cells lead to the idea of a multi-V_T implementation. In the multi-V_T implementation, HVT cells are chosen to get low leakage currents and SVT cells are chosen in the critical paths

Table 8.4.: Extracted Parameter for the Synthesized Implementations

Arch.	Cells	k_{leak}	k_{cap}	k_{crit}	μ_e	Area
Org	HVT	1113	835	127	0.727	1124
	SVT	1181	803	112	0.723	1163
	LVT	1619	875	101	0.671	1208
	HVT+SVT	1197[1]	829[2]	81[3]	0.750	1022
	HVT+LVT	1303[1]	766[2]	85[3]	0.980	1051
Uf-2	HVT	1695	1375	127	0.708	1836
	SVT	1971	1553	116	0.620	1871
	LVT	4485	1434	105	0.720	2069
	HVT+SVT	2531[1]	2476[2]	68[3]	0.600	1851
	HVT+LVT	2747[1]	2446[2]	66[3]	0.770	2504
Uf-4	HVT	3172	2798	164	0.703	3275
	SVT	3199	2709	150	0.710	3390
	LVT	8524	2721	133	0.760	3750
	HVT+SVT	4296[1]	4015[2]	89[3]	0.650	3110
	HVT+LVT	4655[1]	3925[2]	90[3]	1.100	4123
Uf-8	HVT	5924	5422	232	0.890	6170
	SVT	5990	5250	217	0.900	6385
	LVT	9660	5780	196	0.810	6698
	HVT+SVT	6902[1]	6350[2]	129[3]	0.890	6565
	HVT+LVT	7370[1]	6202[2]	136[3]	0.881	6684

[1] calculated for k'_{leak}, [2] calculated for k'_{cap}, [3] calculated for k'_{crit}

to get higher speed. However, the induction of these cells increases the overall leakage. Therefore, to find out if multi-V_T designs have any advantage over circuits with only one type of threshold cells, a multi-V_T analysis is important. As the LVT cells have very high leakage, which is not particularly suitable for low energy requirements, all simulations results for circuits synthesized with LVT cells are not included in the main discussion.

The k parameters for the unfolded implementations are not proportional to the unfolding factor j since the number of internal registers remain unchanged from the basic implementation, although there is an increase in the number of

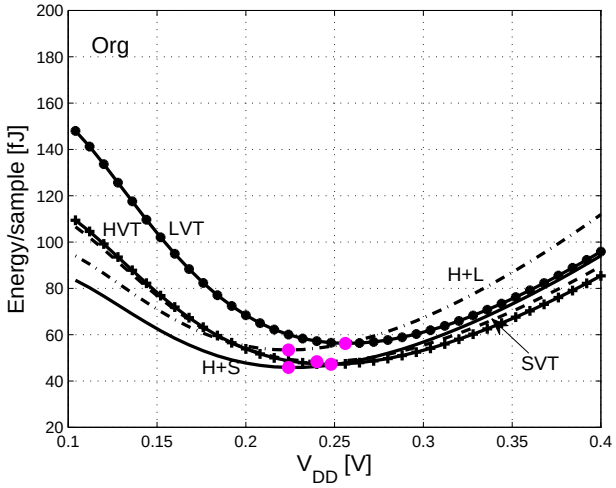


Figure 8.8.: Energy vs V_{DD} per sample simulation plots of simplified HBD filter (*Org*) architectures

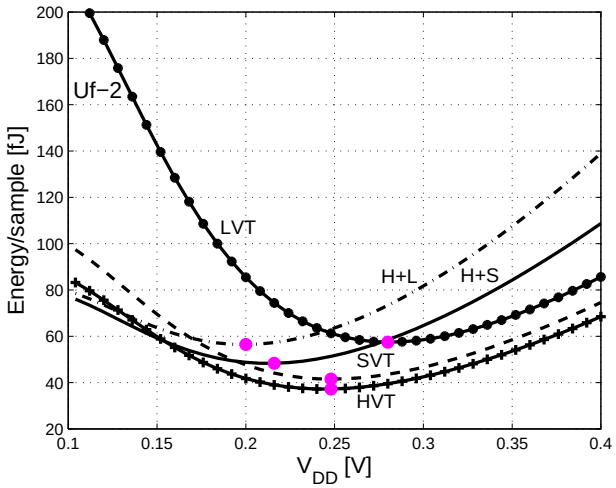


Figure 8.9.: Energy vs V_{DD} per sample simulation plots of unfolded by 2 HBD filter (*Uf-2*) architectures

input and output registers. However, the number of registers with reference

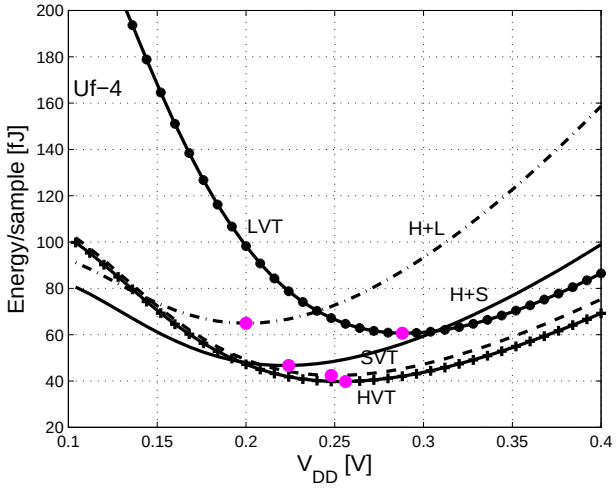


Figure 8.10.: Energy vs V_{DD} per sample simulation plots of unfolded by 4 HBD filter (*Uf-4*) architectures

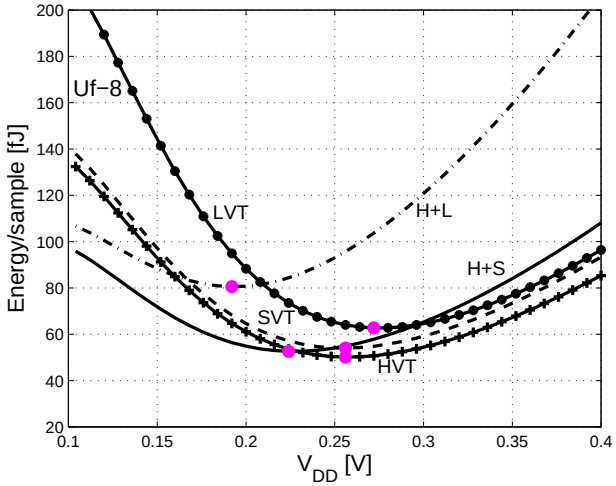


Figure 8.11.: Energy vs V_{DD} per sample simulation plots of unfolded by 8 HBD filter (*Uf-8*) architectures

to operation per sample remain unchanged.

Table 8.5.: Ratios for the H+S Synthesized Implementations

Arch.	$L_{r,1}$	$L_{r,2}$	$C_{r,1}$	$C_{r,2}$	$TL_{0,1}$	$TL_{0,1}$
Org	0.14	0.84	0.57	0.43	0.10	0.90
Uf-2	0.18	0.82	0.70	0.30	0.16	0.83
Uf-4	0.17	0.83	0.65	0.35	0.14	0.86
Uf-8	0.13	0.86	0.58	0.42	0.03	0.97

At extremely low V_{DD} the circuits are very slow, and therefore the overall leakage current increases per operation. As V_{DD} is increased, the static energy dissipation decreases and the proportion of switching energy increases. This phenomena leads to minimizing the energy per operation with respect to the V_{DD}. That gives the so called *Energy Minimum Voltage* (EMV) point [39]. The threshold voltage for these 65 nm transistors is around 450 mV for LL-SVT and around 500 mV for LL-HVT. The designs' energy characteristics, over a scaled V_{DD} per sample are presented in Figure 8.8. The energy dissipation is calculated under the assumption that the designs operate at critical path speed. Figure 8.8, shows the energy dissipated per output sample by gate-level implementations of *Org* for H+L, H+S, LVT, SVT, and HVT implementations. Similarly, Figure 8.9, Figure 8.10, and Figure 8.11, show the energy dissipation curves for the *Uf-2*, *Uf-4*, and *Uf-8* architectures. The minimum point on the curves indicate EMV for each architecture and threshold voltage type. Secondly, in [42], it is found that the supply voltage that realizes operation with less than 0.001 failure rate for a 65 nm process is 250 mV for HVT cells. No failure rates were observed when HVT cells are operated at higher supply voltages. The failure rates for SVT cells are lower at 250 mV compared to HVT cells. However, in this study 250 mV value is taken as the minimum reliable operating voltage (ROV). It is vital to know ROV, as if the EMV is observed below ROV, other optimization options need to be considered.

In most of the cases the implementations with the HVT cells gives the EMV point. The energy vs voltage figures show that H+S combination does not give any major advantage and in most of the case H+S under performs compared to single threshold implementations. One of the reasons for such a behaviour is a speed miss-match among cells. This miss-match leads to false transitions that increase the dynamic energy dissipation. At lower voltages the difference of speed between the two selected devices is not significant, therefore a small advantage is observed. However, the miss-match of speed increases with the increase in supply voltage, therefore, high switching activity increases the overall energy dissipation and the H+S synthesis loses.

Table 8.6.: Ratios for the H+L Synthesized Implementations

Arch.	$L_{r,1}$	$L_{r,2}$	$C_{r,1}$	$C_{r,2}$	$TL_{0,1}$	$TL_{0,1}$
Org	0.03	0.97	0.80	0.20	0.02	0.98
Uf-2	0.03	0.97	0.75	0.25	0.02	0.98
Uf-4	0.03	0.97	0.72	0.28	0.05	0.95
Uf-8	0.03	0.97	0.65	0.35	0.01	0.99

Table 8.5, shows the ratios of cell currents ($L_{r,1}$ for HVT and $L_{r,2}$ for SVT), node capacitances ($C_{r,1}$ for HVT and $C_{r,2}$ for SVT) and the currents within a critical path ($TL_{r,1}$ for HVT and $TL_{r,2}$ for SVT). The capacitance is dominated by HVT cells because of their higher presence in the circuit. Secondly, the difference between the capacitance of two different threshold cells is minor. However, the overall leakage current both for the complete circuit and specifically the critical path are dominated by SVT cells because of higher current leakage. Similarly, Table 8.6, shows the ratios of cell currents ($L_{r,1}$ for HVT and $L_{r,2}$ for LVT), node capacitance ($C_{r,1}$ for HVT and $C_{r,2}$ for LVT) and the currents within a critical path ($TL_{r,1}$ for HVT and $TL_{r,2}$ for LVT). The H+L implementation has similar characteristics to the H+S implementations

Table 8.7, presents the EMVs of each gate-level implementation, including the maximum clock frequency at EMV, the corresponding throughput in samples per second, and the energy dissipated per sample. These simulation results are calculated by (4.6) for single threshold implementations and by (4.26) for multi-V_T implementations. The estimated throughput calculated deviates 30 % from the actual speed of the circuit, as confirmed by spice simulations. However, the information is good enough for design space exploration. The comparison of different architectures for the EMV point shows that *Uf-2* HBD filter appears to be the architecture that dissipates least energy per sample in most of the cases. The *Uf-2* architectures dissipate least energy per sample at EMV. The simulations show that the SVT and H+S implementations are able to operate at a moderate frequency at EMV with mimimal energy dissipation compared to their counterparts. The reason for this behavior is higher currents in the cells, both drive currents and leakage currents. Increased leakage, and drive current, pushes the frequency higher. Therefore, required throughputs are achieved at lower voltages that helps in reduction of energy per sample. In the case of LVT the energy dissipation is high as the leakage currents are higher in this implementations, which also pushes the EMV point to a slightly elevated supply voltage. The H+L implementations suffers from high switching activity due to false switching that causes the EMV to shift below 200 mV

Table 8.7.: Characterization of the Implementations at EMV

Arch.	**Cells**	**EMV** [mV]	**Freq.** [kHz]	**Throughput** [ksamples/s]	**E/smp** [fJ]
Org	HVT	241	23	23	45
	SVT	237	398	398	46
	LVT	251	3710	3710	55
	HVT+SVT	220	250	250	45
	HVT+LVT	216	2133	2133	51
Uf-2	HVT	238	23	46	35
	SVT	242	383	767	40
	LVT	271	6600	13200	55
	HVT+SVT	206	182	364	46
	HVT+LVT	190	1450	2904	54
Uf-4	HVT	247	22	88	38
	SVT	241	297	1180	41
	LVT	280	6500	26000	59
	HVT+SVT	215	170	680	44
	HVT+LVT	193	1104	4416	62
Uf-8	HVT	251	15	120	48
	SVT	248	250	2000	52
	LVT	272	3570	28560	61
	HVT+SVT	220	149	1192	50
	HVT+LVT	189	616	4928	77

and the dynamic energy increase exponentially with the increase in the supply voltage. This results in higher energy dissipation at Reliable operating voltage for H+L implementation.

The simulations show that the maximum clock frequency increases exponentially with increasing supply voltage, i.e., the current increases exponentially in the cells, which leads to a further analysis on energy dissipation versus throughput.

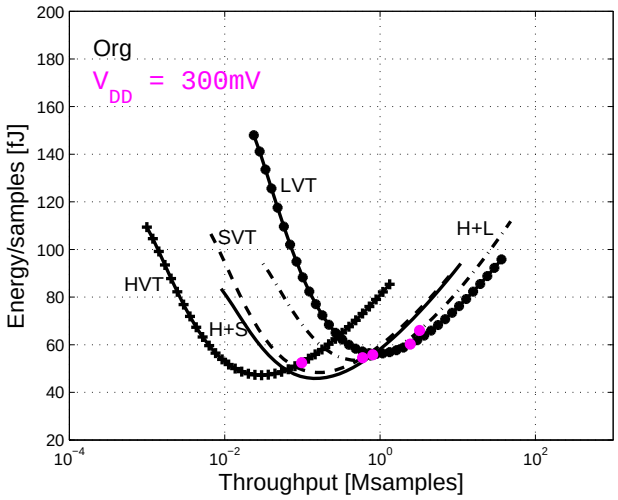


Figure 8.12.: Energy vs Throughput simulation plots of simplified HBD filter (*Org*) architectures

8.3.1. THROUGHPUT CONSTRAINTS

Figure 8.12, shows the energy dissipated versus the throughput by gate-level implementations of *Org* for the various threshold voltage options, indicated as H+S, SVT, and HVT. Similarly, Figure 8.13, 8.14, and 8.15, show the energy dissipation versus the throughput curves for the *Uf-2*, *Uf-4* and *Uf-8* architectures. These figures, shows that the SVT and H+S implementations are faster than the HVT implementations, however, are slower than LVT and and H+L implementations. The multi-V_T implementations are also fast, and close to pure SVT or LVT implementations, correspondingly. This result may be explained by the fact that the multi-V_T implementation use both HVT cells for reduced static energy and use SVT or LVT cells in the crictial path s to increase the speed. For example in the case of H+S, HVT cells are slow, however, critical paths are synthesized to get a throughput rate close to pure SVT implementations. Therefore, the speed almost matches the speed of SVT implementation, and the same is applicable for H+L implementations.

The throughput constraints for the system are 2 and 1 Msamples/s for the first two decimation filters and for the last two 500 and 250 ksamples/s. These requirements are fulfilled with least energy dissipation by different architectures, see Table 8.8. The most energy efficient architecture for 2 Msamples/s is *Uf-4* synthesized using SVT cells. *Uf-2*, synthesized with SVT cells are the

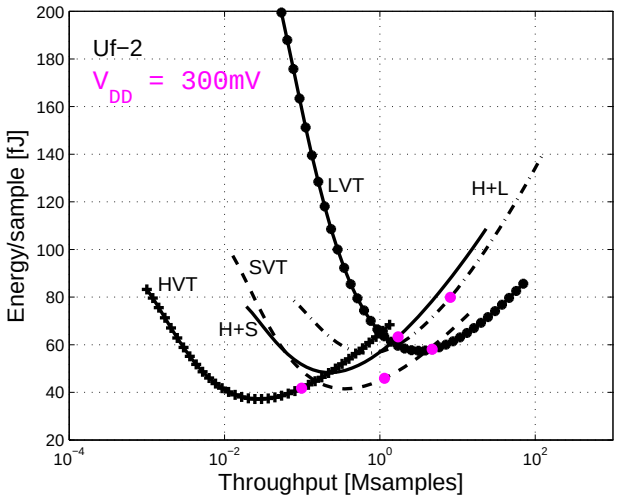


Figure 8.13.: Energy vs Throughput simulation plots of unfolded by 2 HBD filter (*Uf-2*) architectures

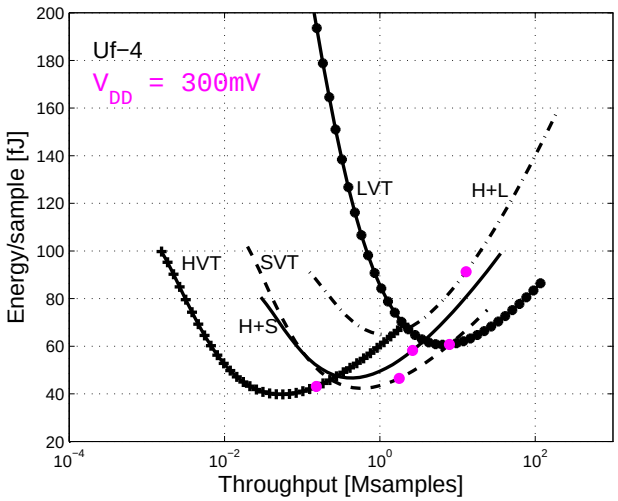


Figure 8.14.: Energy vs Throughput simulation plots of unfolded by 4 HBD filter (*Uf-4*) architectures

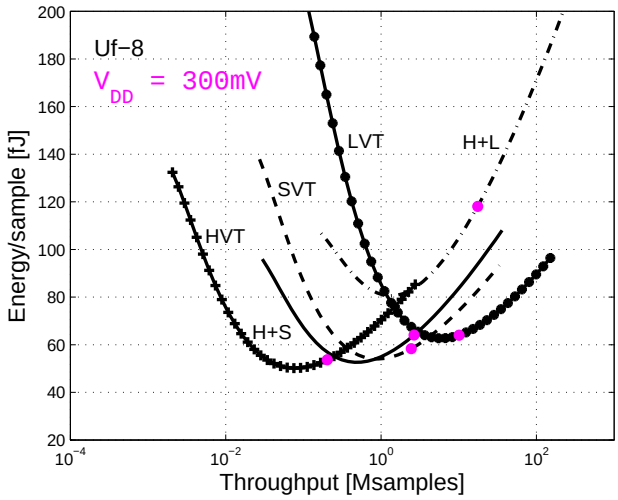


Figure 8.15.: Energy vs Throughput simulation plots of unfolded by 8 HBD filter (*Uf-8*) architectures

energy efficient for 1 Msamples/s throughput requirements. For the throughput requirement of 500 ksamples/s, *Uf-2* synthesized with SVT cells is the most energy efficient. Lastly, throughput requirement of 250 ksamples/s, Uf-dissipates least energy per output sample. Table 8.8, presents the energy dissipation per sample for the required throughputs at corresponding supply voltages. The architectures are selected for the best threshold options. The optimal values are shown in the bold in Table 8.8. The total energy dissipation per output sample for the filter chain is 164 fJ.

8.3.2. SUPPLY VOLTAGE AND THROUGHPUT CONSTRAINTS

The simulations show that the required throughput for the first, second, third, and fourth decimation filter are fulfilled by various implementations of *Uf-4* and *Uf-2* at various voltages. Having multiple supply voltage levels would require DC-DC voltage level converters. Therefore the complexity increases and the cost with respect to area, and overall energy dissipation increases [63], which therefore, is not desired. A single supply voltage at 300 mV is introduced as another constraint on the system. The selection of this voltage is based on the analysis that the first filter with a higher throughput constraint is fulfilled by using *Uf-4* at 300 mV. Therefore, 300 mV is selected as a supply voltage constraint and all the filters will operate at 300 mV. The assump-

Table 8.8.: Characteristics of the HBD Filter at Required Throughput and Fixed Supply Voltage

Throughput samples/s	**Arch.**	**Best Opt.** Cells	$\mathbf{V_{DD}}$ [mV]	**E/smp[fJ]** @EMV	**E/smp [fJ]** @300 mV
2 M	*Uf-8*	SVT	288	57	62
	Uf-4	SVT	**296**	**47**	**49**
	Uf-2	SVT	320	50	-
	Org	SVT	344	63	-
1 M	*Uf-8*	H+S	260	55	80
	Uf-4	SVT	280	44	52
	Uf-2	SVT	**290**	**45**	**49**
	Org	SVT	390	65	-
500 K	*Uf-8*	H+S	250	52	100
	Uf-4	SVT	250	42	62
	Uf-2	SVT	**240**	**41**	55
	Org	H+S	270	50	59
	Org	SVT	285	51	**53**
250 K	***Uf-8***	HVT	300	60	**60**
	Uf-4	SVT	218	45	88
	Uf-2	SVT	**240**	**41**	**66**
	Org	H+S	235	46	69

tion that the data is provided to the filter at critical path speed is not valid any more. Therefore, equation (4.2) for clock constrained systems is used to find the energy dissipation [42]. The last column in Table 8.8, shows energy dissipation per sample at 300 mV, for the four architectures at the required throughputs. Using a single power domain will have an impact on the criteria of selection of the suitable filter structures that are least energy dissipating.

The energy dissipation of the first filter remains unchanged, as *Uf-4* is operated at critical path speed. The energy dissipation for the second filter increases, as the implementation is clocked slower than the critical path delay, and therefore, there is an increase in leakage energy. In this case *Uf-2* implemented with SVT cells is still the most suitable filter architecture at these throughput and supply voltage constraints. The most suitable architecture

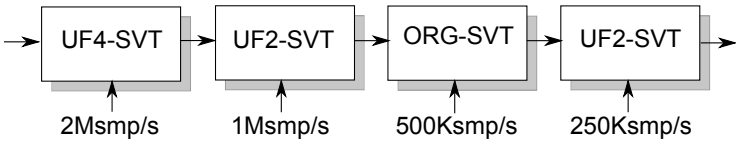


Figure 8.16.: Filter Chain optimized for V_{DD} = 300 mV

for the throughput requirements of 500 ksamples/s is *Org* synthesized with SVT cells. The *Org* architecture has the least area cost, and therefore, when not operated with critical path speed, *Org* has an advantage of less energy dissipation because of lesser leakage currents. For 250 ksamples/s the *Uf-8* synthesized with HVT cells gives the lowest energy dissipation. However, the *Uf-2* synthesized with SVT cells has a relatively low energy dissipation with smaller area, as shown in Table 8.8. Hence, with all the requirements in place, all filters in the chain have SVT implementations, with the first filter being *Uf-4*, the second *Uf-2*, the third *Org* and the last filter is *Uf-2* architectures, as shown in Figure 8.16. The total energy dissipation per output sample for the filter chain is 205 fJ. These simulation results show that for the required throughput constraints, the most suitable decimation filter chain dissipates 164 fJ per output sample. However, when single power domain constraint is applied, the most suitable decimation filter chain dissipates 205 fJ per output sample. Therefore, there is a loss of 42 fJ, that is equivalent to a *Uf-4* HBD filter implementation that gives an output of 2 Msamples/s at 300 mV. This analysis compels to find efficient ways for application of multiple power domains that dissipates less than the energy lost due to single power domain constraint. Furthermore, another option for low energy with moderate throughput requirements may be achieved by circuits operated slightly above V_T [17]. Another advantage of moderate inversion region is that the delay variation is lower than sub-V_T region. Therefore, further analysis should be carried out where V_{DD} is slightly higher than V_T. However, in this case the energy equation will change and a new energy model is needed.

8.4. SUMMARY

In this chapter various HBD filter structures are evaluated for minimum energy dissipation in the sub-V_T domain for a throughput and voltage constrained system. Scaling of V_{DD} degrades the speed of the circuit, any degradation is counteracted by parallelism techniques. An analysis on the architectures with respect to speed, and energy dissipation is vital to find the appropriate design that fulfills all the requirements with the least energy dis-

sipation. The energy model helps the design to analyze and characterize the designs, that leads to a better identification of the appropriate design. In this chapter different unfolding factors are used to achieve the required performances. First, all filter structures are implemented and simulated using 65 nm LL-HVT, LL-SVT and LL-LVT standard cells. Secondly, the design space is increased by utilization of combination of LL-HVT + LL-SVT and also LL-HVT + LL-LVT cells. The analysis with sub-V_T energy model leads to the conclusion that different architectures are suitable for different constraints. A suitable design is a synergy between parallelism, and utilization of various threshold options. However, with stringent low energy dissipation requirements combined with moderate throughput requirements unfolded architectures synthesized with SVT cells are the most appropriate option. In this analysis the multi-V_T, implementations did not show a major advantage over single V_T implementations.

9

Sub-V_T Measurements of a 65 nm CMOS Decimation Filter Chain

Measurements of a sub-threshold (sub-V_T) decimation filter, composed of four halfband digital (HBD) filters in 65 nm CMOS are presented in this chapter. Different unfolded architectures are analyzed and implemented to combat the speed degradation as presented in Chapter 7. The reliability in the sub-V_T domain is analyzed by Monte-Carlo simulations. The simulation results are validated by measurements and demonstrate that low-power standard threshold logic (LP-SVT) and different architectural flavors are suitable for a low-power implementation. Silicon measurements prove functionality down to 350 mV supply, with a maximum clock frequency of 500 kHz, having an energy dissipation of 102 fJ/cycle. The work in this Chapter has been published in [64].

The decimation filters are used in systems where data rate has to be reduced. A receiver may require to down- sample data from a high speed delta-sigma analog-to-digital converter (ADC), therefore, the decimation filters will be used. In this test case the main task of the decimation filter circuit is to re-sample the data received from the ADC at a rate of (N) ksamples/s to (N/8) ksamples/s and employed in the system proposed by [53]. Down-sampling of signals require anti-aliasing filters. In this project IIR filters are chosen instead of FIR filters, as they can be implemented with fewer coefficients for the required alias suppression. Another property of these filters is that they operate with high stability when the order of the filter is low [14]. Therefore, instead of having a high order filter, a chain of low order decimation filters are implemented. The following aspects of the circuits are discussed:

1) Analysis of process variations and delay variations due to mismatch of

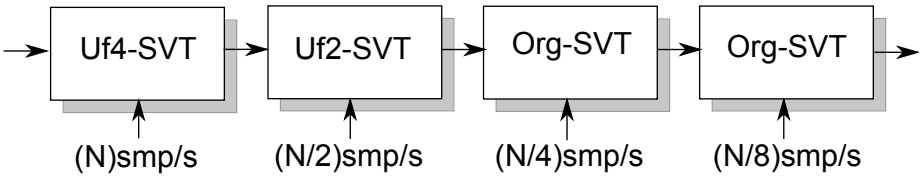


Figure 9.1.: Filter chain block diagram.

the design based on the SVT technology option.

2) Silicon fabrication of a sub-V_T ASIC.

3) Validation of the simulation results by measurements.

The rest of the chapter is structured as follows: Sec. 9.1, describes the filter chain implementation with its corresponding floor-plan. In Sec. 9.2, the simulation and measurements results obtained from the halfband (HBD) filters are shown and discussed, and finally, the summary is presented in Sec. 9.3.

9.1. HARDWARE MAPPING OF DECIMATION CHAIN

The filter chain has been synthesized with the LP-SVT standard cells option. The reason for this selection is based on a theoretical pre-study presented in [37], where the main constraints were maximum. throughput, lowest energy dissipation, and using a single power domain. The analysis performed on the designs show that the SVT implementation is able to operate at higher clock rates with a penalty of slightly higher energy dissipation. The outcome of this theoretical design space exploration was the filter chain presented in Figure 9.1. The filter chain has the first filter implemented as unfolded by 4 (*Uf-4*), the second filter as unfolded by 2 (*Uf-2*) and the last two as original (*Org*) filter architectures [37]. The inputs to the filter chain has 3-bits. The first filter is designed with 5-bit to handle the overflows. The second filter is designed with 7 bits, the third filter designed with 9-bits, and finally, the fourth filter for 11-bits. This ensures that the design maintains a sufficient level of precision and accuracy. Thereby, a decimation chain that provides a downsampling of 8 times is realized.

Tight synthesis constraints are set to achieve minimum area, minimum leakage, and a short critical path. Table 9.1, presents the normalized ratio of combinational and sequential cells in the filters with respect to *Org*. The ratio increases with unfolding factor mainly due to an increased number of adder cells. The synthesized netlist is placed and routed for fabrication. During place and route the filter chain core (FCC) is placed as a separate block

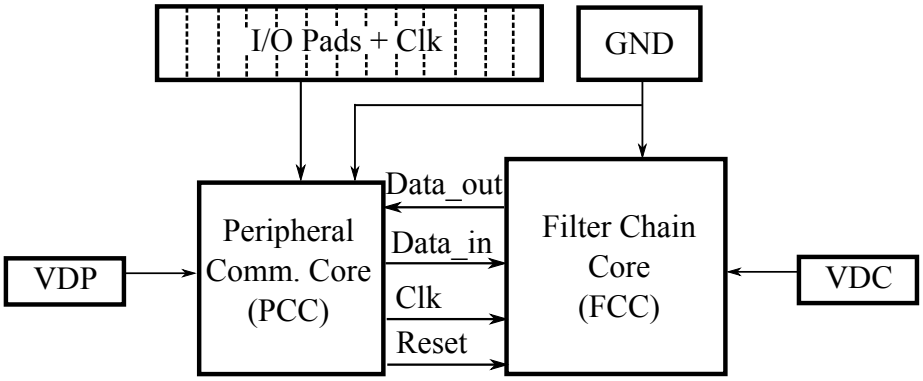


Figure 9.2.: Conceptual floor-plan.

together with a peripheral communication core (PCC) block. The purpose of the PCC is to provide communication between the FCC and the external test environment. The PCC also generates the required frequency divided clocks for the filters at decimated nodes. The benefit of an isolated FCC is that it is operated at lower supply voltages than the PCC block. Secondly, the current measurements can be performed for the FCC block stand-alone. As the PCC is connected to the pads of the chip to communicate with the test environment, it is operated at a minimum 600 mV supply voltage. The pads are directly driven from the PCC block, thus pad power supply is not needed. The FCC is operated at lower voltages compared to peripheral block. The connection between them is voltage level converter less. Further discussion is presented in Sec. 9.2.

A conceptual floorplan of the chip is shown in Figure 9.2. The design is placed on a multi-project die and the total available area for this design is 1 mm x 0.2 mm. These dimensions placed a maximum limit on the number of pads allowed, in this case 12 custom designed small pads are possible to place. The input is of 3 bits and the output of 11 bits. However, due to the pad limitations only two bits are connected with the output pads to get the functional verification. Therefore, during functional verification zero error toleration policy is observed. There is also a clock and a reset pad, two control signal pads to select from four filter outputs, and three supply pads. One of the supply pads is for the source voltage for FCC, the second for PCC, and the third is a common ground. Figure 9.3, shows the photograph of the fabricated design. The FCC and PCC are indicated together with pads for ground (GND), supply voltage for FCC and PCC i.e., VDC and VDP, respectively.

Table 9.1.: Normalized ratio of combinational and sequential cells in filters

Archi.	***Org***	***Uf-2***	***Uf-4***
Rn(C/S)	1	1.86	3.75

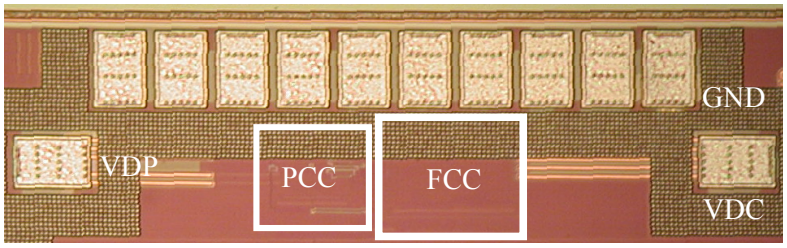


Figure 9.3.: Chip Photograph

9.2. PROCESS VARIATION AND MEASUREMENTS RESULTS

This section presents the simulation and silicon measurement results of the FCC fabricated in 65 nm LL-SVT CMOS technology. The FCC is evaluated for maximum frequency and energy dissipation for a given supply voltage. First, a simulation based analysis on the process variations and its effects on timing are presented.

9.2.1. PROCESS VARIATIONS

In the sub-V_T domain the timing is very sensitive to mismatches and variations in process and temperature [29] [30]. Therefore, 1000 (point) Monte-Carlo based simulations are performed to cover the timing analysis of the circuit. Initially, delay variation is analyzed on a minimum sized inverter. The cell selected in this case has minimum dimensions for its transistors. Figure 9.4(a) and Figure 9.4(b), show the delay variation normalized to the mean delay (μ), due to process variations and mismatches @ 400 mV and 300 mV supply, respectively. At lower voltages, the delay variation is higher than the delay variation at higher voltages. The mean delay @ 300 mV is about 20 ns and the worst case is around 80 ns, that is around 4 times the mean delay. The mean delay (μ) @ 400 mV is about 2.5 ns and the worst case is around 7.8 ns, that is around 3 times the mean delay.

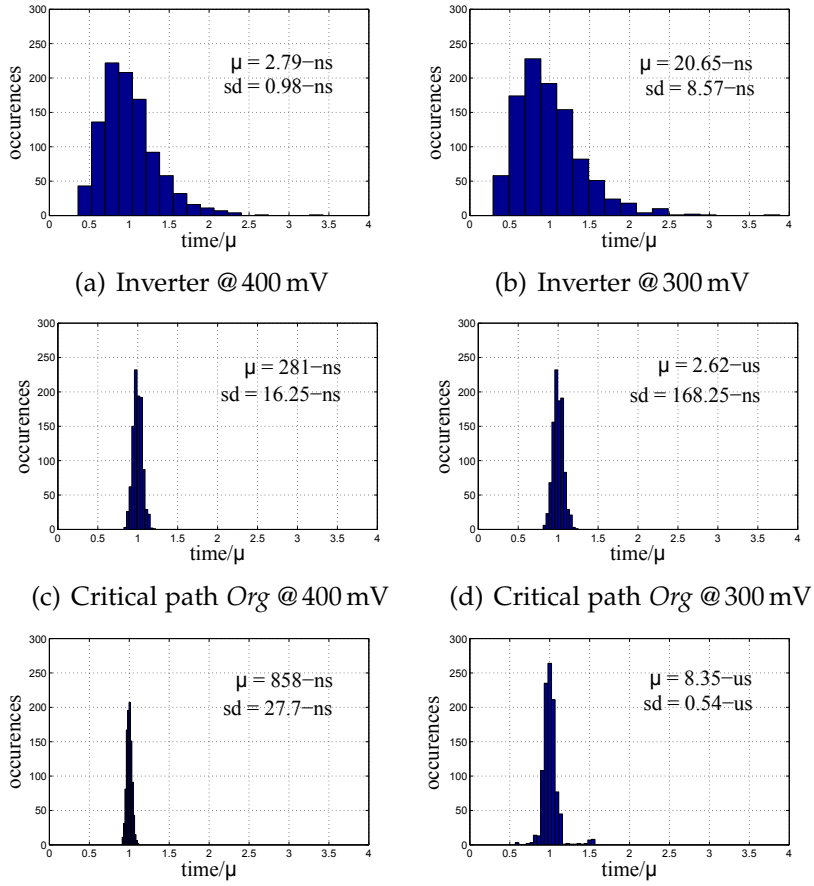
Secondly, for the filters, Figure 9.4(c) and 9.4(d), shows delay variation

spread of the critical path of the *Org* filter @400 mV and @300 mV, respectively. At 300 mV the mean delay is 2.6 μs and the standard deviation (σ) is of 168 ns, with worst case 1.3 times the mean delay. At 400 mV The mean delay is 281 ns and the standard deviation (σ) is of 16.24 ns, with worst case 1.2 times the mean delay. The deviations are acceptable in this case. This has two explanations. First the critical path is longer than for an inverter, and mismatch then tends to average out. Secondly, the transistors of the full-adder cells are almost 3 times larger than the transistors used in the inverter cells, corresponding to less mismatch [29]. As expected, the simulation results show that the transistor with minimum dimensions experiences a higher degradation when operated in the sub-V_T domain. For even longer critical paths, less relative delay variation is observed. The standard deviation (σ) of *Uf-4* filter is 27.7 ns @400 mV, with the mean delay of 0.85 μs as shown in Figure 9.4(e). The delay variation is slightly higher for longer combinatorial paths at lower voltage as shown in Figure 9.4(f). The standard deviation (σ) of *Uf-4* filter is 0.54 μs @300 mV, with the mean delay of 8.5 μs as shown in Figure 9.4(f).

For even longer critical paths the chain of same adder cells experience a slightly higher variation. The standard deviation (σ) of *Uf-4* filter is 27.7 ns @400 mV, with the mean delay of 0.85 μs as shown in Figure 9.4(e). The delay variation is slightly higher for longer combinatorial paths at lower voltage as shown in Figure 9.4(f). The standard deviation (σ) of *Uf-4* filter is 0.54 μs @300 mV, with the mean delay of 8.5 μs as shown in Figure 9.4(e). This indicates that the longer combinatorial paths for sub-threshold operations may experience higher timing variations.

9.2.2. MEASUREMENT SETUP

The functionality of the core is first verified for a minimum measurable supply voltage and the maximum clock frequency resulting in zero error rate. The test vectors and the system clock are supplied through the pattern generator. The functionality verification has been performed over 4000 samples. The currents are measured for a given voltage and frequency once the functionality is completely verified and there are no error in the output data bits. The output data is recorded with a Agilent 16822A logic analyzer. As the logic analyzer requires a minimum of 550 mV voltage swing to detect logic, the VPC is kept at 600 mV, whereas, the VDC is varied. The current drawn by the FCC is measured by a nano-ampere-meter. Furthermore, an oscilloscope is also connected to monitor the clock and data bits.



(a) Inverter @ 400 mV (b) Inverter @ 300 mV

(c) Critical path *Org* @ 400 mV (d) Critical path *Org* @ 300 mV

(e) Critical path *Uf-4* @ 400 mV (f) Critical path *Uf-4* @ 300 mV

Figure 9.4.: Delay Variation normalized to the mean delay (μ), based on 1000 point Monte-Carlo simulations.

9.2.3. SUB-V_T ENERGY MEASUREMENTS

The energy dissipation per cycle is measured by sweeping the supply voltage V_{DD} of FCC from 350 mV to 400 mV, in steps of 10 mV. The minimum clock period with zero error rate at 350 mV was found to be 2.0 μs. The clock period is kept constant and the voltage is varied to measure the average current. In simulations it was noted that the SVT cells may detect input levels that are around 300 mV lower than the maximum supply voltage of 550 mV. However, there is a degradation of rise and fall-time of the signals. At higher operational

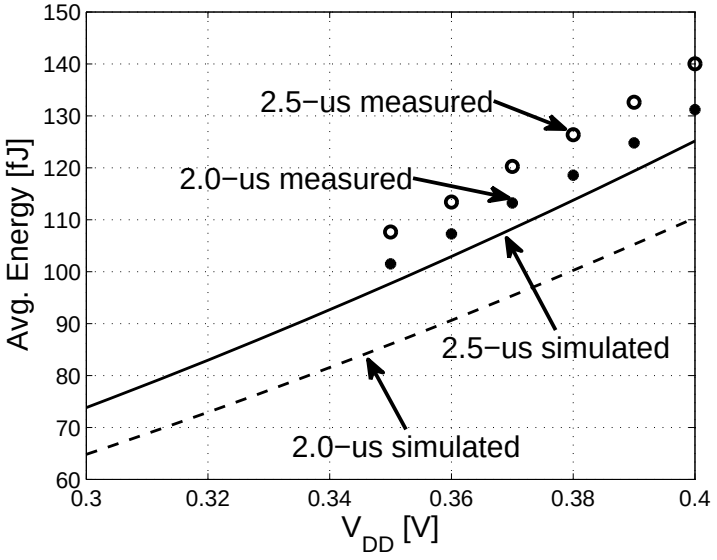


Figure 9.5.: Measured and simulated energy dissipation at 27 °C.

Table 9.2.: Measured Energy per Cycle for FCC

V_{DD} **[V]** @T	**I [μA]** = 2.5 μs	**E/c [fJ]** = 2.5 μs	**I [μA]** = 2.0 μs	**E/c [fJ]** = 2.0 μs
0.35	0.123	107.6	0.145	101.5
0.36	0.126	110.3	0.149	107.2
0.37	0.130	113.8	0.153	113.2
0.38	0.133	116.3	0.156	118.5
0.39	0.136	119.0	0.160	124.8
0.40	0.140	122.5	0.164	131.2

rates the rise-time slope is adversely effected.

Table 9.2, presents the measured results for this particular test case together with a test case where the clock period is 2.5 μs. The results in the Table 9.2 show that the energy per operation for a higher clock frequency is lower than

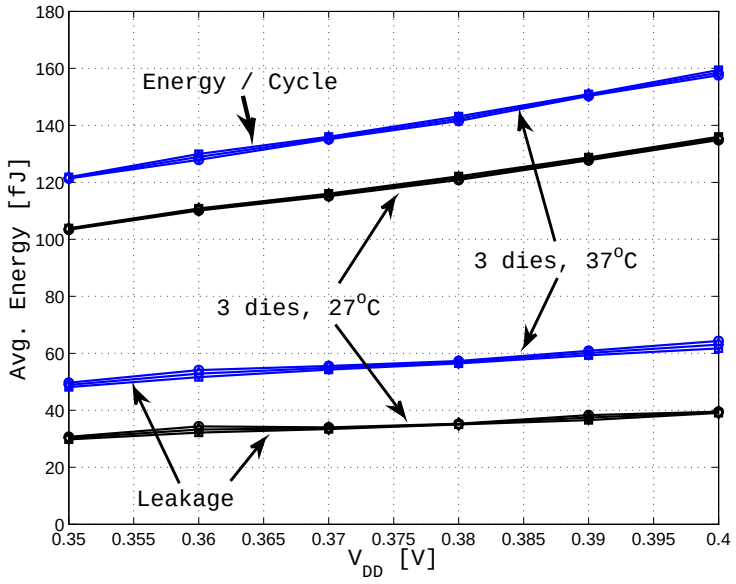


Figure 9.6.: Measured avg. energy/cycle and Measured leakage energy dissipation, at 27 ° and 37 °C.

using a slow clock frequency. As expected, the circuit if operated below the maximum frequency it will dissipate more energy per operation due to idle time, where the circuit leaks. Figure 9.5, shows the measured energy dissipation vs V_{DD} compared with simulated energy dissipation. The sub-V_T characterization are based on the energy model in [35]. The energy measurements deviate 10 % to 15 % from the simulation results mainly due to parasitic capacitance in the fabricated design. Furthermore, Figure 9.6, shows the average dynamic energy per cycle for the design operated at room temperature 27 °C and body temperature 37 °C for three dies. For these measurements the clock is kept constant at 500 kHz.

In the sub-threshold domain, leakage current is the operating current and as shown through measurements it increases at higher temperature. With the increased current, the speed of the gates to charge and discharge the output nodes increases. Therefore, the circuits can be operated at higher frequencies at higher temperatures. However, without increase in the operating frequency of the circuit, the results show that the energy dissipation increases with higher temperature, at 350 mV the rise in energy dissipation is about 17 %. Leakage energy dissipation is measured when the circuit is idle and no

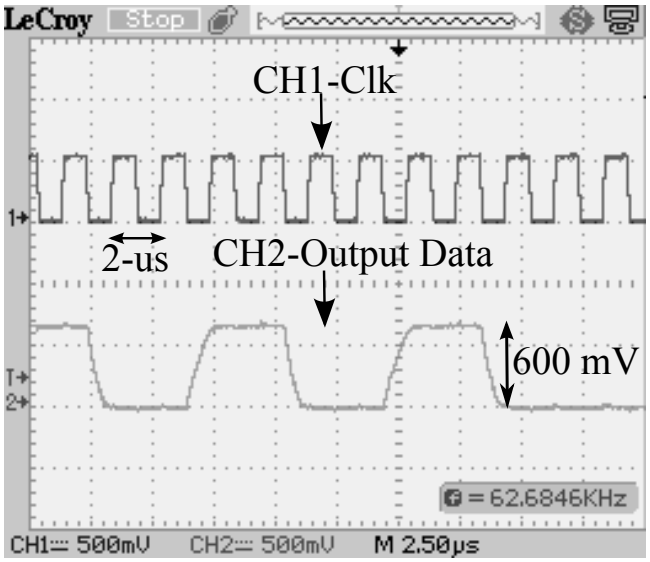


Figure 9.7.: Measured Signals.

clock or input data is supplied. Figure 9.6, also shows the average leakage energy dissipation at 27 ° and 37 °C for three dies. At 350 mV the leakage energy is around 30 fJ and 48 fJ for measurements at 27 ° and 37 °C, respectively.

Expectedly, in idle mode, when no clock or input data is supplied, an increase in energy dissipation of around 60 % is observed at body temperature. This indicates that for body area specific designs its important to shut down the device or power gate the circuits to avoid excessive energy dissipation. Figure 9.7, show the activity on one of the bits of the data line obtained from the chip, recorded using an Oscilloscope at channel 2 (CH2) and the clock supplied to the chip is recorded at CH1. The clock and the output are both at 600 mV level as the PCC receives and supplies the data at this level. Here, the clock frequency is 500 kHz.

The data obtained has a slow rise time. The reason being that the FCC is operated at 350 mV and without voltage level converters the signals experience slope degradation at PCC. Furthermore, the maximum frequency that the FCC could operate with zero errors was seen to be 500 kHz@350 mV, which deviates by 15 % from simulations. Moreover, for higher speeds it has been shown in [33] that the low-power/general-purpose LP/GP technology option provides with higher operation speed with reduced energy dissipation. Therefore, the speed of the designs can be increased by utilization of LP/GP

option in 65 nm.

9.3. SUMMARY

In this chapter a decimation filter chain designed for sub-V_T operations, evaluated for throughput, minimum energy dissipation, and a single voltage constrained system is presented. Scaling of the supply voltage (V_{DD}) degrades the speed of the circuit, any degradation is counteracted by parallelism techniques. Various unfolded filter architectures are therefore utilized to implement the filter chain. A theoretical energy model is used for initial simulations to analyze and characterize the designs. This leads to a better identification of the appropriate circuit architecture and transistor choice. The designs are analyzed for the effects on the delay spread due to process variations and mismatch in the sub-V_T domain. A filter chain synthesized with low power standard cells is fabricated in 65 nm CMOS. Measurements of the ASIC intended for sub-V_T is carried out. A close match between the simulations and measurements results is observed.

Part IV

Standard Cell Based Memories (SCM) in Sub-V_{T} Domain

This part consists of a chapter that provide an analysis on Standard cell based memories (SCM) that are operated in the sub-V_{T} region. This part includes material published in the following paper.

- P. Meinerzhagen, **S. Sherazi**, A. Burg, J. Rodrigues: , "Benchmarking of standard-cell based memories in the sub-V_{T} domain in 65 nm CMOS technology", *Journal of Emerging and Selected Topics in Circuits and Systems*, Vol. 1, No. 2, pp. 173-182, 2011.

The material in this chapter originates from the article and is mutually used by the authors

10

Analysis on Standard Cell Based Memories (SCM) in Sub-V_T

Standard-cell based memories (SCMs) are proposed as an alternative to full-custom sub-V_T SRAM macros for ultra-low-power systems requiring small memory blocks in [47]. The energy per memory access as well as the maximum achievable throughput in the sub-V_T domain of various SCM architectures are evaluated by means of a gate-level sub-V_T characterization model described in Chapter 4. The characterization of SCM is based on data extracted either from synthesized or a fully placed, routed, and back-annotated netlists. The reliable operation at the energy-minimum voltage of the various SCM architectures in a 65 nm CMOS technology considering within-die process parameter variations is demonstrated by means of Monte-Carlo circuit simulation. Finally, the energy per memory access, the achievable throughput, and the area of the best SCM architecture are compared to recent sub-V_T SRAM designs.

As an alternative to variation-tolerant full-custom circuit design, the authors in [39] [65] [66] promote the design of sub-V_T circuits based on conventional standard-cell libraries. In such conventional standard-cell based designs, embedded memory macros may limit the scalability of the supply voltage, and thus the minimum achievable energy per operation, as the noise margins gradually decrease with the supply voltage, which leads to write and read failures in the sub-V_T regime [67].

The main options for embedded memories which may be operated reliably in the sub-V_T domain are: 1) specially designed SRAM macros, and 2) storage arrays built from flip-flops or latches. Standard SRAM designs require non-trivial modifications to function reliably in the sub-V_T regime [3] [18] [68–72]. However, flip-flop and latch arrays, commonly referred to as *standard-cell based*

memories (SCMs), originally intended for super-V_T operation [73], and easily synthesized with standard digital design tools may directly be adopted in the sub-V_T domain, where they still are fully functional.

Beside being immediately compatible with voltage scaling until deep into the sub-V_T domain, SCMs bring other advantages over SRAM macros. The use of SCMs described in a hardware description language eases the portability of a design to other technologies and modifications in the memory configuration at design time. Furthermore, designs comprising SCMs can be placed automatically using the standard place-and-route tools. Consequently, SCMs may be merged with logic blocks, which may improve data locality [74] and reduce routing. Also, for reconfigurable designs targeting low power consumption, memories are preferably organized in many small blocks, which can be turned on and off separately. In the context of such fine-granular memory organizations, SCMs provide more flexibility, which may result in a smaller overall area, which are more adequate to reduce the overall power consumption.

In this chapter, the SCM architectures reported in [73] are reconsidered in the sub-V_T regime. The analysis is extended to account for the energy per memory access and the maximum achievable frequency with sub-V_T voltage scaling. By means of Monte-Carlo circuit simulation, it is shown that SCM architectures operate reliably in the sub-V_T domain even in the presence of within-die process parameter variations. Finally, the best SCM architecture is compared to full-custom sub-V_T SRAM designs regarding the energy per memory access, the maximum achievable throughput, and the silicon area.

Sections 10.1 introduces the investigated SCM architectures. The different SCM architectures are characterized and compared by means of this model in Section 10.2. Section 10.3 verifies the reliability of SCMs in the sub-V_T domain, while section 10.4 compares SCMs to full-custom SRAM macros. Section 10.5 gives the summary of the chapter.

10.1. STANDARD-CELL BASED MEMORY ARCHITECTURES

The remainder of this chapter assumes SCMs with a separate read and write port, a word access scheme, and a read and write latency of one cycle, which are typical requirements for memories distributed within dedicated datapaths. As shown in Figure 10.1(a), any such SCM accomodates the following building blocks: 1) a write logic, 2) a read logic, and 3) an array of storage cells. Different ways to implement the write and read logic are presented in Sections 10.1.1 and 10.1.2, respectively, assuming flip-flops as storage cells. The use of latches instead of flip-flops as storage cells is discussed in Section 10.1.3.

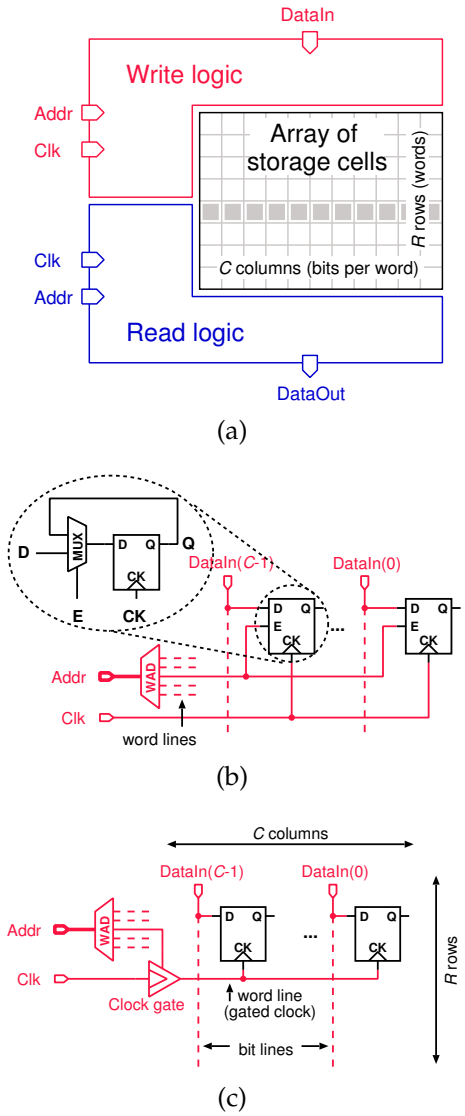


Figure 10.1.: (a) Building blocks of a generic standard-cell based memory architecture. (b) Write logic relying on enable flip-flops. (c) Basic flip-flops in conjunction with clock-gates.

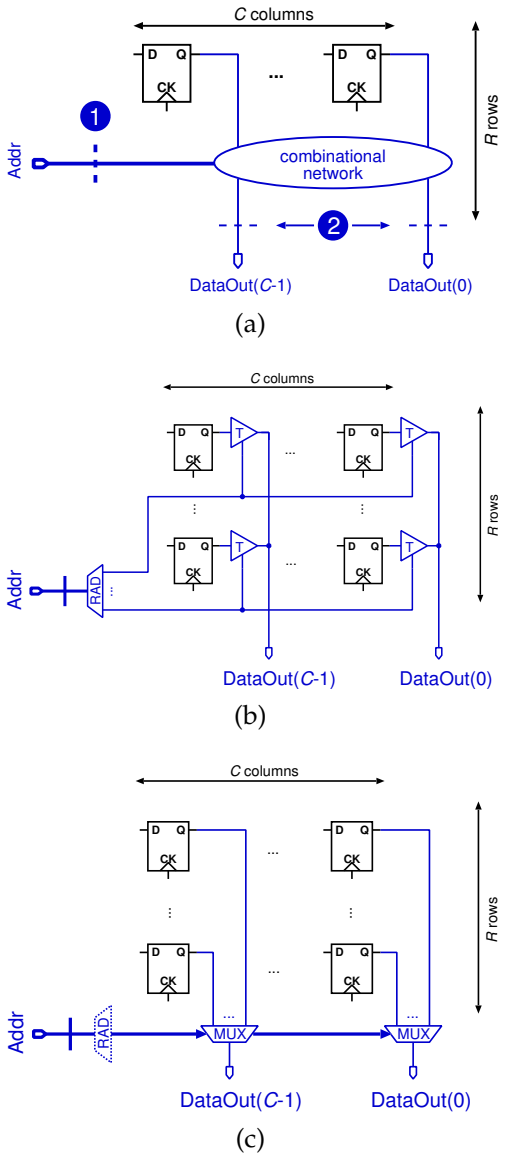


Figure 10.2.: (a) Achieving typical one-cycle read latency. (b) Read logic relying on tri-state buffers. (c) Read logic relying on multiplexers.

10.1.1. WRITE LOGIC

Consider an array of $R \times C$ flip-flops, where R and C denote the number of rows (words) and the number of columns (bits per word), respectively. Assuming a word-access scheme and a write latency of one cycle, the write logic needs to select one out of R words, according to the given write address, and update the content of the corresponding flip-flops on the next active clock edge. Accordingly, the *write address decoder* (WAD) produces one-hot encoded row select signals, which select one row of the flip-flop array. Next, the flip-flops in the selected row need to update their state according to the data to be written. One option is to use flip-flops with an enable feature or with a corresponding logic, as shown in Figure 10.1(b). A second option is to use basic flip-flops in conjunction with clock-gates, as shown in Figure 10.1(c), which generate a separate clock signal for each row so that only the currently selected row receives a clock pulse to sample the provided data, while all other rows receive a silenced clock, thereby keeping their current state.

10.1.2. READ LOGIC

As shown in Figure 10.2(a), the read logic may be purely combinational or contain sequential elements, which leads to a read latency. Assuming a word access scheme, one out of R words needs to be routed to the data output, according to the read address. The typical one-cycle latency is obtained by inserting flip-flops either at the read address input, see case (1) in Figure 10.2(a), or at the data output, see case (2) in Figure 10.2(a). The former and latter case require $\text{ceil}(\log_2(R))$ and C additional flip-flops, impose gentle and hard read address setup-time requirements, and cause considerable and negligible output delays, respectively. The task of routing one out of R words to the output is accomplished using either tri-state buffers or multiplexers.

TRI-STATE BUFFER BASED READ LOGIC

This approach asks for a *read address decoder* (RAD) to produce one-hot encoded row select signals, and $R \cdot C$ tri-state buffers, i.e., exactly one per storage cell, as shown in Figure 10.2(b). Notice that it is generally difficult to buffer tri-state buses [75], which might be necessary to maintain reasonable slew rates if these buses are routed over long distances.

MULTIPLEXER BASED READ LOGIC

C parallel R-to-1 multiplexers are required to route an entire word to the output, as shown in Figure 10.2(c). The R-to-1 multiplexer may be implemented

in many ways. Binary selection tree multiplexers do not require one-hot encoded row select signals and can therefore save the RAD. However, some glitches or activity on unselected data inputs can propagate all the way to the input of the last stage, giving rise to unnecessary power consumption. A better approach is to use a glitch-free RAD to mask (AND operation) unselected data at the leaf-level of an OR-tree to realize the multiplexer functionality.

10.1.3. ARRAY OF STORAGE CELLS

Instead of flip-flops, latches can be used as storage cells, while the previous discussions on the write and read logic remain valid. However, setup-time requirements on the write port become considerably more stringent when using latches. The reason for this is that when sticking to a single-edge-triggered one-phase clocking discipline and a duty cycle of 50%, the WAD together with the clock-gates in the latch-based design can use only the first half of a clock period to generate one clock pulse and $R-1$ silenced clocks, which will make the latches in one out of R rows transparent and keep the latches in all other rows non-transparent, during the second half of the clock period. The latches, which receive a clock pulse, store the applied input data on the next active clock edge.

Furthermore, if the currently transparent latches are also selected by the output multiplexers, the SCM becomes transparent from its data input to its data output, and combinational loops through external logic can arise. To avoid this problem, a restriction on the choice of read and write addresses needs to be imposed. If such a restriction is not desired, latches which are non-transparent during the second half of the clock period needs to be inferred at either the SCM's data input or output, or alternatively, registers needs to be inserted into any path that feeds the output data from SCM's back to the input of the SCM's.

10.2. SCM ARCHITECTURE EVALUATION

After the presentation of different architectural choices for SCMs and the sub-V_T characterization model, the aim is now aim at identifying the SCM architecture that performs best in terms of energy, but also in terms of throughput, and silicon area. All SCMs are mapped to a 65 nm CMOS technology with low-power (LP) high threshold-voltage (HVT) transistors (V_T is above 450 mV) and the results are based on fully synthesized, placed, and routed netlists with back-annotated layout parasitics. The average switching activity μ_e is obtained using voltage change dumps (VCDs) for 1000 write and read cycles. All inputs of the SCMs are driven by buffers of standard driving strength;

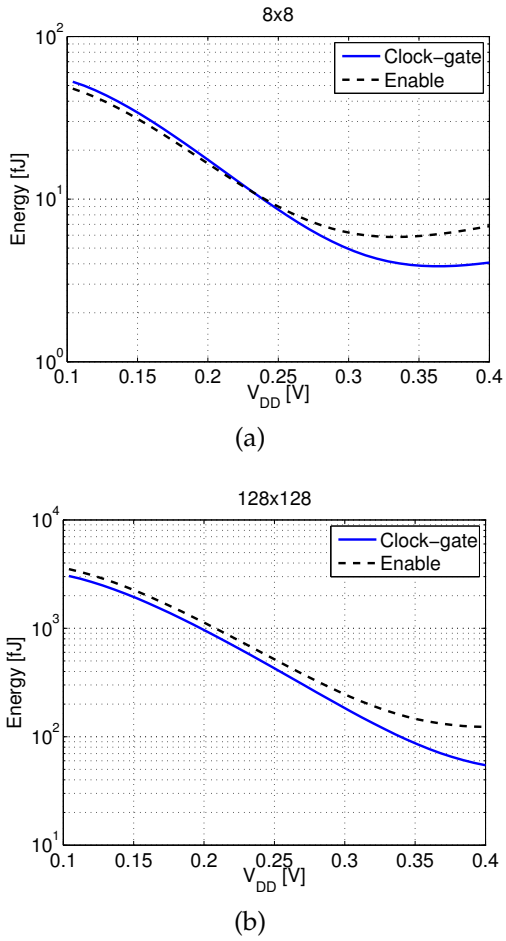


Figure 10.3.: Energy versus V_{DD} for different write logic implementations, namely *enable flip-flops* and *basic flip-flops in conjunction with clock-gates*, assuming a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$.

highly capacitive nets such as the bit lines are buffered inside the SCMs. For the comparisons between SCMs of different sizes $R \times C$, energy figures are reported as *energy per written bit* and *energy per read bit*, commonly referred to as *energy per accessed bit*. In Sections 10.2.1 and 10.2.2 the different implemen-

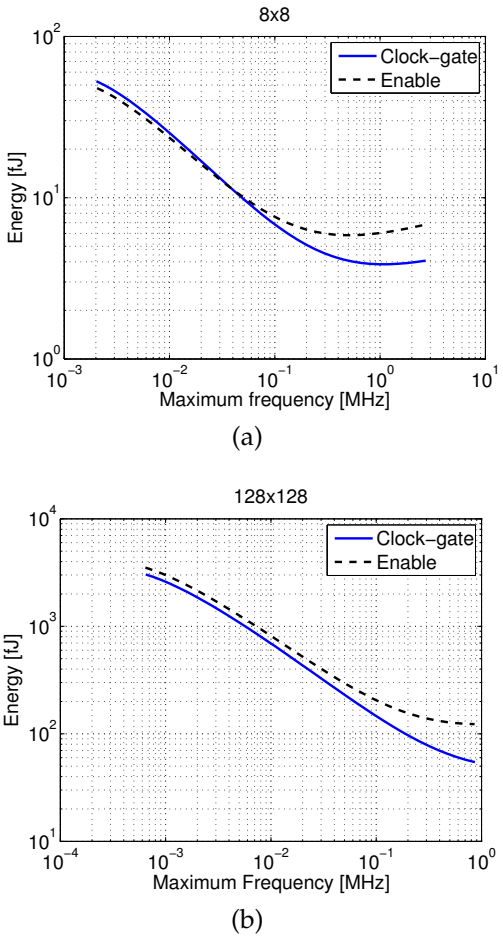


Figure 10.4.: Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (a) and (b).

tations of the write and read ports are compared and in Section 10.2.3 flip-flop arrays are compared with latch arrays.

10.2.1. COMPARISON OF WRITE LOGIC IMPLEMENTATIONS

In order to compare different write logic implementations, a multiplexer-based read logic and flip-flops as storage cells are chosen. Two memory con-

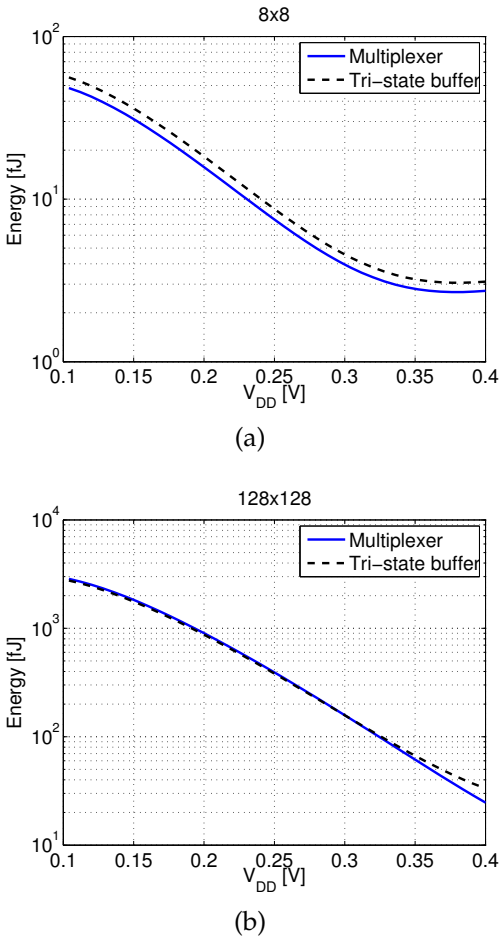


Figure 10.5.: Energy versus V_{DD} for different read logic implementations, namely *tri-state buffers* and *multiplexers*, assuming a clock-gate based write logic and latches as storage cells, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$.

figurations ($R = 8$, $C = 8$ and $R = 128$, $C = 128$) are considered, which are expected to have a smaller and to full-custom sub-V_T SRAM designs comparable area cost, respectively.

Figure 10.3(a) and Figure 10.3(b) show the energy per written bit as a func-

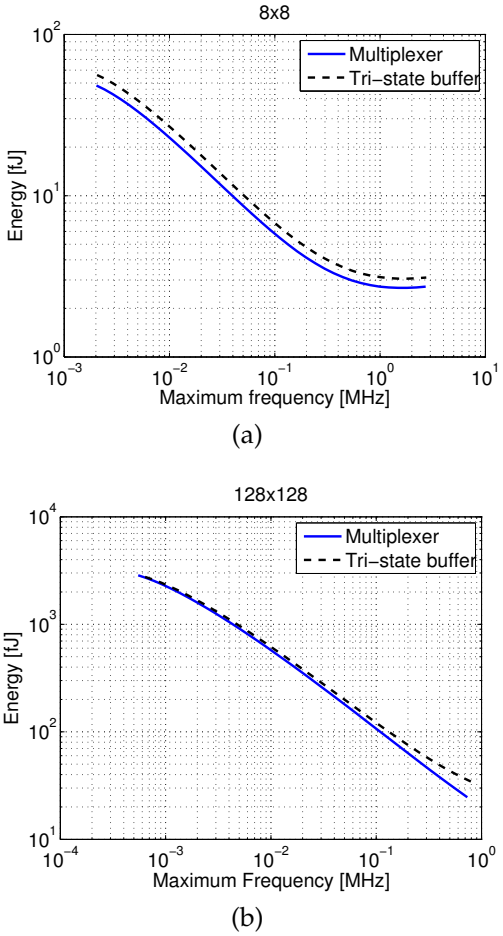


Figure 10.6.: Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (a) and (b).

tion of the supply voltage V_{DD} for the small and the larger memory configuration, respectively. In both cases, the write logic, relying on clock-gates in addition to basic flip-flops, exhibits lower energy per written bit than the architecture that employs flip-flops with enable, for the range around the energy-minimum supply voltage. In the sub-V_T regime, there are two main reason for this behavior: First, the architecture based on clock-gates dissipates

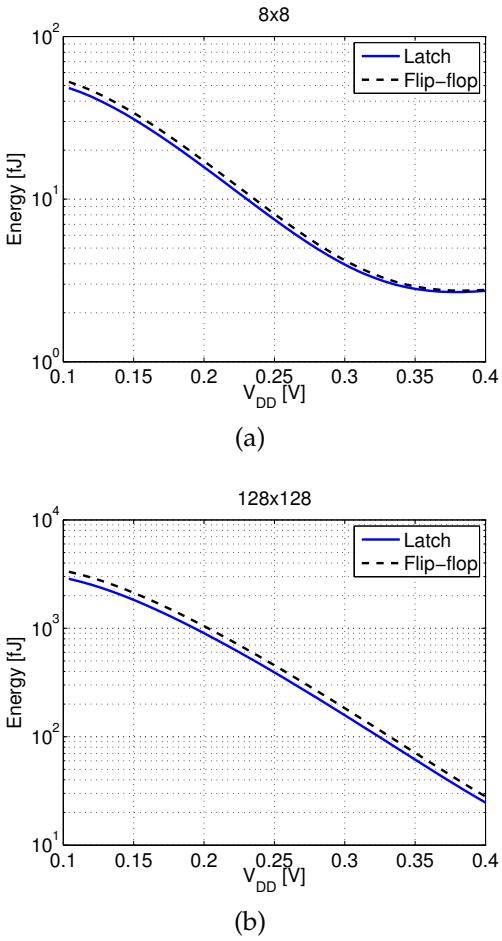


Figure 10.7.: Energy versus V_{DD} for different storage cell implementations, namely *latches* and *flip-flops*, assuming a clock-gate based write logic and a multiplexer based read logic, for (a) $R = 8$ and $C = 8$ as well as for (b) $R = 128$ and $C = 128$.

less active energy than the architecture based on enable flip-flops, as the latter distributes the clock signal to each storage cell, while the former silences the clock signal of all, but the selected row. The second reason is more visible for the larger storage array whose energy dissipation is dominated by leakage.

This leakage is larger for the case of the more complex storage cells that require additional circuitry to realize the enable for each cell in a standard-cell based implementation.

For systems that require a constrained memory bandwidth, the energy dissipation at a given frequency may also be of interest. Figure 10.4(a) and Figure 10.4(b) shows that the energy per written bit as a function of the maximum achievable operating frequency of the corresponding SCM. The frequency range on the x-axis is obtained by sweeping V_{DD} from 0.1 V to 0.4 V. It can be seen that both architectures have the same maximum operating frequencies, as the critical path is in the read logic through the output multiplexers.

With respect to area, the results in [73] show that the clock-gate architecture yields smaller SCMs than the enable architecture if only $C \geq 4$. This statement is true for many different CMOS technologies and standard-cell libraries.

In summary, the clock-gate architecture exhibits lower energy, equal throughput, and smaller area compared to the enable architecture and is therefore generally preferred.

10.2.2. COMPARISON OF READ LOGIC IMPLEMENTATIONS

In order to compare different read logic implementations, the clock-gate based write logic and a latch-based storage array are chosen for again a small and a larger SCM configuration. Figure 10.5(a) and Figure 10.5(b) show that the multiplexer based read logic with RAD has a small advantage over the tri-state buffer based read logic in terms of energy per read bit, at least around the energy-minimum supply voltage. Figure 10.6(a) and Figure 10.6(b) show that there is no significant difference between the two read logic implementations as far as the maximum achievable operating frequency is concerned. Indeed, the delay of the tri-state buffer is quite long and comparable to the delay through the entire multiplexer as all R tri-state buffers in one column are connected to the same net, which consequently has a high capacitance.

In summary, multiplexer based SCMs have a small energy and an area advantage [73], compared to the tri-state buffer approach and are therefore preferred.

10.2.3. COMPARISON OF STORAGE CELL IMPLEMENTATIONS

In order to compare different storage cell implementations, the best write and read logic implementations and again a small and a larger SCM block are considered. Figure 10.7(a) and Figure 10.7(b) show that latch arrays have less energy per accessed bit than flip-flop arrays, due to smaller leakage currents drained in each storage cell and due to lower active energy of the latch implementation. However, the energy savings of using latches instead of flip-flops

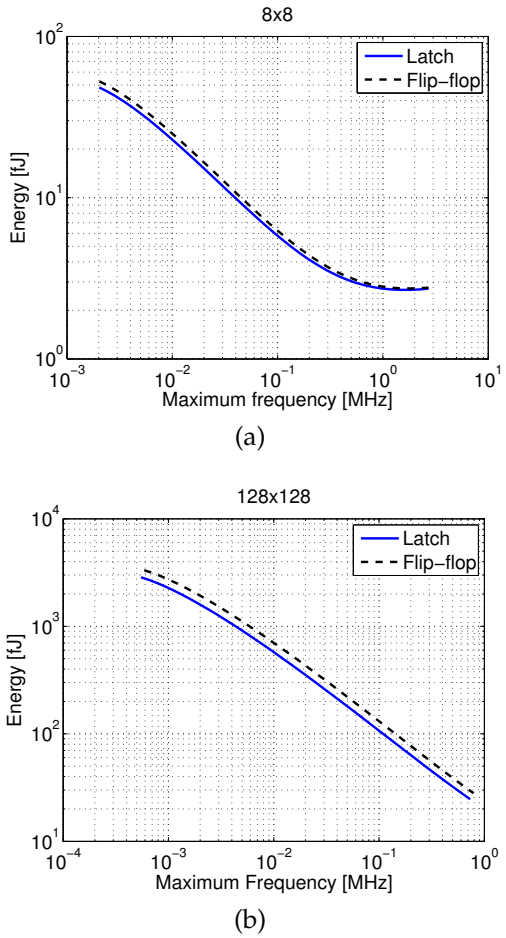


Figure 10.8.: Energy versus maximum achievable frequency for the same memory architectures and sizes is shown in (a) and (b).

are only small: a latch has around 2/3 the leakage of a flip-flop in the considered standard-cell library, but only around 2/3 of all cells in an SCM are storage cells, which accounts for the approximately 22 % energy reduction visible from Figure 10.8(b).

Figure 10.8(a) and Figure 10.8(b) show that there is no significant difference in terms of maximum frequency. In fact, the storage cells are not in the critical

Table 10.1.: Standard-cell area A_{SC} and area $A_{P\&R}$ of fully placed and routed latch and flip-flop arrays for different configurations $R \times C$, clock-gate based write logic, and multiplexer based read logic.

		Latch array		Flip-flop array	
R	C	A_{SC} [μm^2]	$A_{P\&R}$ [μm^2]	A_{SC} [μm^2]	$A_{P\&R}$ [μm^2]
8	8	738	984	811	1.1k
8	32	2.5k	3.3k	2.8k	3.7k
8	128	9.5k	12.7k	10.6k	14.1k
32	8	2.9k	3.8k	3.1k	4.2k
32	32	9.9k	13.2k	10.9k	14.6k
32	128	37.9k	50.6k	42.1k	56.2k
128	8	11.2k	15.0k	12.3k	16.4k
128	32	39.4k	52.5k	43.7k	58.3k
128	128	152.2k	202.9k	169.0k	225.4k

path, since the critical path of any SCM is through the RAD and the tri-state buffers or the multiplexers. However, flip-flops as storage cells allow for shorter write address setup-times than latches, as described in Sec. 10.1.3.

Latch arrays have only slightly smaller area than flip-flop arrays [73]. Table 10.1 shows the standard-cell area A_{SC} and the area $A_{P\&R}$ of fully placed and routed latch and flip-flop arrays for different configurations $R \times C$, the clock-gate based write logic, and the multiplexer based read logic. Notice that $A_{P\&R} = A_{SC}/0.75$, as the SCMs have been successfully placed and routed with a typical initial floorplan utilization of 75 %. An approximation of the area $A(R, C)$ for an arbitrary memory configuration $R \times C$ can be found according to

$$\begin{aligned} A(R,C) &= \beta_1 + \beta_2 R + \beta_3 C + \beta_4 RC + \\ &\quad \beta_5 \text{ceil}(log_2(R)) + \beta_6 \text{ceil}(log_2(C)). \end{aligned} \tag{10.1}$$

The coefficients $\beta_1 \ldots \beta_6$ are obtained through a least squares fit to a set of reference configurations in the technology under consideration such as the ones provided in Table 10.1.

To summarize, latch arrays have slightly less energy per accessed bit, achieve the same frequency, and are smaller compared to flip-flop arrays.

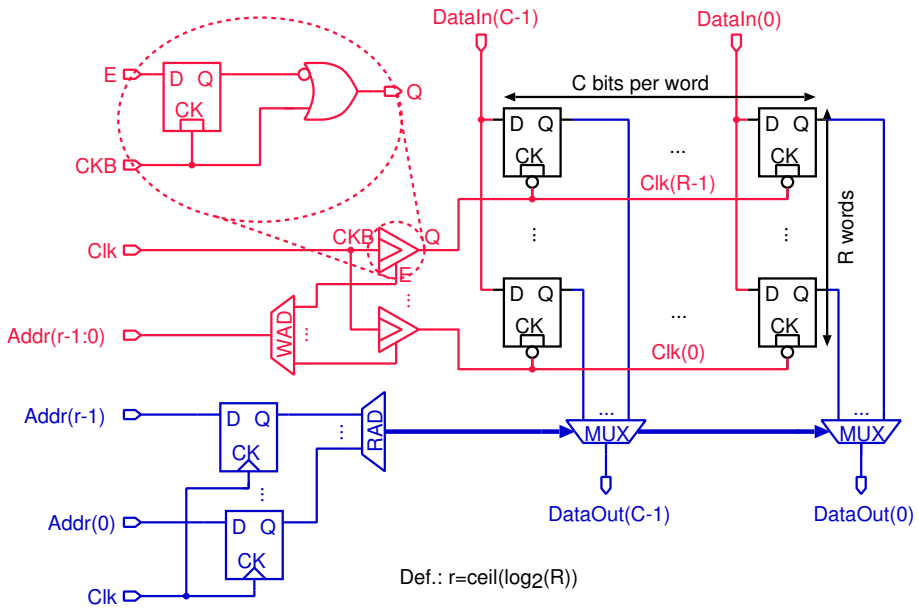


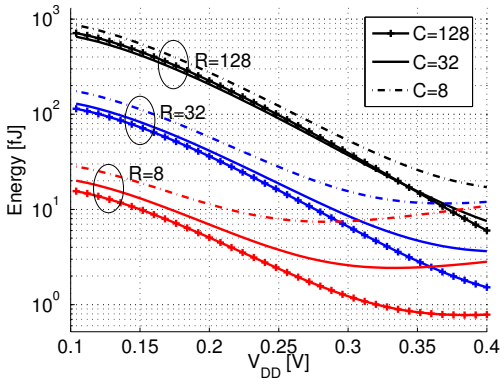
Figure 10.9.: Schematic of latch based SCM with clock-gates for the write logic and multiplexers for the read logic.

10.2.4. BEST PRACTICE IMPLEMENTATION

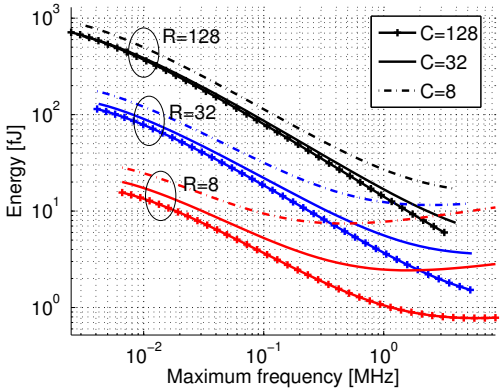
Figure 10.9 shows the schematic of the best SCM architecture. This architecture uses latches without enable feature as storage cells, clock-gates for the write logic, and multiplexers for the read logic.

With respect to the energy efficiency, it is noted that a significant switching activity is required to find an energy-minimum, which occurs only for the smallest memory configurations. However, for the large memory configurations, the overall switching activity is very low and the energy dissipation is clearly dominated by the integration of the leakage power over the access time, which decreases with increasing V_{DD} if always operating at maximum speed. Consequently, the energy-minimum supply voltage within the sub-V_T domain approaches the threshold voltage V_T when increasing the memory size.

For different memory configurations with the same storage capacity ($R \cdot C = \text{const.}$), it is observed from Figure 10.10(a) and Figure 10.10(b) that the energy-efficiency improves for a larger number of columns C and a smaller number of rows R. The reason for this behavior is that the maximum operat-



(a)



(b)

Figure 10.10.: Energy versus V_{DD} (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for different memory configurations.

ing frequency increases as R decreases, which again reduces the contribution of the energy consumed due to leakage power in each access cycle.

10.3. RELIABILITY ANALYSIS

Besides the desire to operate at the energy-minimum, one of the limiting factors with respect to voltage scaling in the sub-V_T domain is the reliability of

the circuit. Reliability issues arise mainly from within-die process variations and are aggravated in deep submicron technologies. Consequently, ensuring robust operation in the sub-V_T regime has been one of the most important concerns in the design of full-custom sub-V_T storage arrays.

Compared to full-custom designs, SCMs are compiled from conventional combinational CMOS logic gates, such as NAND, NOR, or AOI gates, and from sequential elements, i.e., latches and/or flip-flops. The reliability issue therefore corresponds to the discussion down to which supply voltage a given standard-cell library can operate reliably. This point limits in the same way the operation of the combinational and sequential logic and of the embedded SCMs for a given process corner.

To determine the range of reliable operation of the SCMs, a distinction between the combinational and the sequential cells in the library, used to construct the storage array. Previous work shows that when gradually scaling down the supply voltage, the sequential cells fail earlier than the combinational CMOS logic gates [66], provided that the combinational logic is built without transmission gates. Therefore, the focus is on the analysis of the sequential elements in the following.

The peripherals of SCM storage arrays, i.e., the read and write logic, are built from combinational CMOS gates and are thus less sensitive to process variation than the array of storage cells itself. Also, delay variations in SCM peripherals induced by process variation are unproblematic due to the used single-edge-triggered one-phase clocking discipline where path delays do not necessarily need to be matched. Compared to SCM peripherals, the peripherals of SRAM arrays are more sensitive to process variation: delay variations may cause the sense amplifiers to be triggered at the wrong time, and mismatch in the sense amplifiers can further compromise reliability, especially at very low supply voltages.

10.3.1. SENSITIVITY OF SCMS TO VARIATIONS

Reliability issues in both sequential standard-cells and in dedicated SRAM storage cells essentially arise from mismatch between carefully sized transistors due to *within-die* process variations [76]. In a conventional 6T-SRAM cell, such mismatch manifests itself in three types of failures: a) read failures, b) write failures, and c) hold failures. The read failures result from the direct access of the read bit line to the storage node, which is not present in a standard latch design such as the one shown in Figure 10.11, where the output is isolated from the internal node with a separate driver. The write failures in a 6T-SRAM cell are caused by the inability to flip storage nodes that suffer from an unusually strong keeper. The standard-cell latch avoids this issue by turn-

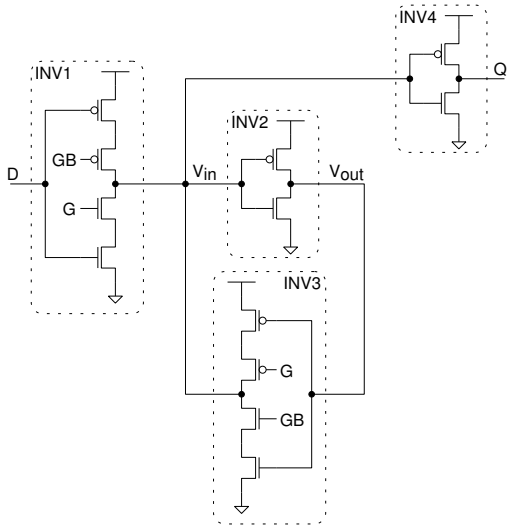


Figure 10.11.: Simplified schematic of the latch used in the best SCM architecture.

ing off the feedback path during write operation. The only remaining issue are hold failures which occurr in the non-transparent phase of a latch during which the circuit behavior essentially resembles that of a basic 6T-SRAM cell. Hence, a conventional standard-cell latch may be viewed as a very conservative SRAM cell design [3] where the reliability is determined by the risk of experiencing hold failures.

10.3.2. HOLD FAILURE ANALYSIS

Figure 10.11 shows a simplified schematic of the latch, which was chosen by the logic synthesizer from a commercial standard-cell library in order to minimize leakage and area of the latch arrays, described in this chapter. The development of new libraries with special latch topologies is beyond the scope of this study.

A latch needs to be able to hold data in the non-transparent phase. In this phase, INV2 and INV3 in Figure 10.11 act as a cross-coupled inverter pair. The stability of the state of this pair is usually defined by the *static noise margin* (SNM) that is required to hold data in the presence of voltage noise on the storage nodes [77]. This SNM is extracted as the side of the largest embedded square for the butterfly curves shown in Figure 10.12 for different supply

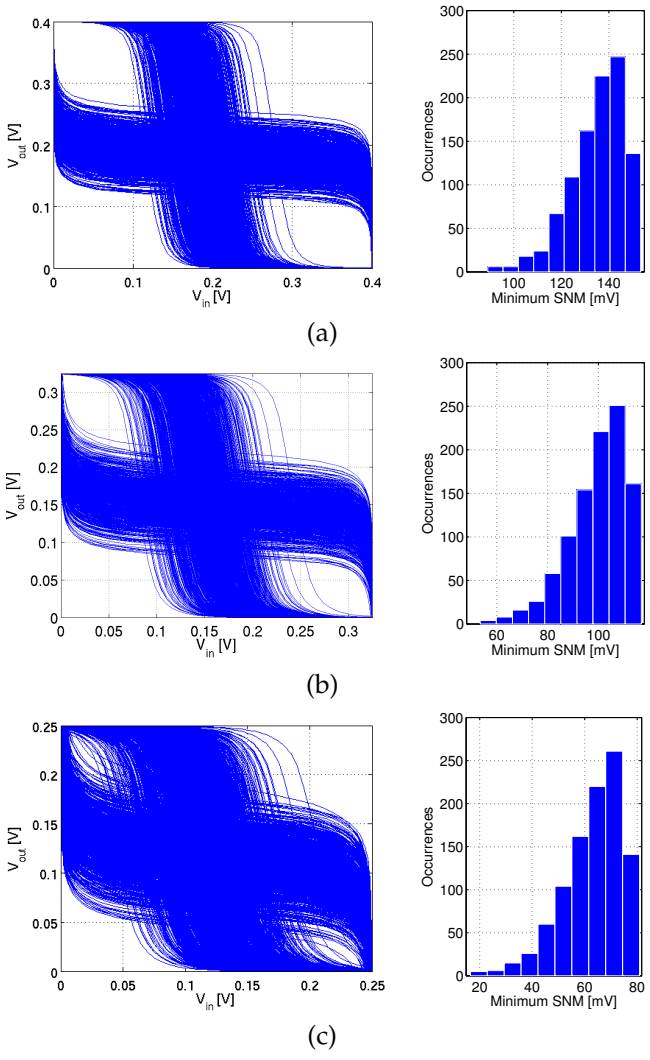


Figure 10.12.: Butterfly curves (left) and distribution of minimum hold SNM (right) of the latch used in the best SCM architecture for (a) $V_{DD} = 400\,mV$, (b) $V_{DD} = 325\,mV$, and (c) $V_{DD} = 250\,mV$.

voltages in the sub-V_T domain. For each butterfly curve, there is an SNM associated with the top-left and the bottom-right eye, referred to as *SNM high*

and *SNM low*. The probability distribution functions on the right-hand side of Figure 10.12 are always for the minimum of *SNM high* and *SNM low*. The butterfly curves and the corresponding minimum SNM distributions are obtained from a 1000-point Monte Carlo circuit simulation assuming within-die process parameter variations for the typical process corner at a temperature of 25 °C. All common parameters of the BSIM4 transistor simulation models are subject to variation according to statistical distributions provided by the foundry.

The distributions in Figure 10.12 show that the SNM values decrease with the supply voltage. As can be seen in Figure 10.12(a), there is a clear separation between the voltage transfer characteristic (VTC) of inverter INV2 and the inverse VTC of inverter INV3 corresponding to a comfortable SNM for a supply voltage of 400 mV, which also corresponds to the energy optimum supply voltage for most SCM architectures and sizes. Figure 10.12(b) and Figure 10.12(c) show that there is still a separation between the VTCs even at lower supply voltages, indicating that operation is still possible, but the SNMs are small and reliability clearly starts to become critical at 250 mV, limiting the range of operation.

10.4. COMPARISON WITH SUB-V_T SRAM DESIGNS

In this section, the performance and cost of sub-V_T SCMs is compared to a selection of sub-V_T SRAM designs in the literature [3] [68–70] [72]. Section 10.4.1 gives an overview of recent sub-V_T memory implementations including this work. Section 10.4.2 compares the energy and throughput of the smallest SCM architecture with a prominent sub-V_T SRAM design, while Section 10.4.3 compares their area.

10.4.1. OVERVIEW

Table 10.2 presents a selection of recently published sub-V_T memories. V_{DDmin} is defined as the minimum supply voltage, which guarantees reliable write, hold, and read operations. Unless otherwise stated, the maximum operating frequency f_{max} is given for $V_{DD} = V_{DDmin}$. The reported energy includes both active energy for a read operation and the leakage energy of the memory array during the access time. Furthermore, the total energy value is normalized by the width of the data IO bus, thereby reporting the total energy per read bit. Unless otherwise stated, the energy is given for f_{max} at V_{DDmin}.

All sub-V_T SRAM designs [3] [68] [69] realized in a 65 nm CMOS technology have $V_{DDmin} \geq 300$ mV. Monte Carlo simulations indicate that SCMs mapped to the same technology should operate reliably at least down to the

Table 10.2.: Comparison of sub-V_T memories.

Publication	[3]	[68]	[69]	[70]	[72]	This
Capacity [kbit]	256	256	64	8	480	32
Tech. [nm]	65	65	65	90	130	65
Basis of results	ASIC measurements					Post-layout
V_{DDmin} [mV]	380[a]	350[c]	300	160	200	300
f_{max} [kHz]	475 (0.4 V)	25	20 (0.25 V)	200	120	1 000 (0.4 V)
Energy [fJ/bit]	65.6 (0.4 V)	884.4	86.0[d] (0.4 V)	750[e]	4.2	32.7 (0.4 V)
Area [μm^2/bit]	2.9[b]	4.0[b]	7.0[b]	19.5	12.8	12.5

[a]One redundant row and column per 32-kbit block are assumed to guarantee reliable operation at this supply voltage.
[b]Area estimated from die photograph.
[c]Plus 50 mV for boosting of word line drivers.
[d]Estimation extracted from a graph.
[e]Includes the energy dissipation of the package.

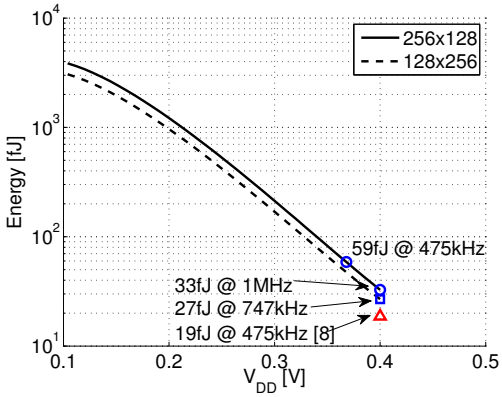
same minimum supply voltage. Two SRAM designs [70] [72] fabricated in older technologies are less sensitive to process parameter variations and are reported to have an even lower V_{DDmin}, i.e., 160 mV and 200 mV, respectively.

At the same technology node and supply voltage V_{DD}, SCMs are faster than SRAM designs, which bares the potential to lower energy dissipation per memory access if 1) speed is traded against energy, or 2) early task completion is honored by power gating. Obviously, older technologies exhibit lower leakage currents which may lead to lower energy per memory access.

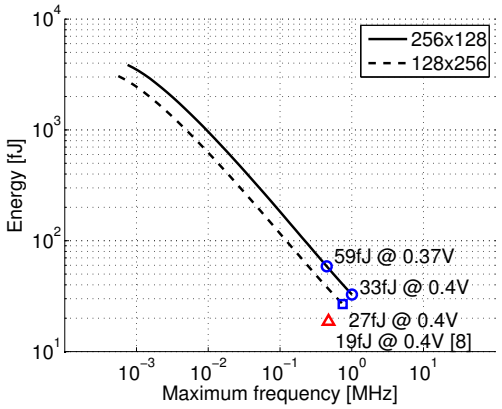
With respect to area, the use of robust latches, available from conventional standard-cell libraries, instead of 8T or 10T SRAM cells, is clearly paid for by a larger area per bit for SCMs, in the same technology.

10.4.2. ENERGY AND THROUGHPUT

A well-cited 256-kbit 10T sub-V_T SRAM [3] in 65 nm CMOS has 8 32-kbit blocks ($R = 256$, $C = 128$), which are served by a single 128-bit data IO bus. The leakage energy of this SRAM macro is divided by 8 to compare one block with the proposed 32-kbit SCM block, while the active energy is taken as is, since only one block is accessed at a time. At 400 mV, the SRAM macro is reported to be operational at $f_{max} = 475$ kHz, and a single 32-kbit block dissipates 19 fJ per accessed bit, as indicated by the triangle in Figure 10.13.



(a)



(b)

Figure 10.13.: Energy versus V_{DD} (a) and energy versus frequency (b) for the *latch multiplexer clock-gate* architecture for $R = 256$, $C = 128$ and for $R = 128$, $C = 256$. The red triangle corresponds to [3].

For comparison, Figure 10.13(a), and Figure 10.13(b), show the energy per accessed bit of the smallest SCM architecture as a function of V_{DD} and f_{max}, respectively. Considering an SCM block with $R = 256$ and $C = 128$, $f_{max} = 475\,\text{kHz}$ is already achieved at $V_{DD} = 370\,\text{mV}$ and the energy per accessed bit for this operating point is 59 fJ, which is more than for the full-custom SRAM macro. However, when operated at the same supply voltage

($V_{DD} = 400\,\text{mV}$), the SCM is able to operate at $f_{max} = 1\,\text{MHz}$, with an energy dissipation of 33 fJ per accessed bit, which is only 1.7× higher compared to the full-custom design. The energy savnings compared to the initial operating point are achieved due to a higher possible clock frequency combined with power gating after earlier completion of a task.

Changing the SCM configuration to $R = 128$ and $C = 256$ while keeping a constant storage capacity $R \cdot C$, the energy per accessed bit of the SCM is further reduced. As shown by the square marker in Figure 10.13, this new SCM configuration is able to run at 747 kHz for $V_{DD} = 400\,\text{mV}$, and dissipates 27 fJ per read bit in this operating point, which is only 1.4× higher than for the full-custom design. This change in the SCM configuration results in lower energy and doubled memory bandwidth at the price of a higher routing congestion during system integration.

10.4.3. AREA

The bitcell of SCMs (flip-flop or latch) is clearly larger than the SRAM bitcell. However, SRAM macrocells have an overhead to accommodate the peripheral circuitry, i.e., precharge circuitry and sense amplifiers [78]. For SRAM macrocells with a small storage capacity, this area overhead may be significant. Hence, SCMs may outperform SRAM macrocells in terms of area for small storage capacities, but become bigger for large storage capacities. In [73], it is shown that the border up to which SCMs still are smaller than SRAM macrocells depends on the *number of words* and the *number of bits per word*, and may be as large as 1 kbit. However, [73] considers only circuit implementations for super-V_T operation, i.e., SRAM macros based on the 6T bitcell and SCMs synthesized with a given timing constraint. When considering circuit implementations specifically optimized for sub-V_T operation, SRAM macrocells become significantly larger due to the need for 8 T [68] or 10 T [3] bitcells and the additional assist circuits required for reliable sub-V_T operation. As opposed to this, SCMs may be synthesized with relaxed timing constraints (and still reach 1 MHz in the current study) as speed is not of major concern for typical ultra-low-power applications and may therefore have a reduced area cost compared to super-V_T implementations.

In the present case, considering a storage capacity of 32 kbit, the SCM is 4.3 times larger than a corresponding SRAM block [3]. For some applications, this area increase may be acceptable for the benefit of lower energy per memory access and higher throughput.

10.5. SUMMARY

In this chapter it is show that for standard-cell based ultra-low-power designs which need to operate in the sub-V_T regime, standard-cell based memories (SCMs) are an interesting alternative to full-custom SRAM macros, which must be specifically optimized to guarantee reliable operation. The main advantages of SCMs are the reduced design effort, reliable operation for the same voltage range as the associated logic, high speed (when compared to corresponding full-custom macros), and reasonably good energy efficiency for maximum-speed operation. The drawbacks are the area penalty (for storage arrays larger than a few kbit) and a loss in energy efficiency compared to full-custom designs when operating at the same clock frequency.

Energy-efficient SCM design is driven by the fact that most of the energy is consumed due to leakage while active energy plays only a minor role, especially for large configurations. A design based on latches using clock-gates for the write logic and glitch-free multiplexers for the read logic achieves the best energy efficiency and has the smallest silicon area. For the same maximum throughput but smaller write address setup-times, the latches may be replaced by flip-flops.

Part V

Future Work

11

Future Work

There are many aspects of the sub-V_T design space that need further exploration. Techniques such as complete power shut down together with retention memory needs to be studied for the gains and drawback with respect to energy dissipation, area cost, and speed penalty. Furthermore, all the designs discussed are based on complementary CMOS gates, gate design techniques such as pass-transistor logic or transmission gate logic can be useful for sub-V_T operation, and therefore need further exploration. The analysis for the pipelining and unfolding techniques is missing a generalized formulation where an optimum structure for a circuit can be predicted by plucking in some design parameters. The dual-V_T gate technique proposed in [34] is studied for only simple gates. An analysis on larger circuits and architectural level is needed to gauge the benefits of dual-V_T gate properly. In the case of SCMs better flip-flop/latches are needed that have low leakage current consumption, as majority of the cells are not in active use and cause high static energy dissipation. Furthermore, an analysis is required to identify better clock tree structures is needed for low energy dissipation.

References

[1] I. D. ITWG. (2011, Dec) Itrs winter public conference. [Online]. Available: http://www.itrs.net/links/2011Winter/6_Design.pdf

[2] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Transactions on Circuits and Systems I: Regular Papers, TCAS-I*, vol. 59, no. 1, pp. 3–29, 2012.

[3] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 42, no. 3, pp. 680–688, Mar 2007.

[4] M. Alioto, "Understanding dc behavior of subthreshold CMOS logic through closed-form analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers, TCAS-I*, vol. 57, no. 7, pp. 1597–1607, 2010.

[5] P. van der Meer, A. van Staveren, and A. van Roermud, *Low-Power Deep Sub-Micron CMOS Logic*. Kluwer Academic Publishers, 2006, ch. 2.

[6] J. M. Rabaey and *et al.*, *Digital Integrated Circuits*. Prentice Hall, 2003, ch. 5.

[7] P. van der Meer, A. van Staveren, and A. van Roermud, *Low-Power Deep Sub-Micron CMOS Logic*. Kluwer Academic Publishers, 2006.

[8] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. Dutton, "Impact of gate direct tunneling current on circuit performance: a simulation study," *IEEE Transactions on Electron Devices*, vol. 48, no. 12, pp. 2823–2829, 2001.

[9] J. M. Rabaey and *et al.*, *Digital Integrated Circuits*. Prentice Hall, 2003, ch. 11.

[10] ——, *Digital Integrated Circuits*. Prentice Hall, 2003, ch. 10.

[11] E. Beigne, F. Clermidy, S. Miermont, and P. Vivet, in *International Symposium on Networks-on-Chip, NoCS*.

[12] I. Koren, *Computer Arithmetic Algorithms*. Prentice Hall, 2002, ch. 5.

[13] J. N. Rodrigues, "Development and implementation of cardiac event detectors in digital CMOS," PhD Thesis, Dept. EIT, LTH, Lund University.

[14] L. Wanhammar, *DSP Integrated Circuits*. Academic Publishers, 2009, ch. 4.

[15] J.-J. Kim and K. Roy, "Double gate-mosfet subthreshold circuit for ultra low power applications," *IEEE Transactions Electron Devices*, vol. 51, pp. 1468–1474, 2004.

[16] S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, and A. Fish, "Digital subthreshold logic design - motivation and challenges," in *IEEE 25th Convention of Electrical and Electronics Engineers*, Dec 2008, pp. 702 –706.

[17] D. Markovic, C. Wang, L. Alarcon, T.-T. Liu, and J. Rabaey, "Ultra low-power design in near-threshold region," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 237 –252, Feb 2010.

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

[19] E. Vittoz, *Low-Power Electronics Design*. CRC Press, 2004, ch. 16.

[20] P. van der Meer, A. van Staveren, and A. van Roermud, *Low-Power Deep Sub-Micron CMOS Logic*. Kluwer Academic Publishers, 2006, ch. 3.

[21] M. Van der Tol and S. G. Chamberlain, "Drain-induced barrier lowering in buried-channel MOSFET's," *IEEE Transactions on Electron Devices*, vol. 40, no. 4, pp. 741–749, 1993.

[22] T.-Y. Chan, J. Chen, P.-K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on MOSFET scaling," in *International Electron Devices Meeting*, vol. 33, 1987, pp. 718–721.

[23] M. Akizawa and S. Matsumoto, "The effect of fowler-nordheim tunneling current stress on mobility in n-channel MOSFETs," *IEEE Transactions on Electron Devices*, vol. 35, no. 2, pp. 245–246, 1988.

[24] M. Alioto, "A simple and accurate model of input capacitance for power estimation in cmos logic," in *International Conference on Electronics, Circuits and Systems, ICECS 2007*, Dec 2007, pp. 431 –434.

[25] Y. Pu, J. de Jesus Pineda de Gyvez, H. Corporaal, and Y. Ha, "V_T balancing and device sizing towards high yield of sub-threshold static logic gates," in *International Symposium on Low Power Electronics and Design, ISLPED*, 2007, pp. 355–358.

[26] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal supply and threshold scaling for subthreshold cmos circuits," in *IEEE Computer Society Annual Symposium on VLSI, ISVLSI*, 2002, pp. 5 –9.

[27] A. Tajalli and Y. Leblebici, "Design trade-offs in ultra-low-power digital nanoscale CMOS," *IEEE Transactions on Circuits and Systems I: Regular Papers, TCAS-I*, vol. 58, no. 9, pp. 2189 –2200, Sept 2011.

[28] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," in *International Symposium on Quality Electronic Design, ISQED*, Mar 2005, pp. 516 – 521.

[29] B. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *International Symposium on Low Power Electronics and Design, ISLPED*, Aug 2004, pp. 90 – 95.

[30] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A 0.27v 30 MHz 17.7nJ/transform 1024-pt complex FFT core with super-pipelining," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers, ISSCC*, 2011, pp. 342–344.

[31] S. Luetkemeier, T. Jungeblut, M. Porrmann, and U. Rueckert, "A 200 mV 32b subthreshold processor with adaptive supply voltage control," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers, ISSCC*, 2012, pp. 484 –486.

[32] P. Meinerzhagen, O. Andersson, B. Mohammadi, S. Sherazi, A. Burg, and J. Rodrigues, "A 500 fW/bit 14 fJ/bit-access 4kb Standard-Cell Based sub-V_T Memory in 65 nm CMOS," 2012.

[33] D. Bol, J. de Vos, C. Hocquet, F. Botman, F. Durvaux, S. Boyd, D. Flandre, and J.-D. Legat, "A 25 MHz 7μW/MHz ultra-low-voltage microcontroller soc in 65nm LP/GP CMOS for low-carbon wireless sensor nodes," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers, ISSCC*, 2012, pp. 482–483.

[34] B. Mohammadi, S. Sherazi, and J. Rodrigues, "Sizing of dual-V_T gates for sub-V_T circuits," in *IEEE Subthreshold Microelectronics Conference, Sub-V_T*, 2012, pp. 1–3.

[35] O. Akgun, J. Rodrigues, Y. Leblebici, and V. Öwall, "High-level energy estimation in the sub-V_T domain: Simulation and measurement of a cardiac event detector," *IEEE Transactions on Biomedical Circuits and Systems, TBioCAS*, vol. 6, no. 1, pp. 15–27, 2012.

[36] H. Soeleman, K. Roy, and B. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Transactions on VLSI Systems, TVLSI*, vol. 9, pp. 90–99, Feb 2001.

[37] S. Sherazi, J. Rodrigues, O. C. Akgun, H. Sjöland, and P. Nilsson, "Ultra low energy design exploration of digital decimation filters in 65 nm dual-v_T CMOS in the sub-V_T domain," *Microprocessors and Microsystems*, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.micpro.2012.04.002

[38] O. C. Akgun and Y. Leblebici, "Energy efficiency comparison of asynchronous and synchronous circuits operating in the sub-threshold regime," *Journal of Low Power Electronics*, vol. 4, OCT 2008.

[39] J. Rodrigues, O. Akgun, and V. Öwall, "A <1 pJ sub-V_T cardiac event detector in 65 nm LL-HVT CMOS," *IEEE International Conference on Very Large Scale Integration, VLSI-SOC*, pp. 253 –258, 2010.

[40] S. Sherazi, P. Nilsson, O. Akgun, H. Sjöland, and J. Rodrigues, "Design exploration of a 65 nm sub-V_T CMOS digital decimation filter chain," *IEEE International Symposium on Circuits and Systems, ISCAS*, 2011.

[41] B. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits, JSSC*, 2005.

[42] O. C. Akgun, J. N. Rodrigues, Y. Leblebici, and V. Öwall, "High-level energy estimation in the sub-V_T domain: Simulation and measurement of a cardiac event detector," *IEEE Transactions on Biomedical Circuits and Systems, TBioCAS*, vol. PP, no. 99, p. 1, 2011.

[43] O. Andersson, S. Sherazi, and J. Rodrigues, "Impact of switching activity on the energy minimum voltage for 65 nm sub-V_T CMOS," in *NORCHIP*, 2011, pp. 1–4.

[44] I. Koren, *Computer Arithmetic Algorithms*. Prentice-Hall, 2002.

[45] R. Hussin, A. Y. M. Shakaff, N. Idris, Z. Sauli, R. Ismail, and A. Kamarudin, "An efficient modified booth multiplier architecture," in *International Conference on Electronic Design, ICED*, 2008, pp. 1–4.

[46] P. Meinerzhagen, O. Andersson, S. Sherazi, A. Burg, and J. Rodrigues, "Synthesis Strategies for Sub-V_T Systems," *European Conference on Circuit Theory and Design, ECCTD*, 2011.

[47] P. Meinerzhagen, S. Sherazi, A. Burg, and J. N. Rodrigues, "Benchmarking of standard-cell based memories in the sub-V_T domain in 65-nm cmos technology," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems, JETCAS*, 2011.

[48] L. Nazhandali, B. Zhai, J. Olson, A. Reeves, M. Minuth, R. Helfand, S. Pant, T. Austin, and D. Blaauw, "Energy optimization of subthreshold-voltage sensor network processors," in *International Symposium on Computer Architecture, ISCA.*, 2005, pp. 197–207.

[49] M. Choudhury, V. Chandra, R. Aitken, and K. Mohanram, "Time-borrowing circuit designs and hardware prototyping for timing error resilience," *IEEE Transactions on Computers*, vol. PP, no. 99, pp. 1–1, 2012.

[50] K. K. Parhi, *VLSI Digital Signal Processing Systems*, 1999, ch. 5.

[51] S. Sherazi, J. Rodrigues, O. Akgun, H. Sjöland, and P. Nilsson, "Ultra low energy vs throughput design exploration of 65 nm sub-V_T CMOS digital filters," in *NORCHIP*, 2010.

[52] P. Nilsson, A. Gundarapu, and S. Sherazi, "Power savings in digital filters for wireless communication," in *European Conference on Circuit Theory and Design, ECCTD*, 2013.

[53] H. Sjöland, J. Anderson, C. Bryant, R. Chandra, O. Edfors, A. Johansson, N. Mazloum, R. Meraji, P. Nilsson, D. Radjen, J. Rodrigues, S. Sherazi, and V. OÌĹwall, "A receiver architecture for devices in wireless body area networks," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 1, pp. 82–95, 2012.

[54] P. O. Ottó Nyári, Tibor Szakáll. (2005) IIR half-band filter design with TMS320VC33 DSP. [Online]. Available: http://conf.uni-obuda.hu/sisy2005/Odry.pdf

[55] K. Surma-Aho and T. Saramaki, "A systematic technique for designing approximately linear phase recursive digital filters," in *IEEE International Symposium on Circuits and Systems, ISCAS*, vol. 5, 1998, pp. 399–403 vol.5.

[56] S. Powell and P. Chau, "A technique for realizing linear phase IIR filters," *IEEE Transactions on Signal Processing*, vol. 39, no. 11, pp. 2425–2435, 1991.

[57] M. Lutovac and L. Milic, "Approximate linear phase multiplierless iir halfband filter," *IEEE Signal Processing Letters*, vol. 7, no. 3, pp. 52–53, 2000.

[58] A. Fettweis, "Wave digital filters: Theory and practice," *Proceedings of the IEEE*, vol. 74, no. 2, pp. 270–327, 1986.

[59] P. Nilsson and M. Torkelson, "Method to save silicon area by increasing the filter order," in *Electronic letters*. ACM, NY, USA, 1995.

[60] H. Ohlsson, O. Gustafsson, and L. Wanhammar, "Arithmetic transformations for increased maximal sample rate of bit-parallel bireciprocal lattice wave digital filters," in *IEEE International Symposium on Circuits and Systems, ISCAS*, 2001.

[61] P. Aström, P. Nilsson, and *et al.*, "Power reduction in custom CMOS digital filter structures," *Journal of Analog Integrated Circuits and Signal Processing, AICSP*, vol. 18, pp. 97–105, 1998.

[62] J. Rodrigues, O. Akgun, P. Acharya, A. Calle, Y. Leblebici, and V. Öwall, "Energy dissipation reduction of a cardiac event detector in the sub-V_T domain by architectural folding," *nternational Workshop on Power And Timing Modeling, Optimization and Simulation, PATMOS*, Jun 2009.

[63] Y. Osaki, T. Hirose, N. Kuroki, and M. Numa, "A level shifter circuit design by using input/output voltage monitoring technique for ultra-low voltage digital cmos lsis," in *IEEE International New Circuits and Systems Conference, NEWCAS*, Jun 2011, pp. 201–204.

[64] S. Sherazi, P. Nilsson, H. Sjöland, and J. Rodrigues, "A 100-fJ/cycle sub-V_T decimation filter chain in 65 nm CMOS," in *IEEE International Conference on Electronics, Circuits and Systems, ICECS*, 2012, pp. 448–451.

[65] B. Calhoun, A. Wang, and A. Chandrakasan, "Device sizing for minimum energy operation in subthreshold circuits," in *IEEE Custom Integrated Circuits Conference, CICC*, Oct 2004, pp. 95–98.

[66] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 40, no. 9, pp. 1778–1786, Sept 2005.

[67] J. Chen, L. Clark, and T.-H. Chen, "An ultra-low-power memory with a subthreshold power supply voltage," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 41, no. 10, pp. 2344–2353, Oct 2006.

[68] N. Verma and A. Chandrakasan, "A 65nm 8T sub-V_T SRAM employing sense-amplifier redundancy," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers, ISSCC*, Feb 2007, pp. 328–606.

[69] M. E. Sinangil, N. Verma, and A. P. Chandrakasan, "A reconfigurable 8T ultra-dynamic voltage scalable (U-DVS) SRAM in 65 nm CMOS," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 44, no. 11, pp. 3163–3173, Nov 2009.

[70] S.-C. Luo and L.-Y. Chiou, "A sub-200-mV voltage-scalable SRAM with tolerance of access failure by self-activated bitline sensing," *IEEE Transaction on Circuits and Systems II: Express Briefs, TCAS-II*, vol. 57, no. 6, pp. 440–445, Jun 2010.

[71] M.-F. Chang, J.-J. Wu, K.-T. Chen, Y.-C. Chen, Y.-H. Chen, R. Lee, H.-J. Liao, and H. Yamauchi, "A differential data-aware power-supplied (D^2AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin applications," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 45, no. 6, pp. 1234–1245, Jun 2010.

[72] 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 *IEEE International Solid-State Circuits Conference Digest of Technical Papers, ISSCC*, Feb 2007, pp. 330–606.

[73] P. Meinerzhagen, C. Roth, and A. Burg, "Towards generic low-power area-efficient standard cell based memory architectures," in *Proc. IEEE International Midwest Symposium on Circuits and Systems*, Aug 2010, pp. 129–132.

[74] C. Roth, P. Meinerzhagen, C. Studer, and A. Burg, "A 15.8 pJ/bit/iter quasi-cyclic LDPC decoder for IEEE 802.11n in 90 nm CMOS," in *IEEE Asian Solid-State Circuits Conference, A-SSCC*, Nov 2010.

[75] J. Lillis and C.-K. Cheng, "Timing optimization for multisource nets: characterization and optimal repeater insertion," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, TCAD*, vol. 18, no. 3, pp. 322–331, Mar 1999.

[76] A. Agarwal, B. Paul, S. Mukhopadhyay, and K. Roy, "Process variation in embedded memories: failure analysis and variation aware architecture," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 40, no. 9, pp. 1804–1814, Sept 2005.

[77] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65 nm CMOS," *IEEE Journal of Solid-State Circuits, JSSC*, vol. 41, no. 7, pp. 1673–1679, Jul 2006.

[78] K.-S. Yeo and K. Roy, *Low-Voltage, Low-Power VLSI Subsystems*. McGraw-Hill, 2005.