UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

KLEBER HUGO STANGHERLIN

Energy and Speed Exploration in Digital CMOS Circuits in the Near-threshold Regime for Very-Wide Voltage-Frequency Scaling

Master's thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Prof. Sergio Bampi
Advisor

Prof. Valter Roesler
Co-advisor

Porto Alegre, August 2013

CIP – CATALOGING-IN-PUBLICATION

Stangherlin, Kleber Hugo

Energy and Speed Exploration in Digital CMOS Circuits in the Near-threshold Regime for Very-Wide Voltage-Frequency Scaling / Kleber Hugo Stangherlin. – Porto Alegre: PPGC da UFRGS, 2013.

68 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2013. Advisor: Sergio Bampi; Co-advisor: Valter Roesler.

1. Voltage-frequency scaling. 2. Energy-efficiency. 3. Power savings. 4. Near-threshold. I. Bampi, Sergio. II. Roesler, Valter. III. Energy and Speed Exploration in Digital CMOS Circuits in the Near-threshold Regime for Very-Wide Voltage-Frequency Scaling

“(...) love of family, hard work, and integrity.”

— Kanaan Kano

ENERGY AND SPEED EXPLORATION IN DIGITAL CMOS CIRCUITS IN THE NEAR-THRESHOLD REGIME FOR VERY-WIDE VOLTAGE-FREQUENCY SCALING

RESUMO

Esta tese avalia os benefícios e desafios associados com a operação em uma ampla faixa de frequências e tensões próximas ao limiar do transistor. A diminuição da tensão de alimentação em circuitos digitais CMOS apresenta grandes vantagens em termos de potência consumida pelo circuito. Esta diminuição da potência é acompanhada por uma redução da performance, reflexo da diminuição na tensão de alimentação.

A operação de circuitos digitais no ponto de energia mínima é comumente associada ao regime de operação abaixo do limiar do transistor, trazendo enormes penalidades em performance e variabilidade. Esta dissertação mostra que é possível obter 8X mais eficiência energética com uma ampla faixa dinâmica de tensão e frequência, da tensão nominal até o limite inferior da operação próximo ao limiar do transistor. Como parte deste estudo, uma biblioteca de células digitais CMOS para esta ampla faixa de frequências foi desenvolvida.

A biblioteca de células lógicas foi exercitada em um PDK comercial de 65nm para operação próximo ao limiar do transistor, reduzindo os efeitos da variabilidade sem comprometer o projeto em termos de área e energia quando operando em inversão forte. Para operar próximo e abaixo do limiar do transistor as células devem ser desenvolvidas com um número limitado de transistores em série. Nosso estudo mostra que uma performance aceitável em termos de margens de ruído estático é obtida para um conjunto restrito de células, onde são empregados no máximo dois transistores em série.

Reportamos resultados para projetos de média complexidade que incluem um filtro notch de 25kgates, um microcontrolador 8051 de 20kgates, e 4 circuitos combinacionais/sequenciais do conjunto de avaliação ISCAS. Neste trabalho, é estudada a máxima frequência atingida em cada tensão de alimentação, desde 0.15V até 1.2V. O ponto de mínima energia é demonstrado em operação abaixo do limiar do transistor, aproximadamente 0.29V, oque representa um ganho de 2X em eficiência energética comparado ao regime de operação próximo ao limiar do transistor. Embora o pico de eficiência energética ocorra abaixo do limiar do transistor para os circuitos estudados, nós também demonstramos que nesta tensão de alimentação ultra-baixa o atraso e a potência sofrem um impacto substancial devido ao aumento na variabilidade, atigindo uma degradação em performance de 30X, com respeito à operação próxima ao limiar do transistor.

Palavras-chave: variação de tensão e frequência, eficiência energética, economia de energia, próximo ao limiar do transistor.

ENERGY AND SPEED EXPLORATION IN DIGITAL CMOS CIRCUITS IN THE NEAR-THRESHOLD REGIME FOR VERY-WIDE VOLTAGE-FREQUENCY SCALING

ABSTRACT

This thesis assesses the benefits and drawbacks associated with a very wide range of frequency when operation at near-threshold is considered. Scaling down the supply voltage in digital CMOS circuits presents great benefits in terms of power reduction. Such scaling comes with a performance penalty, hence in digital synchronous circuits the reduction in frequency of operation follows, for a given circuit layout, the VDD reduction.

Minimum-energy operation of digital CMOS circuits is commonly associated to the sub-VT regime, carrying huge performance and variability penalties. This thesis shows that it is possible to achieve 8X higher energy-efficiency with a very-wide range of dynamic voltage-frequency scaling, from nominal voltages down to the lower boundary of near-VT operation. As part of this study, a CMOS digital cell-library for such wide range of frequencies was developed.

The cell-library is exercised in a 65nm commercial PDK and targets near-VT operation, mitigating the variability effects without compromising the design in terms of area and energy at strong inversion. For near-VT or sub-VT operation the cells have to be designed with few stacked transistors. Our study shows that acceptable performance in terms of static-noise margins is obtained for a constrained set of cells, for which a maximum of 2-stacked transistors are allowed. In this set we include master-slave registers.

We report results for medium complexity designs which include a 25kgates notch filter, a 20kgates 8051 compatible core, and 4-combinational/4-sequential ISCAS benchmark circuits. In this work the maximum frequency attainable at each supply for a wide variation of voltage is studied from 150mV up to nominal voltage (1.2V). The sub-VT operation is shown to hold the minimum energy-point at roughly 0.29V, which represents a 2X energy-saving compared to the near-VT regime. Although energy-efficiency peaks in sub-VT for the circuits studied, we also show that in this ultra-low VDD the circuit timing and power suffer from substantially increased variability impact and a 30X performance drawback, with respect to near-VT.

Keywords: Voltage-frequency scaling, energy-efficiency, power savings, near-threshold.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BTI	Bias Temperature Instability
CMOS	Complementary Metal-oxide-semiconductor
DVFS	Dynamic Voltage-frequency Scaling
FET	Field Effect Transistor
FFT	Fast-fourier transform
FF	3-Sigma Process conditions: Fast NMOS, Fast PMOS
FO4	Fan-out of four identical loads
FPU	Floating Point Unit
FS	3-Sigma Process conditions: Fast NMOS, Slow PMOS
HCI	Hot Carrier Injection
IC	Initial Condition
IR	Current-resistance (Voltage)
LER	Line-Edge Roughness
MMMC	Multi-mode Multi Corner
MOSFET	Metal-oxide-semiconductor Field Effect Transistor
NBTI	Negative-Bias Temperature Instability
NMOS	N-type MOSFET
OPC	Optical Proximity Correction
PBTI	Positive-Bias Temperature Instability
PDK	Process Development Kit
PLL	Phased-locked Loop
PMC	Power management control
PMOS	P-type MOSFET
PVT	Process Voltage Temperature
RDF	Random Dopant Fluctuations
RSCE	Reverse-short channel effect

SF	3-Sigma Process conditions: Slow NMOS, Fast PMOS
SNM	Static-noise Margin
SRAM	Static-random access memory
SS	3-Sigma Process conditions: Slow NMOS, Slow PMOS
STA	Static Timing Analysis
SoC	System on chip
TDDB	Time-Dependent Dielectric Breakdown
TT	3-Sigma Process conditions: Typical NMOS, Typical PMOS
V-F	Voltage-frequency
VFS	Voltage-frequency Scaling
VLSI	Very Large Scale Integration
VR	Voltage Regulator
VTC	Voltage-transfer curve
VT	Threshold Voltage

LIST OF SYMBOLS

σ	Standard deviation
σ_{VT}	Threshold voltage standard deviation
μ	Mean value
σ/μ	Coefficient of variation
W	Transistor width
L	Transistor length
W_P	PMOS transistor width
W_N	NMOS transistor width
W_{MIN}	Minimum transistor width
E_{DYN}	Dynamic Energy
E_C	Energy stored in the capacitor
P_{DYN}	Dynamic power
V_{GS}	Gate to source voltage
P_{STA}	Static power
V_{DD}	Supply voltage
i_{sta}	Static current
v_{out}	Output voltage
C_L	Load capacitor
i_{dyn}	Dynamic current
i_{sc}	Short-circuit current
v_{in}	Input voltage
$f_{0\rightarrow 1}$	Zero-to-one frequency
α	Switching activity factor

CONTENTS

1 INTRODUCTION

CMOS circuit design was historically focused on performance. High-end processors frequency kept increasing until the year of 2004, when reached near 4.0GHz – dissipating 120W of power in a single die. The performance-only oriented approaches for chip design no longer sustained the demanded performance increase at reasonable cost. The power scenario change was in fact caused by the failure of Dennard scaling (DENNARD et al., 1974), where intrinsic material parameters like junction potential and device threshold imposed a barrier for scaling the voltage across process nodes as Dennard predicted, thus increasing chip power-density every new technology. For overcoming this barrier, a new power-oriented design methodology has been developed. Techniques for saving power in VLSI designs vary from process optimizations, layout, circuit level, architectural, up to software stack and application.

From the architectural point of view, a significant change occurred in order to continue improving performance: the shift to multicores chips. Even though it required a whole new programming paradigm, multicore designs enabled a continued increase in performance, without compromising the design in terms of dissipated power. However, recent research points out that regardless of chip organization and topology, multicore scaling is also power limited; at the 22nm node, 21% of the transistors in a chip should be powered-off to provide a chance of dissipating the heat, and this number grows to more than 50% at the 8nm node (ESMAEILZADEH et al., 2011). In addition to this very critical power-density issue, an increasing number of applications demand ultra-low-energy solutions, able to maximize the battery lifetime for autonomous systems like sensor networks, wearable devices, lightweight sensors, ubiquitous environment monitoring, and so on.

For minimizing the power and energy of CMOS circuits, several techniques have been proposed such as: clock gating, for reducing unnecessary clock toggles in group of registers; operand isolation, for avoiding unnecessary glitch propagation through datapaths; power gating, for reducing static current and thus improving static energy consumption; multi-threshold devices, for reducing static energy in non-critical paths; and finally, voltage-frequency scaling (VFS), which reduces voltage and frequency, seeking to explore the bursty nature of most workloads. These techniques enable the design of processors that serve a wide spectrum of performance and energy dissipation range, providing highly configurable solutions to trade-off energy and performance – without compromising time-to-market.

This work dedicates special attention to the VFS technique, which enables a *dynamic* change in key aspects of the core such as voltage and frequency. Thus, providing the integrated run-time capability of power-performance exploration according to the workload. Notably, in most applications the workload varies from full performance (like in processing/transmit modes), to power-savvy operation in idle modes. In this scenario of varying workloads, dynamic VFS emerges as a preeminent energy-saving technique, causing a dramatically, monotonic decrease in power, with one order of magnitude energy-savings. The industry however, explores a narrow range of dynamic VFS in high-end processors, which usually translates to a 2X – 3X frequency variation.

The limited employment of the VFS technique in industry occurs due to process and environmental variability effects. These variability issues are also present at nominal voltage, but present a dramatically increased impact at lower voltages, where MOSFET devices work in sub- and near-VT regimes. When operating in those regimes, the current shows an exponential dependence on device's threshold voltage, which in turn, presents a very significant coefficient of variability in recent technology nodes (sub-100nm) (ZHAI et al., 2005). To illustrate, a design that operates at 250mV, should account for 2.5X increase in cycle-time due to 3-sigma process variation, 1.7X due to 10% supply voltage variation, and 23X for the -40^{o}C temperature corner – resulting in a total timing variation of 99X. The same design at nominal voltage, under the same conditions, should account only for an overall 1.5X variation.

This master thesis seeks to demonstrate the challenges and advantages of operating robust digital CMOS circuits in the near-VT regime, assuming a dynamic VFS case. In order to overcome the variability effects at ultra-low voltages, and ensure robust operation with respect to process and environmental variations, circuits that operate in sub- and near-VT regimes require the careful design of a robust, variation aware, digital cell-library. The designed cells should take into account all the low voltage effects that emerge when operating in such reduced supply voltages, e.g., reduced voltage swings, degraded static-noise margins, and higher dynamic/static current ratio. This new cell-library designed for ultra-low voltages may present slight penalties in performance and area when operating at nominal voltage with respect to commercial libraries, but the achieved energy-savings at lower voltages are substantial, and offer a new perspective for low-power CMOS circuit design.

Related works in (CHANDRAKASAN; SHENG; BRODERSEN, 1992), search for the lowest possible supply voltage, while the throughput degradation is compensated with architectural parallelism. In (CALHOUN; WANG; CHANDRAKASAN, 2005), the authors derive equations for the minimum energy point and show that, theoretically, minimum size devices are optimum for reducing energy. The work in (WANG; CHANDRAKASAN, 2005) demonstrates, using a FFT processor, the design methodology for a sub-VT cell-library and SRAM memory. The FFT processor is fabricated into a 0.18u technology, achieving for 16-bit/1024 points design with minimum energy point at 180mV and 10kHz frequency. In (KWONG; CHANDRAKASAN, 2006), the authors propose a new transistor sizing methodology, based on Monte Carlo simulations and static-noise margin measurements. The results showed that upsizing is a need for robust sub-VT operation. The simulation-based sizing methodology proposed in (KEANE et al., 2006) uses SPICE DC simulations to find the best transistor widths in order to improve static-noise margins. The author also derives a closed form solution for stacked devices width. The sizing of such ultra-low voltage circuits may also depend on process technology, i.e., in (KIM et al., 2007) an optimal sub-VT sizing is showed to use a larger than minimal transistor channel length due to reverse-short channel effect (RSCE).

Regarding ultra-low voltage SRAM design, a 10T SRAM cell is proposed in (CALHOUN; CHANDRAKASAN, 2007). The SRAM memory is demonstrated working properly at 380mV in 65nm CMOS. In (KIM et al., 2008) the 10T cell is showed to work at

0.2V in 130nm CMOS. From the circuit topology point of view, the 10T SRAM cell separates the read and write lines, and inserts an additional read buffer for improved bitline sensing during read operations. Even though the 10T cell operates well in deeply scaled supply voltages, the increased area due to the higher number of transistors in the cell still defines a strong limitation for its widely adoption. Even the later improved 8T cell design proposed in (VERMA; CHANDRAKASAN, 2008) is not as dense as the traditional 6T cell. In (HE et al., 2010) a SIMD processor that uses a 10Mbit SRAM memory is presented; memory density issues lead the designers to use an onchip 6T SRAM memory, which lies in a separate voltage island that does not scale.

Other strategies have also been recently developed for variability mitigation at circuit level. In (LIU; RABAEY, 2013) a neural signal processor is proposed using asynchronous design techniques, thus the critical path delay may theoretically be the minimum for any process and environmental conditions. Another demonstration of circuit level techniques for variability mitigation is presented in (LEFURGY et al., 2011), where a critical path monitor coupled with the clock generation circuitry optimizes the processor guardband time, which depends on many sources of timing deviations, e.g. process and environmental conditions, aging, workload, etc.

Recent works have demonstrated processors working under dynamic VFS. In (JAIN et al., 2012), an IA-32 microprocessor fabricated in 32nm CMOS technology works from 280mV up to 1.2V. The performance varies 3-orders of magnitude, from 3MHz up to 915MHz, enabling a 4.7X energy-efficiency variation. In (HSU et al., 2012), a reconfigurable 4-way to 32-way SIMD vector permutation engine is demonstrated on 22nm CMOS working from 240mV up to 1.1V. The performance varies from 15MHz up to 2.5GHz, while minimum energy VDD point provides a 9X greater energy-efficiency than that attainable at the nominal supply voltage.

This thesis is organized as follows: Chapter 2 discusses the major sources of variability, including process, environmental, and aging effects. This chapter also includes a quantitative analysis of variability for ultra-low power circuits from a design perspective. Chapter 3 brings a brief overview of the major techniques for ultra-low power CMOS circuit design, including clock gating, operand isolation, power gating, multi-VT devices, and dynamic VFS. Chapter 4 shows the design methodology used to develop this near-VT cell-library. Chapter 5 reports the methodology used for simulation, and list the obtained energy savings achieved with our developed near-VT cell library for dynamic VFS. Finally, Chapter 6 presents the conclusions of this work.

2 PROCESS, AGING, AND ENVIRONMENTAL VARIABILITY

Advanced process nodes present an increasing challenge for the circuit designer with respect to variability. This chapter discusses the main sources of variation in CMOS circuits, including systematic sources through lithography, and also statistical as the dopant originated variations, random dopant fluctuations (RDF). The environmental variations are also discussed, and focus on thermal and supply voltage conditions. Regarding CMOS circuits reliability, temporal variations are also briefly explored through bias temperature instability, hot carrier injection, and time-dependent dielectric breakdown. The end of this chapter presents some quantitative analysis of process and environmental variability using a study case of a ring oscillator.

2.1 Process Variations

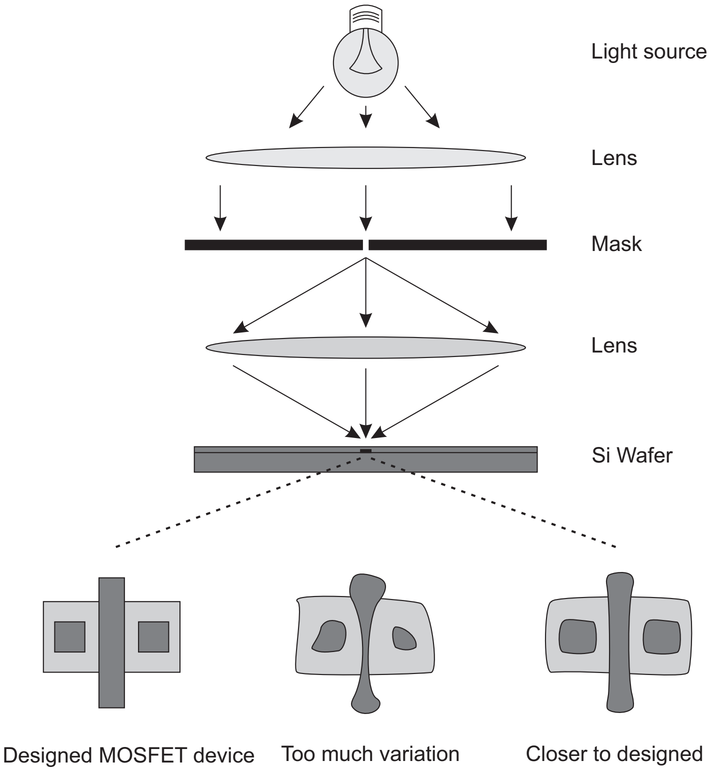
The process variations occur due to limited control over the semiconductor fabrication processes. Even though variations during fabrication have always occurred, its impact on circuit design has significantly increased in decananometer CMOS (sub-100nm). These variations affect the devices electrical characteristics, resulting in variations which may decrease performance or cause a failure, reducing the yield and increasing the cost. They can be categorized into *systematic* and *statistical* natures. The systematic variations are often related to lithography, and repeat from chip-to-chip, causing a constant offset on device characteristics. The statistical variations, on the other hand, produce chips with different characteristics according to some statistical distribution. This section presents a brief overview of the two major sources of process variations, both systematic and statistical: lithografic and dopant, respectively.

2.1.1 Lithografic Variations

Variations due to lithography depend on the physical design, and are related to imperfections on the employed tool. Figure 2.1 shows a simplified view of the lithografic apparatus required to pattern a MOSFET device, including gate, active diffusion region, contacts, vias, and interconnects. The lithography is currently employing a light source of 193nm wavelength, which is a very large wavelength considering the device features being printed on decananometer CMOS. Thus, for achieving high yield lithography, advanced lens systems and mathematical algorithms, such as optical proximity correction (OPC) are used to enable a reliable device patterning over the silicon wafer (BHUNIA; MUKHOPADHYAY, 2011).

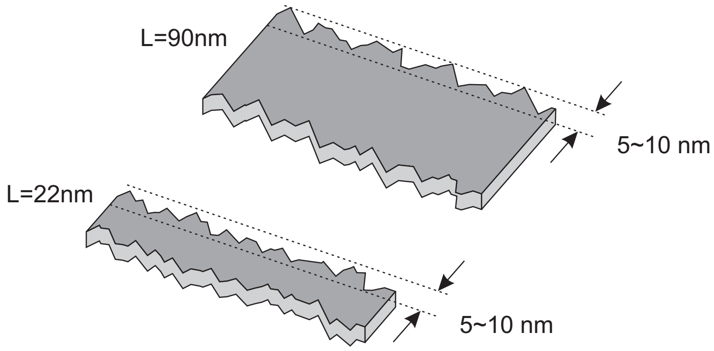
One of the major sources of variations related to lithography is Line-Edge Roughness

Figure 2.1: Systematic variations from lithography process.



Source: Modified from (BHUNIA; MUKHOPADHYAY, 2011).

Figure 2.2: Line edge roughness (LER) for polysilicon gate on devices with L=22nm and L=90nm.



Source: Modified from (BHUNIA; MUKHOPADHYAY, 2011).

(LER), which appears as an edge roughness on the gate of MOSFET devices (shown in Figure 2.2). In addition to lithography itself, this effect is also caused by the etching tools and depends on material properties. The traditional polysilicon material for example, has varied grain sizes from 5 to 10nm. Similar roughness is expected on metal gates as well. The impact of LER is much more significant in advanced nodes due to this intrinsic variation, e.g., considering the same LER roughness, a 90nm gate length device will present a smaller variation impact than a 20nm gate length. From a circuit designer perspective, the LER impact may be seen as an increased threshold voltage variation. In order to mitigate such variation, one approach is to increase the device area, since it is well known that the standard deviation of threshold voltage is proportional to the inverse of square root of the area, as stated by Equation 2.1. (ASENOV et al., 2002).

$$\sigma_{VT} \propto \frac{1}{\sqrt{WL}} \tag{2.1}$$

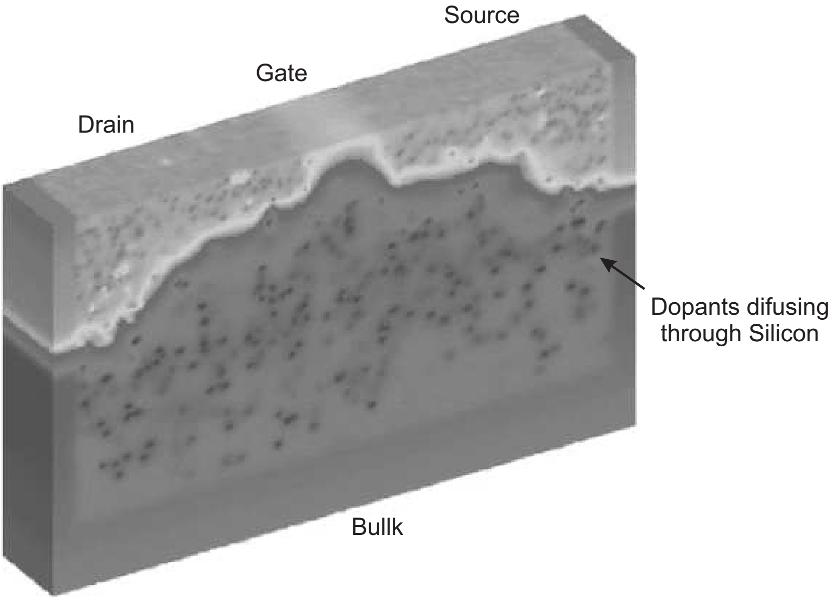
2.1.2 Random Dopant Fluctuations (RDF)

FET devices are doped with impurities to control the transistor's electrical properties, like short-channel effects and threshold-voltage. The statistical nature of the ion implantation process together with the decreasing feature size of MOSFETs made the Random Dopant Fluctuation (RDF) the major contributor on threshold-voltage variability. The variations are not only caused by the number of dopants on the channel, but also by their positions. When a FET is doped with impurities, the number of dopant atoms and their placements can not be controlled precisellly, posing an enormous challenge for decananometer CMOS with very small device areas. Figure 2.3 shows the result of a 3D simulation for a 30nm by 30nm MOSFET device, where the impurity atoms clearly diffuse across the silicon wafer.

2.2 Temperature and Supply Voltage Variations

Besides process originated variabilities, a circuit may also be influenciated by environmental conditions with respect to temperature and supply voltage. The device current is dictated by the voltages on its terminals and hence influenciated by the supply voltage.

Figure 2.3: Random dopant fluctuations (RDF) simulated in 3D for a 30nm by 30nm MOSFET device.



Source: Modified from A. Brown/University of Glasgow.

In addition, the threshold voltage and channel mobility also heavily depend on the operating temperature. Thus, these environmental conditions contribute for the overall design variability, requiring careful analysis from the circuit designer. This section reviews the major environmental variation sources; temperature and voltage.

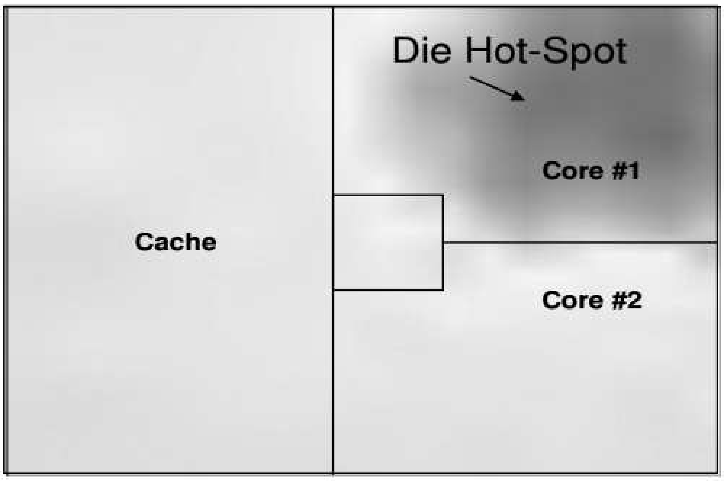
2.2.1 Thermal Management

The thermal profile at circuit level depends on switching activities and capacitive load. High activity blocks such as clock trees will present higher temperatures than low-activity ones, such as SRAMs. A large capacitive load on a block also implies a higher current, resulting in increased temperatures as well. From a system designer perspective, the workload may influenciate the spatial thermal distribution, e.g., consider a multi-core system where some of the cores are idle while others are performing intense computations; the power management algorithm should distribute tasks properly in order to balance this temperature distribution. Regarding the 3D stacking technology, much has been said about the cooling efficiency, which is very degraded for the devices far away from the heat sink. One approach for handling this issue is to place the lower activity circuits far from the heat sink, while the power-hungry components will be closer to it. Figure 2.4 shows the thermal distribution of a single core application that stresses the floating point unit (FPU) running on a Intel Core Duo processor (ROTEM et al., 2006). Note that in order to generate this thermal gradient, more than one temperature sensor is needed into the same chip, hence it is possible to capture the temperature gradient between cache, Core #2 and Core #1 (which is depicted as the hot spot).

2.2.2 Supply Voltage Tolerance

The power grid plays a very important role with respect to timing variability. A correctly designed power distribution network delivers a stable supply voltage for all por-

Figure 2.4: Thermal distribution of Intel Core Duo processor running a single core application.



Source: (ROTEM et al., 2006).

tions of the chip, with adequate current densities to mitigate electromigration (see Section 2.3.4). One of the major concerns for a robust chip design is to minimize the voltage drop across the power grid, also known as IR-drop. An IR-drop aware design must address both the steady and the static currents, creating a power grid with adequate resistance for the current demands of the design, which may vary substantially between blocks of a single chip. Figure 2.5 shows the voltage drop of a POWER6 dual core processor during start-up. The start-up was splitted into two phases, where the Core #2 turns on 4ns after Core #1. This approach was adopted in order to reduce the voltage drop due to high start-up current. Simulation measurements indicate that the noise wave from Core #1 takes only 4ns to reach Core #2, 9mm away from the first core.

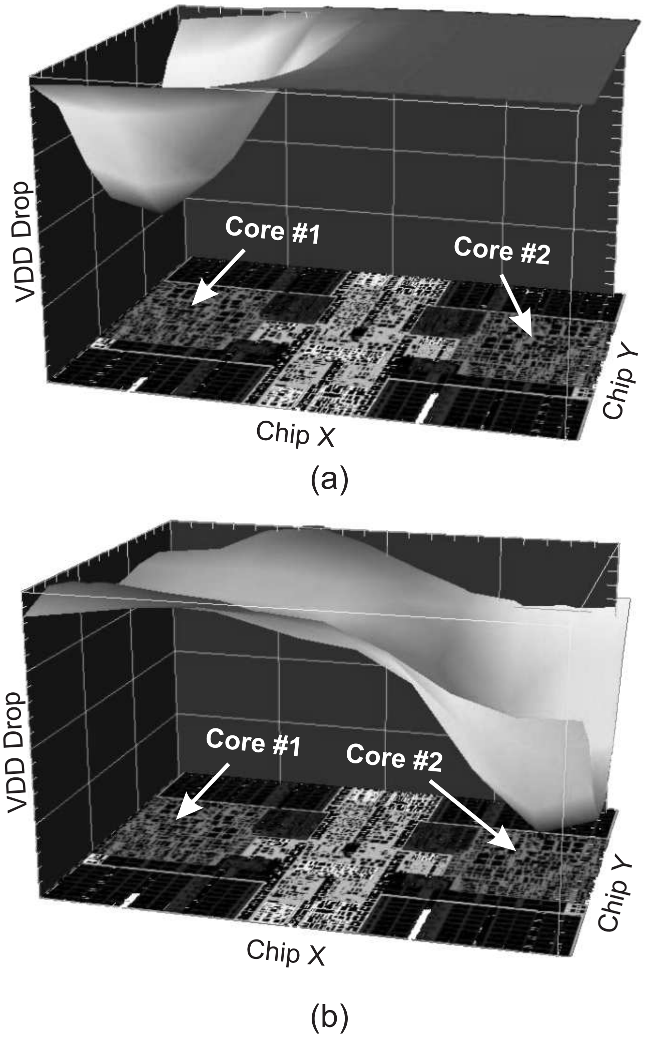
2.3 Temporal Variations: Aging

Once a chip is fabricated, tested, packaged, and shipped to the costumers, it is expected to work during its life time. Even though a chip does not have moving parts as mechanical machines do, it still suffers from aging effects. The movement of electrons and holes in a semiconductor may produce physical changes which affect the FET's behavior. It is important to mention that every aging mechanism accelerates under large electric fields and high temperature – promoting near-VT operation from the reliability point of view due to its low voltages and most likely lower temperature. The major contributors to aging in FET devices are briefly covered in this section, including Bias Temperature Instability, Hot Carrier Injection, and Time-Dependent Dielectric Breakdown. In addition to this FET related effects, another important phenomenon also discussed in this section is named electromigration and may produce severe injures to chip interconnect, especially when associated with high temperatures.

2.3.1 Bias Temperature Instability (BTI)

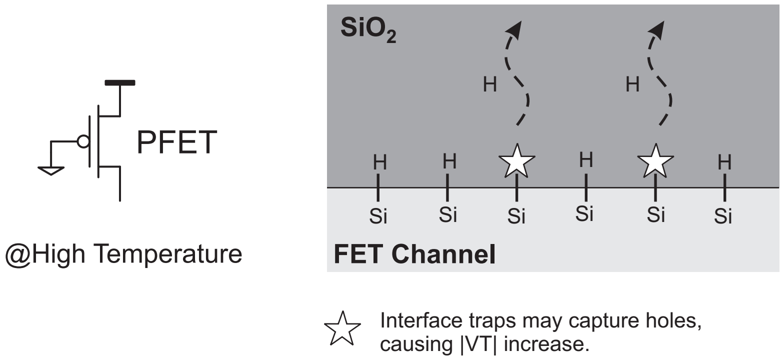
The Bias Temperature Instability (BTI) is the phenomenon where high electric fields and/or high temperature stress the MOSFET gate changing the VT of the device (DEAL

Figure 2.5: POWER6 dual core voltage drop during start-up; (a) Core #1 (left) as it turns on, and (b) Core #2 (right) as it turns on 4ns later.



Source: Modified from (FRANCH et al., 2008).

Figure 2.6: Negative-Bias Temperature instability (NBTI) for PFET devices working under high electric fields and/or high temperatures.



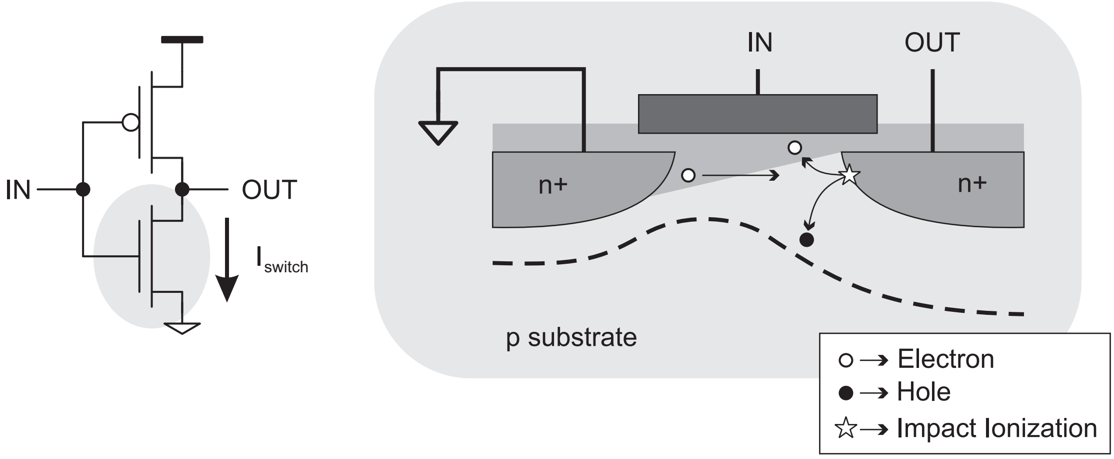
Source: Modified from (BHUNIA; MUKHOPADHYAY, 2011)

et al., 1967). It is classified into two categories: Negative-BTI (NBTI) and Positive-BTI (PBTI), one that accounts for PMOS and the other for NMOS devices, respectively. Both BTI types relate to the instability induced in a FET due to generation of traps at FET channel and gate dielectric interface. Those traps occur due to the interface of two very different materials: the highly ordered crystalline Si at the FET channel and the amorphous SiO_2 or high-k dielectric. This rough interface results in some dangling silicon atoms from the channel due to unsatisfied chemical bonds. These dangling atoms are interface traps which can lead to poor performance due to charge trapping and scattering. For mitigating those traps in a PMOS device and thus enhance device performance, FETs are hydrogen annealed during fabrication after the formation of the interface between Si and SiO_2. Hydrogen gas diffuses to the interface binding themselves to the dangling silicon atoms, thus passivating the interface traps. Figure 2.6 shows the NBTI phenomenon, which takes place when PMOS devices are under high electric fields and/or high temperature stress, under such conditions the $Si - H$ bonds can break, and this dissociation will re-generate those interface traps that were passivated during the fabrication process. These new interface traps can capture holes in an inverted PFET channel resulting in performance degradation, which may be modeled as an increase in threshold voltage. The PBTI on the other hand, relates to the NFET devices, and has become relevant only in advanced nodes where high-k dielectric is introduced.

2.3.2 Hot Carrier Injection (HCI)

Considering an NFET of a static CMOS inverter as shown in Figure 2.7. At the begining of the input switching from high voltage to ground, the output node starts to discharge through the NMOS device. The first few carriers passing through the inverted channel are accelerated under a very high electric field. Some of the accelerated electrons (hot electrons) gain high kinetic energy and, near the drain junction, generate an additional secondary electron-hole pair through impact ionization. Both the primary and secondary carriers may gain enough energy to be injected into the $Si - SiO_2$ interface, or even accumulate traps in the gate dielectric itself. These traps are electrically active and capture carriers, resulting in increased threshold voltage. The HCI occurs in both NMOS and PMOS, being more prominent in the NMOS due to the lower potential barrier at the gate dielectric interface compared to holes. Note in Figure 2.7 that the hole generated by

Figure 2.7: Hot electrons are accelerated through the inverted channel, generating electron-hole pairs through impact ionization which in turn may result in traps at the interface or gate dielectric.



Source: Modified from (BHUNIA; MUKHOPADHYAY, 2011).

impact ionization flows back to the substrate. Therefore, excessive substrate current may be an indication of HCI degradation (BHUNIA; MUKHOPADHYAY, 2011).

2.3.3 Time-Dependent Dielectric Breakdown (TDDB)

The TDDB effect refers to the gate dielectric breakdown due to high vertical electric field, which induces a time and usage dependent formation of traps at the gate dielectric. These traps may join together creating a conductive path from gate contact to the inverted channel, causing the gate dielectric breakdown. This phenomenon is of increasing concern due to continued scaling of gate dielectric thickness, i.e. less defects required to produce a breakdown. One may measure this aging mechanism by sensing gate current, which should hold an increase over the usual gate tunneling current.

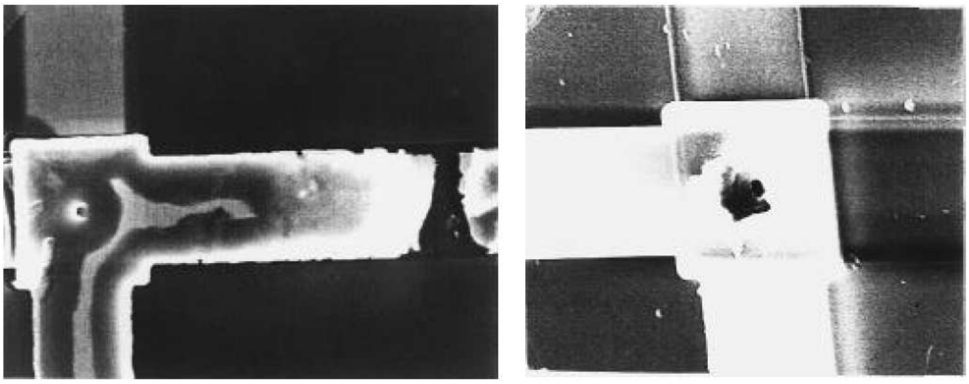
2.3.4 Electromigration

High *direct* current density in a metal wire, over a substantial time period, may cause transport of ions. Eventually, this causes the wire to break or to short circuit with another wire as show in Figure 2.8. This phenomenon takes a considerable amount of time to occur, and it is named electromigration. Even though the movement of material occurs in all wires of a circuit, electromigration only affects the wires under direct current stress, e.g. supply rails. In addition to current density, high temperature also increases the rate of degradation. Thus, from a design perspective, the most effective way to address electromigration is to use strict wire-sizing guidelines, according the estimated chip's temperature ranges (RABAEY; CHANDRAKASAN; NIKOLIC, 2003).

2.4 Impact on Circuit Performance

The process and environmental variabilities impose great challenges on circuit design for near-VT operation. This section first evaluates the variability impact on circuit performance for the employed Process Development Kit (PDK): 65nm CMOS Bulk from IBM.

Figure 2.8: Electromigration effects due to material movement on metal (a) wires and (b) vias.



(a) Line-open failure

(b) Open failure in contact plug.

Source: (RABAEY; CHANDRAKASAN; NIKOLIC, 2003).

The study case circuit for variation-aware designs is a ring-oscillator with fan-out of 4 (FO4). Typical design specifications are used for constraining the analysis: 10% supply voltage variation is allowed, 3-sigma process variations, and temperature corners at -40°C and 125°C. Also, this section brings a brief description of voltage swings and static-noise margins, which are crucial for evaluating the robustness of an ultra-low voltage design.

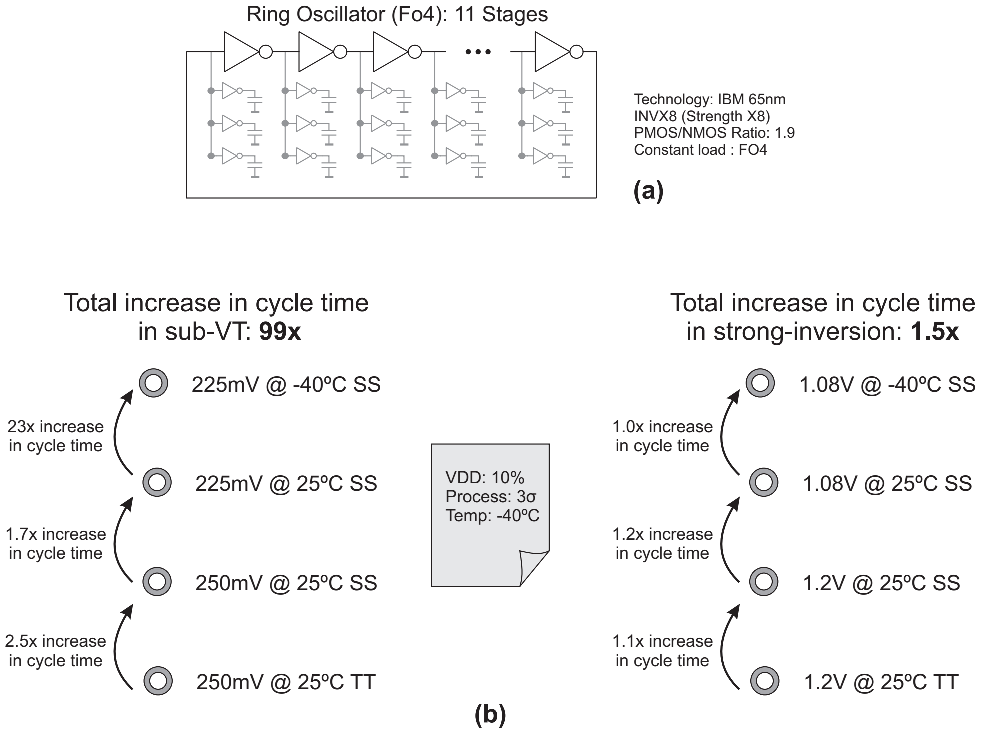
2.4.1 Ring-oscillator FO4

The ring-oscillator depicted in Figure 2.9 (a) is used as study case for evaluating the variability impact. The simulations are performed using a SPICE simulator and a 65nm CMOS Bulk PDK from IBM. The transistor model used is BSIM4. The inverter ratio (W_P/W_N) is tuned for 1.9, resulting in nearly symmetrical rise and fall transitions at 150mV. The inverter size is equal to 8 times the minimum size (strength X8). The total number of stages is 11, a prime number to avoid harmonics during operation. Every oscillator stage has a FO4 load. This whole ring-oscillator configuration was thought to be a reasonable approximation of timing paths found in typical VLSI designs. Considering the process and environmental specifications, the circuit was evaluated at the -40°C and 125°C, with 3-sigma process variation, and 10% voltage variation.

Figure 2.9 (b) presents a summary of the timing impact on sub-VT operation at 250mV compared to strong inversion at 1.2V. The analysis starts on the bottom of the first column, from typical process conditions (TT), typical voltage for this sub-VT analysis (250mV), and room temperature (25°C). Only by assuming 3-sigma process variation, the cycle-time increases 2.5X. Another 1.7X for 10% voltage reduction, and finally, 23X for the -40°C temperature corner. Thus the sub-VT design should be aware that a 99X timing deviation is indeed possible. The strong-inversion design, on the other hand, faces much smaller timing variations. The total increase in cycle-time for strong-inversion is 1.5X. Even though the variation in strong inversion is much smaller than sub-VT, it still shows a 50% impact on typical circuit timing for this 65nm PDK.

Figure 2.9 (b) shows that the temperature variation has a dramatical impact on timing at sub-VT voltages. In Figure 2.10, a detailed view on this temperature impact is provided for the studied ring-oscillator. A 620X timing variation at 150mV is observed,

Figure 2.9: Design corners performance evaluation for sub-VT and strong-inversion. (a) Ring-oscillator with 11 stages and fan-out of 4; (b) Cycle time increase for each design corner considered.


Ring Oscillator (Fo4): 11 Stages
Technology: IBM 65nm
INVX8 (Strength X8)
PMOS/NMOS Ratio: 1.9
Constant load : FO4
(a)
Total increase in cycle time in sub-VT: 99x
225mV @ -40ºC SS
23x increase in cycle time
225mV @ 25ºC SS
1.7x increase in cycle time
250mV @ 25ºC SS
2.5x increase in cycle time
250mV @ 25ºC TT
VDD: 10%
Process: 3σ
Temp: -40ºC
Total increase in cycle time in strong-inversion: 1.5x
1.08V @ -40ºC SS
1.0x increase in cycle time
1.08V @ 25ºC SS
1.2x increase in cycle time
1.2V @ 25ºC SS
1.1x increase in cycle time
1.2V @ 25ºC TT
(b)

from 195kHz in the fast case at 125^{o}C to 315Hz in the slow case at -40^{o}C. Figure 2.10 also shows the *temperature inversion* phenomenon, which happens at 1V and swap the fast/slow temperature corners. This inversion point is dictated by two different effects: the decrease of mobility and VT with high temperature. The first slows the circuit down at high temperature, while the second speeds the circuit up at high temperature. The decrease of VT with high temperatures dominates at voltages lower than 1V, while the decrease of mobility with high temperatures dominates at voltages higher than 1V.

Figure 2.11 explores the statistical nature of the timing data extracted from the analyzed circuit. Figure 2.11 shows the timing distribution of the same ring-oscillator operating at 250mV and at 1.2V. This analysis was performed running 5000 Monte Carlo simulations. The horizontal axis was normalized for better visualization. Regarding the 1.2V analysis, it shows a well shaped bell curve, centered at 1.1GHz with 84MHz frequency deviation and σ/μ of 0.07. The 250mV analysis on the other hand, shows the shape of a lognormal distribution, centered at 91kHz, with 38kHz frequency deviation and σ/μ of 0.4.

Figure 2.12 shows a the fast and slow cases for the ring-oscillator frequency, considering -40^{o}C and 125^{o}C for temperature variation and 3-sigma for process variation. The plot clearly delimits a wide timing variability zone, which presents a 1076X increase on cycle time (from slow to fast) at the 150mV voltage corner. Designs operating at the near-VT regime should be aware of this significantly increased timing variability.

Figure 2.10: Ring-oscillator evaluated under different temperature corners. The inset shows the temperature inversion phenomenon.

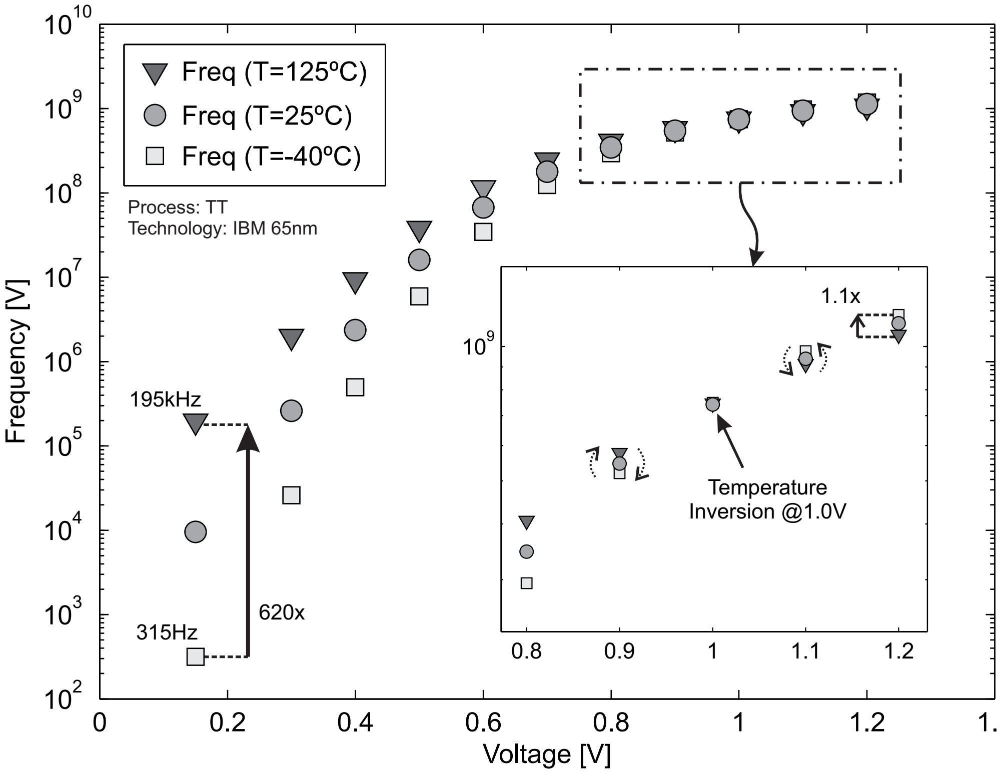


Figure 2.11: Timing variation histogram for the ring-oscillator in two voltage corners, 250mV and 1.2V.

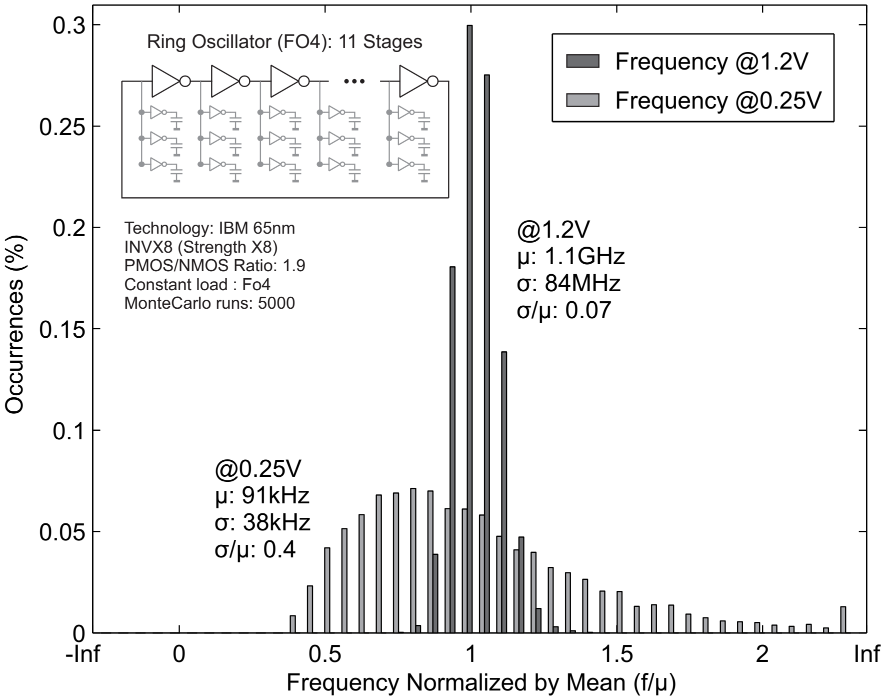
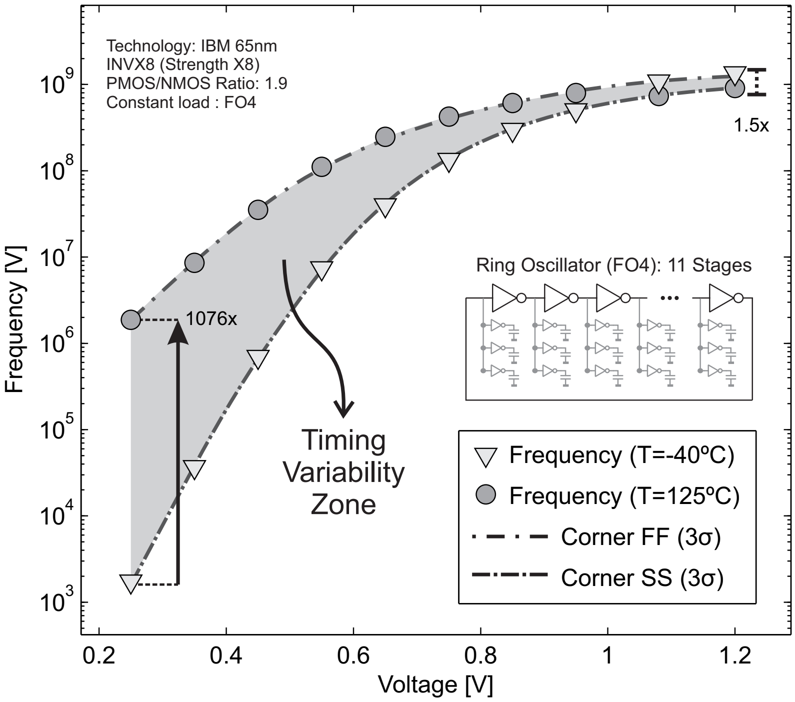


Figure 2.12: Ring-oscillator timing variability zone considering both process and temperature variations.



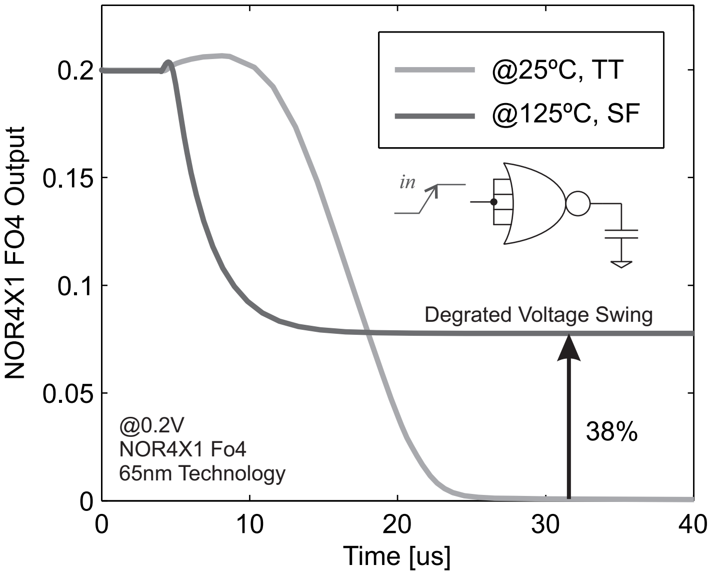
2.4.2 Voltage Swing and Static-noise Margin (SNM)

The design of CMOS digital circuits that operate in ultra-low voltages is very challenging from the robustness point of view. The cells may present different driving strengths for pull-up and pull-down network, causing insufficient *voltage swings* or *degraded static-noise margins* (SNM). Both are very important metrics that ensure cell functionality under the target every operating condition. The SNM is a measure of cell stability, and became popular due to the widely adoption of SRAM cells for cache memory. The voltage swing, on the other hand, relates to the range of output values that are found when a cell is switching.

Figure 2.13 shows the output node of a 4-input NOR gate. The situation depicted in Figure 2.13 refers to a high-to-low transition at supply voltage of 0.2V. The temperature and process conditions differ from one curve to another; the light gray curve runs at 25ºC and typical process (TT), while the dark gray runs at 125ºC and slow NMOS/fast PMOS (SF) process. One may note that the light gray presents a *full* voltage swing, from supply voltage to zero, while the dark gray presents a very degraded zero logic value. The degraded zero value is called a *weak zero,* and occurs due to high static current from pull-up network allied to the low driving current from the pull-down network. In addition to the process conditions that enabled this limited voltage swing, note that the temperature is also higher than room temperature – increasing even more the static currents in off transistors.

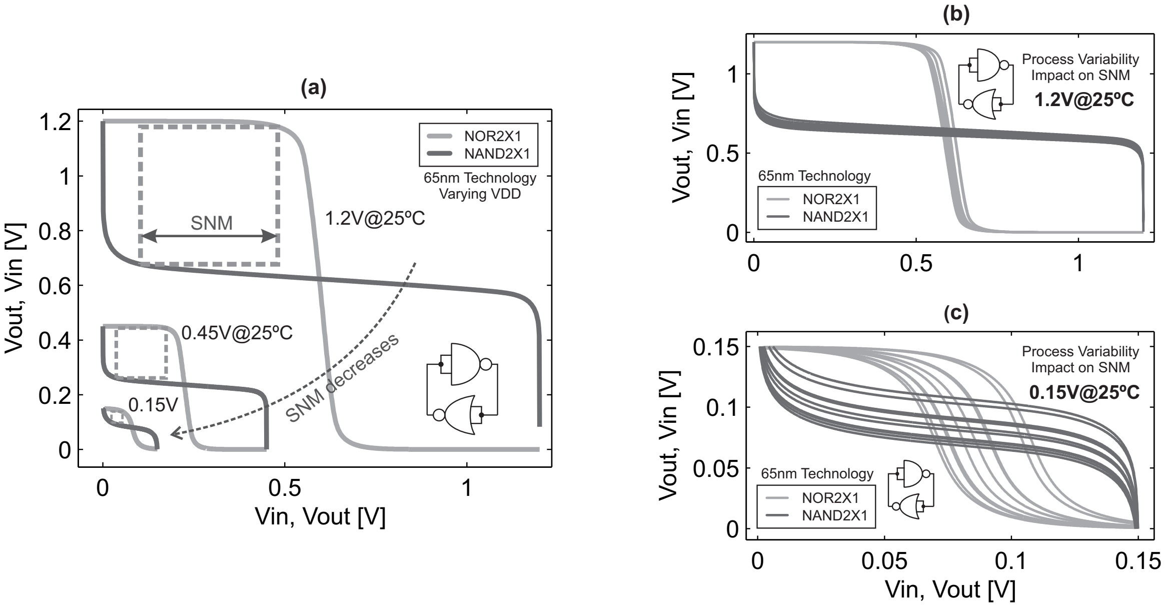
Regarding the SNM measure, it was shown in (LOHSTROH; SEEVINCK; GROOT, 1983) that two back-to-back gates represent the maximum noise that can be applied before failure to an infinitely long chain composed of the same two gates alternatedly. For a pragmatical definition of SNM, Figure 2.14 (a) shows it as the side length of the largest

Figure 2.13: Voltage swing degradation with process variation and high temperature. The pull-up static current generates a weak zero at the output.



inscribed square fitting the worst-case side of the butterfly plot. This thesis adopt the method proposed in (SEEVINCK; LIST; LOHSTROH, 1987) for measuring the SNM using a simulation based approach. Note in Figure 2.14 (a) that the absolute value of SNM decreases at low voltages. Figures 2.14 (b) and (c) show the effects of process variability into the butterfly plots, and consequently on the SNM. Analogous to the voltage swing, the major factor that degrades the SNM is a not well balanced pull-up and pull-down networks. Therefore, the data presented in Figure 2.14 refers not to the basic back-to-back inverter configuration, but a more complex configuration using a 2-input NOR and NAND cells. The two stacked PMOS of the NOR gate and two stacked NMOS of the NAND provide a worst-case analysis with respect to cell selection for the SNM metric.

Figure 2.14: Voltage transfer curves (VTC) and static-noise margins (SNM) plots, (a) multiple voltage SNM; (b) SNM variability at 1.2V; and (c) SNM variability at 0.15V.



3 ULTRA-LOW POWER TECHNIQUES FOR DIGITAL CMOS CIRCUITS

The design of ultra-low power circuits requires power optimizations in all levels, i.e, process, circuit, micro-architectural, and system level. The high level optimizations present large energy impact, but are very specific for a certain case or application. This chapter seeks to provide background information on low-power circuit level design techniques, which are not application dependent and may be applied to a large number of designs. The chapter begins with a brief explanation of the main power components, and follows with several circuit level techniques for low-power digital CMOS circuit design, including clock gating, power gating, operand isolation, multi-VT devices, and finally, voltage-frequency scaling.

3.1 Power and Energy Components

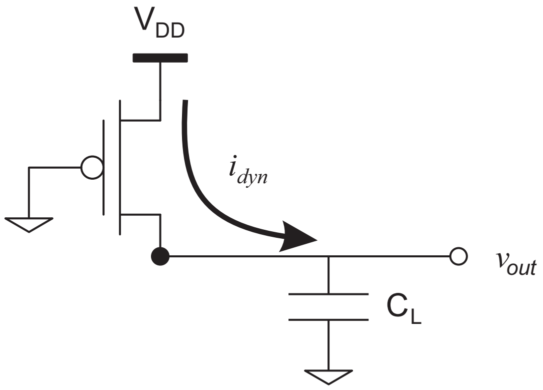
The power and energy are mainly classified into three different categories: dynamic, short-circuit, and static. Each one of them refers to a different kind of energy consumption that helps in identifying how the power profile of a VLSI design can be improved. This section covers each one of those energy components, providing analytical models when it is suitable. This master thesis follows the standard presented in this section for referring to energy and power.

3.1.1 Dynamic Energy

Also known as *switching energy*, the dynamic energy is draw from the power supply each time the capacitor in Figure 3.1 is charged (zero to one transition). Current i_{dyn} passes through the PMOS transistor, dissipating part of the energy as heat, and storing the remaining energy at the capacitor C_L. When the capacitor C_L discharges (one to zero transition), the energy stored in the capacitor is then dissipated into the NMOS transistor, generating a current flow back to the ground rail.

For modeling the dynamic energy drawn from the power supply during the zero to one transition at the output node, we first make the assumption of zero slope time. Then, the circuit shown in Figure 3.1 is valid. The energy is derived from the integration of the instantaneous power, resulting in Equation 3.1. Note that only part of this energy will be stored into the capacitor C_L, and the other part will be dissipated into the PMOS device during the capacitor charging process. For estimating how much energy will stored into the capacitor, we integrate the instantaneous power with respect to the output voltage v_{out}. Equation 3.2 shows that half of the total dynamic power will be stored into C_L, and this energy will be later dissipated into NMOS transistor during the one to zero discharge.

Figure 3.1: Dynamic power equivalent circuit for zero slope times.



Source: Modified from (RABAEY; CHANDRAKASAN; NIKOLIC, 2003).

Thus, assuming zero slope time, the total dynamic energy E_{DYN} drawn from power supply during zero to one transition is given by,

$$\begin{aligned} E_{DYN} &= \int_0^\infty i_{dyn}V_{DD}dt = V_{DD}\int_0^\infty C_L\frac{dv_{out}}{dt}dt \\ &= C_L V_{DD}\int_0^{V_{DD}} dv_{out} = C_L V_{DD}^2 \end{aligned} \tag{3.1}$$

While the energy E_C in fact stored into the capacitor is

$$\begin{aligned} E_C &= \int_0^\infty i_{dyn}v_{out}dt = \int_0^\infty C_L\frac{dv_{out}}{v}_{out}dtdt \\ &= C_L\int_0^{V_{DD}} v_{out}dv_{out} = \frac{C_L V_{DD}^2}{2} \end{aligned} \tag{3.2}$$

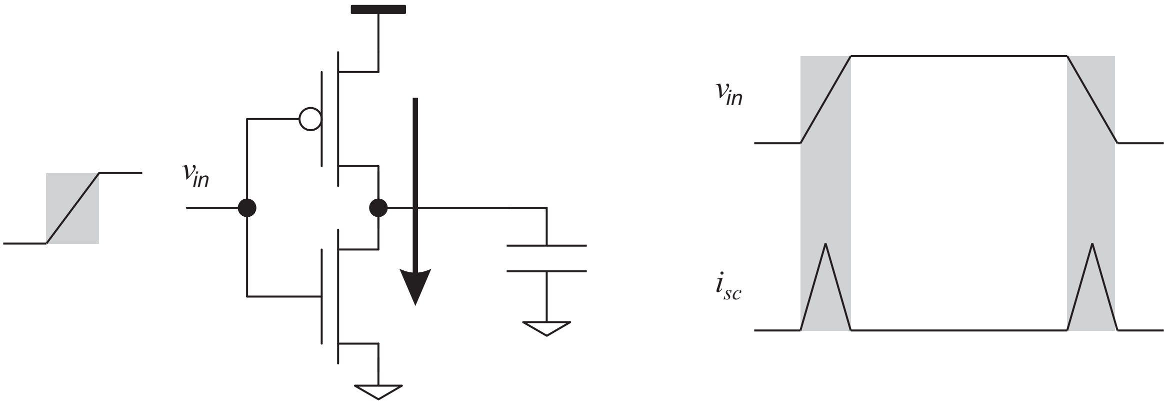
The dynamic power P_{DYN} component of a circuit shall be estimated simply by multiplying the dynamic energy E_{DYN} and the frequency $f_{0\to1}$ in which the cell change state from zero to one. In a large VLSI design however, a cell rarely changes its state as fast as the system clock, thus it is common in literature the adoption of an α term that accounts for logic cell activity. From Equation 3.3, one may conclude that the dynamic power varies quadractically with voltage and linearly with frequency.

$$P_{DYN} = C_L V_{DD}^2 f_{0\to1} \tag{3.3}$$

3.1.2 Short-circuit Energy

It is well known the actual circuit slopes are not as ideal a step function. The slope at voltage input v_{in} causes a direct current i_{sc} to flow from power supply to ground while the input switches. This is due to the middle way value of the gate to source voltage on both PMOS and NMOS devices. The energy consumed through the direct current is known as short-circuit energy. It is directly proportional to the slope time at the input. Figure 3.2 shows that short-circuit power is dissipated during both the rise and fall transitions of the

Figure 3.2: Short-circuit energy due to direct current with non-zero input slopes.



Source: Modified from (RABAEY; CHANDRAKASAN; NIKOLIC, 2003).

input voltage. Several works have tried to model short-circuit power, however the results are overly complicated and are out of scope of thesis (VEMURU; SCHEINBERG, 1994; BISDOUNIS; NIKOLAIDIS; LOUFOPAVLOU, 1998; NOSE; SAKURAI, 2000).

3.1.3 Static Energy

All current that flows from the VDD rail to the ground in the absence of a transition at the input is known as static current. Thus the static power dissipation P_{STA} may be calculated by Equation 3.4. The static current i_{sta} occurs through the following mechanisms:

- **Gate oxide tunneling:** electrons can tunnel across the gate oxide. The probability increases exponentially with oxide thickness, becoming relevant for advanced process nodes.

- **Junction leakage:** the diffusion regions together with wells form reversed biased diodes, which are subject to reverse current as well. These currents are induced by thermally generated carriers, thus increasing exponentially with temperature.

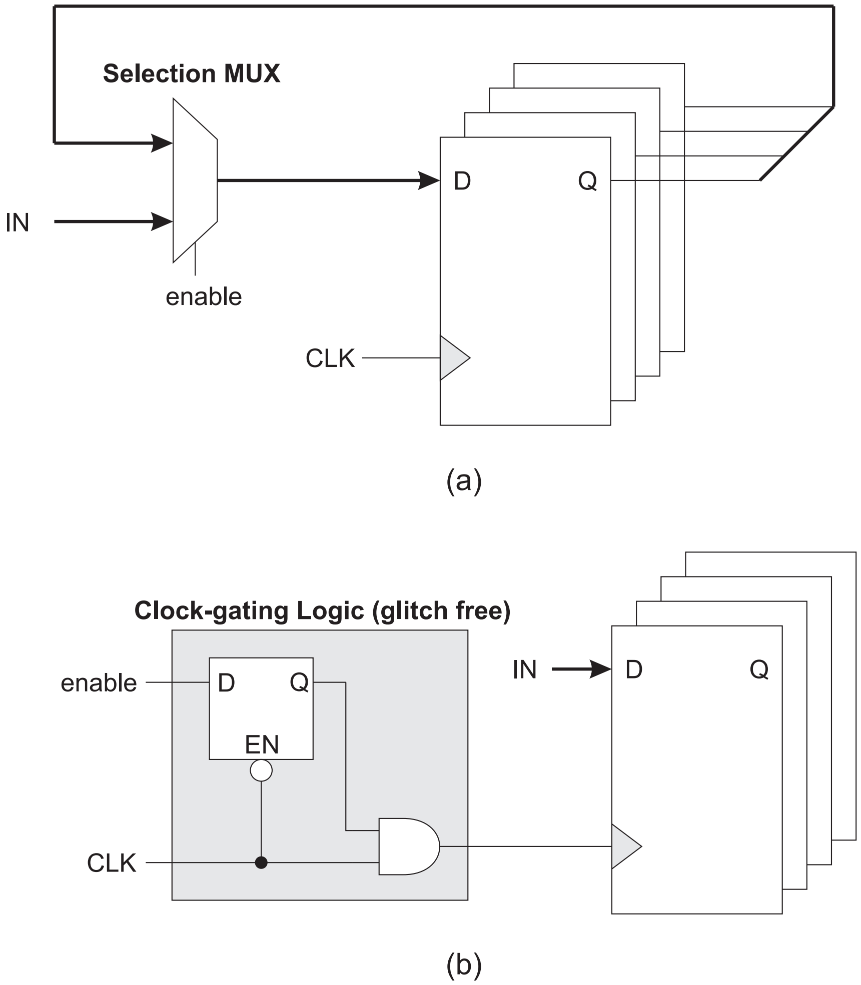
- **Sub-VT conduction:** MOSFET devices under sub-VT bias present an exponential decrease of current with V_{GS} reduction. This exponential relation ensures that a current, orders of magnitude smaller than above-VT currents, still persists between source and drain at V_{GS} below threshold voltage.

$$P_{STA} = i_{sta} V_{DD} \tag{3.4}$$

3.2 Circuit-level Techniques

This section brings a brief overview of circuit level energy-saving techniques, including clock gating, power gating, operand isolation, multi-VT devices, and dynamic VFS. All techniques presented here require additional steps in the circuit design flow. They are applicable not only to a certain specialized application, but most of VLSI designs.

Figure 3.3: Clock gating technique poor saving energy at the clock tree; (a) group of registers without clock gating, (b) group of registers with gated clock.



Source: Modified from (CADENCE, 2011).

3.2.1 Clock Gating

Most of the registers in a complex VLSI design load new data very infrequently, much less frequently than the system clock period. Even though, clock toggles every cycle and dissipates dynamic power. The insertion of an additional control logic to gate the clock of *group of registers* enabled by the same control signal may significantly decrease clock energy consumption.

Figure 3.3 (a) shows the control logic where the clock signal drives the whole group of registers, dissipating supply energy every rising edge of the clock. In Figure 3.3 (b) the clock gating control logic is inserted, and the clock signal no longer drives the whole group of registers. They are now driven by a *gated clock* through the enable signal. The latch is inserted only for glitch suppression purposes.

It is important to notice that clock gating is not an attractive technique for very high activity registers, since the clock would be rarely gated. In those cases, the insertion of additional logic may cause an increase in the overall design energy. Only the infrequently used registers should benefit from the clock gating technique.

3.2.2 Operand Isolation

The operand isolation technique seeks to reduce power dissipated in combinational logic blocks controlled by an enable signal. This is accomplished through the insertion of additional control logic that inhibits the propagation of unnecessary glitches and combinational data through the datapath. In Figure 3.4 (a) a MUX selects the data values to be stored into Register C. Whenever the enable signal selects data from Register B, unnecessary data propagation occurs on the multiplier path. Figure 3.4 (b) shows the operand isolation logic which basically gates the datapath inputs, according to the enable signal. The wasted power at the additional logic might not be significant compared to the power saved at the multiplier. Note that the clock gating technique can not be employed in this scenario, since Register C always loads new data, either from Register B or from the multiplier.

3.2.3 Power Gating

The power gating technique is employed for reducing static energy consumption. Prior to the insertion of high-k dielectric materials, the static current in VLSI designs was 35% in 90nm and 55% in 65nm processes (CADENCE, 2011). The power gating technique basically shuts off portions of the chip when they are not being used. To accomplish such a capability, the portions of the design to be shut off are powered by a local supply rail, which is connected to the global supply rail through one or more large transistors, also known as power gates. When those transistors are in off state, the static energy consumption of the block reduces significantly, and no switching activity propagates through the block.

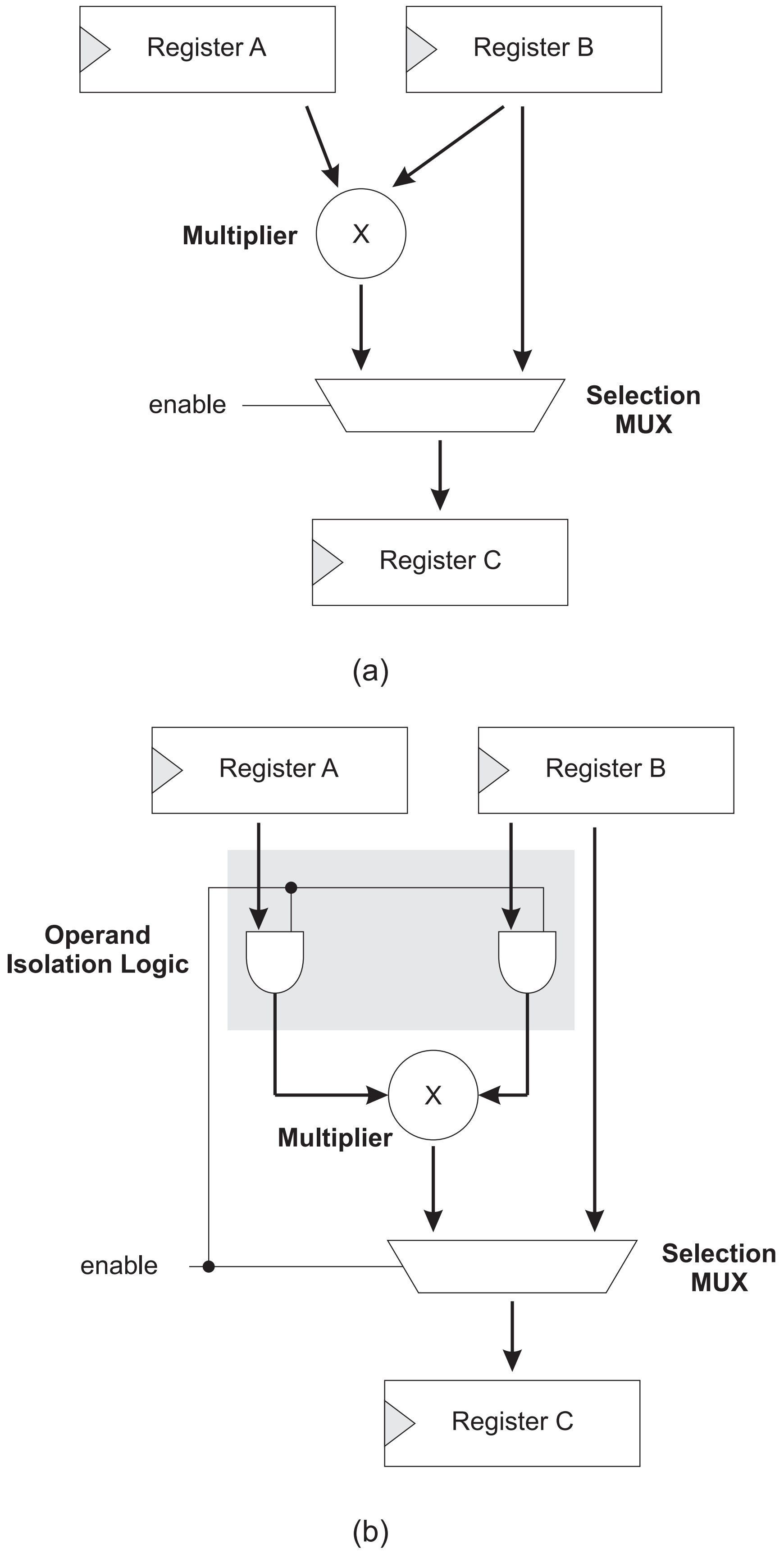
Figure 3.5 shows two different methodologies for applying power gating in a design. In Figure 3.5 (a) coarse-grain sleep transistors are used to control the power to a large block of logic, while Figure 3.5 (b) shows the use of fine-grain sleep transistors integrated into every standard-cell design. Note that both techniques address the static energy consumption through the insertion of two series transistors, which will indeed degradate performance when the circuit is operating at full throughput mode. The two methodologies for power-gating implementation have benefits and drawbacks, which are listed in Table 3.1.

Table 3.1: Power gating methodologies comparison.

	COARSE-GRAIN	FINE-GRAIN
Static current control	*High*	*Low*
Area cost	*Low*	*High*
SNM degradation	*High*	*Low*
Timing accuracy	*Low*	*High*
Full-custom	*Yes*	*No*
Library support	*No*	*Yes*

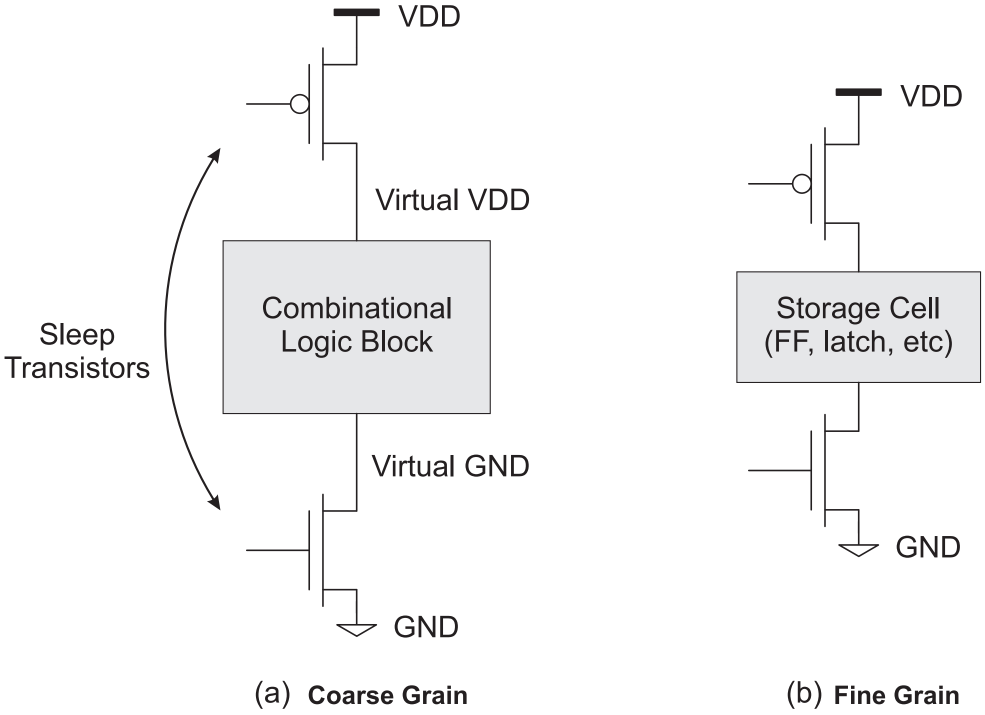
As summarized into Table 3.1, the coarse-grain power-gating provides better control over static current due to the possibility of using multiple transistor as power gating devices, i.e., some of the transistors might be on, while others remain off. Fine-grain power gating achieves better results with respect to timing accuracy and SNM, due to the presence of power gating transistors during cell library design and characterization. In terms

Figure 3.4: Operand isolation technique for energy-saving at combinational logic blocks.



Source: Modified from (CADENCE, 2011).

Figure 3.5: Power gating technique for reducing the static power consumption, with (a) coarse-grain and (b) fine-grain methodologies.



Source: Modified from (CADENCE, 2011).

of area, coarse grain power gating has better results, but requires a much more layout and full-custom based implementation, while fine-grain only requires the appropriated library support – thus, enabling a fully automated process.

Regarding the recovery of the last logic state, the two methodologies have different approaches. The coarse-grain logic uses additional latches connected to auxiliary supply rails to store input variables state. The fine-grain logic achieves full recovery through the use of especially designed state retention storage elements that are also connected to an auxiliary supply rail.

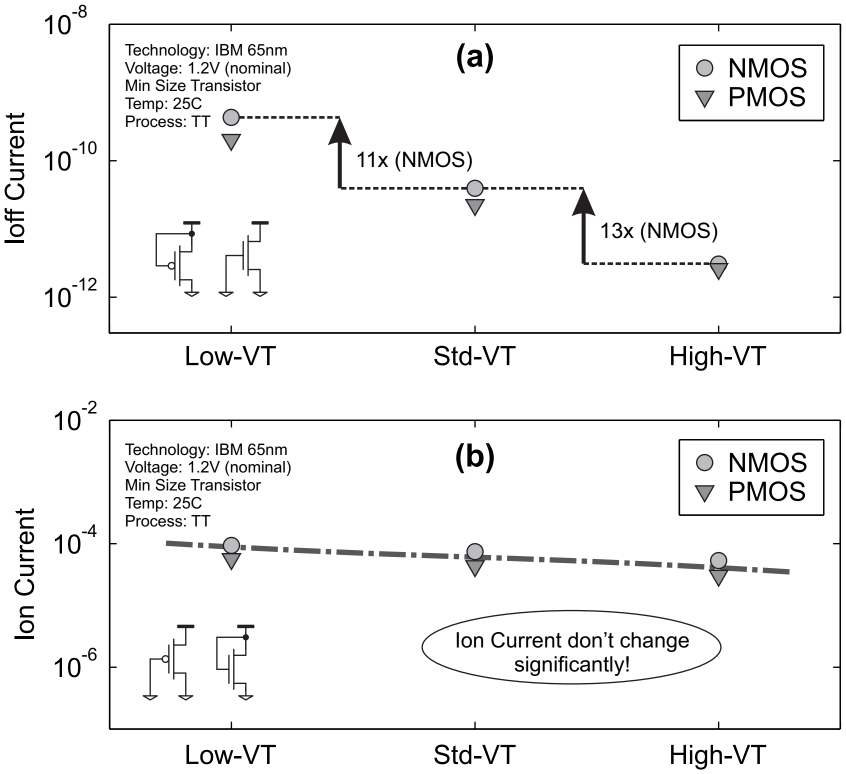
3.2.4 Multi-VT Devices

Some processes offer the possibility of multi-VT MOSFETs. These devices, when offered, usually come into three different categories: low-VT, standard-VT, and high-VT. As the name implies, low-VT devices present lower threshold voltage, consequently, lower latency and higher static energy consumption. High-VT devices, on the other hand, present high threshold voltage, thus higher latency but lower static energy consumption. Standard-VT is in the middle, with balanced performance and static energy consumption.

The addition of these multi-VT devices increases the processing steps and mask cost during manufacturing, but enables the design of libraries with multi-VT cells. Such a multi-VT cell-library provides the capability of power-performance optimization at circuit level, where cells with low static power can be used in non-critical timing paths, while low latency cells are placed in critical timing paths.

Table 3.2 lists the threshold voltage values for each of the devices offered in the 65nm CMOS Bulk PDK, extracted from (IBM, 2009). Figure 3.6 presents the device charac-

Figure 3.6: Multi-VT device electrical characteristics for (a) on current and (b) off current.



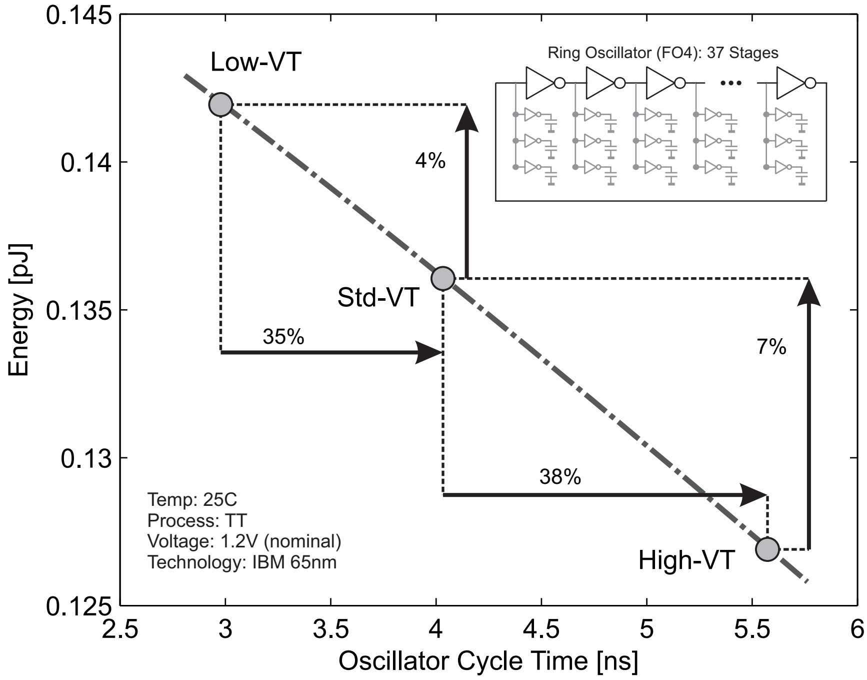
teristics from a circuit designer perspective. In Figure 3.6 (a) the off current is shown to change one order of magnitude between each device, resulting in two orders of magnitude less static current for the high-VT compared to the low-VT device. The on current, shown in Figure 3.6 (b), remains nearly the same for the three different devices. The analyses were performed at typical process (TT), nominal voltage of 1.2V, and room temperature at 25^{o}C.

Table 3.2: Threshold voltage for minimum device size in 65nm CMOS Bulk PDK from IBM. Values extracted from (IBM, 2009).

	NMOS	PMOS
Low-VT	*270mV*	*-280mV*
Std-VT	*428mV*	*-400mV*
High-VT	*585mV*	*-587mV*

Regarding an energy-speed analysis, Figure 3.7 shows the energy-speed characteristic of a 37 stages ring oscillator at typical process (TT), nominal voltage of 1.2V and room temperature at 25^{o}C. Three different implementations were evaluates, varying only the device type, low-VT, standard-VT, and high-VT. The plot shows that performance with lower VT devices improves significantly from one VT to another. Note that the energy data however does not change significantly due to high activity nodes in the ring oscillator. Thus, the energy-savings associated with multi-VT devices are bound to very low activity circuits.

Figure 3.7: Energy-speed improvements though multi-VT devices applied to a 37-stages ring-oscillator.



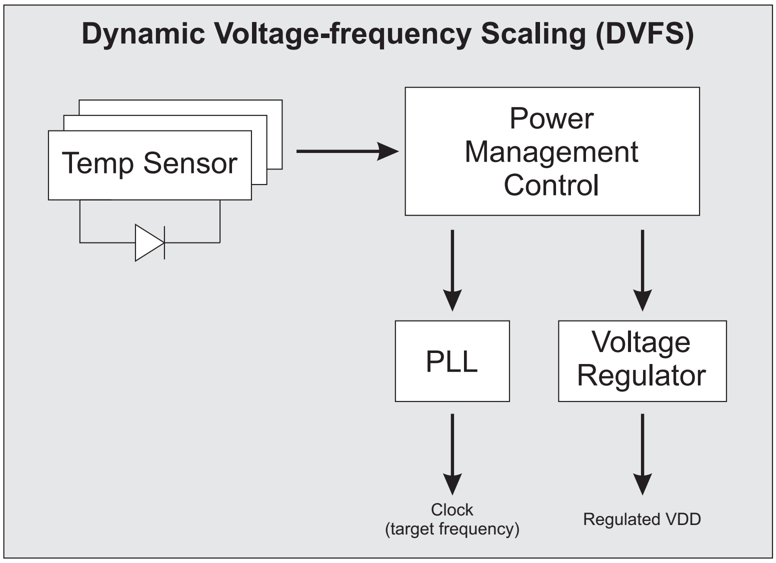
3.2.5 Voltage-Frequency Scaling (VFS)

The voltage-frequency scaling (VFS) technique reduces energy by decreasing chip voltage and frequency as well. Its foundation relies over the bursty nature of most workloads found in battery operated devices, i.e. smartphones, portable computers, wearable devices, ubiquitous sensors, etc. The bursty workload requires maximum performance for a very short period of time, remaining in a low activity mode most of the time. This scenario opens up an opportunity for dynamic VFS, i.e. to change design voltage and frequency in runtime according to workload requirements.

Figure 3.8 shows the basic building blocks for supporting dynamic VFS. The block responsible for controlling the DVFS is named Power Management Control (PMC). It may have inputs from across the software layer, and from several temperature sensors spread over the chip area. When a new voltage/frequency value should be set, the PMC changes first the frequency at the Phase Locked Loop (PLL), and then the supply voltage at the Voltage Regulator (VR).

Despite of the PMC, adjustable PLL and VR, the DVFS technique also requires a cell-library which supports operation on the scaled supply voltages. In other words, the cells should have suitable static-noise margins and voltage swings for the whole range of supply voltage values being scaled. In addition to simply work at low voltage, a cell-library must tolerate process and environmental variability at low voltages as well. Chapter 2 clearly demonstrates that variability has dramatically increased impact in circuits operating at low voltages, thus requiring special design effort for mitigating those effects.

Figure 3.8: Dynamic voltage-frequency scaling block diagram for support hardware.



4 NEAR-VT CELL-LIBRARY FOR DYNAMIC VFS

State of the art chip design in decananometer CMOS may easily reach hundreds of million transistors in a single die. The design of such highly integrated VLSI circuits requires several levels of abstraction. In a traditional digital design flow, the standard-cells may be considered the bridge between logic and layout, separating the discrete digital from the continuous analog domain. Cell-libraries encapsulate logic functions into very small layout pieces. These regularly designed layouts implement a logic function with corresponding inputs and outputs, thus providing the required design abstraction for higher layers. Another advantage of this very block-based approach refers to the automatic placement and routing, which is tremendously simplified due to layout regularity. This Chapter covers the near-VT cell-library design topics with respect to transistor sizing for ultra-low voltages. Variability issues that affect SNMs and voltage swings are also evaluated.

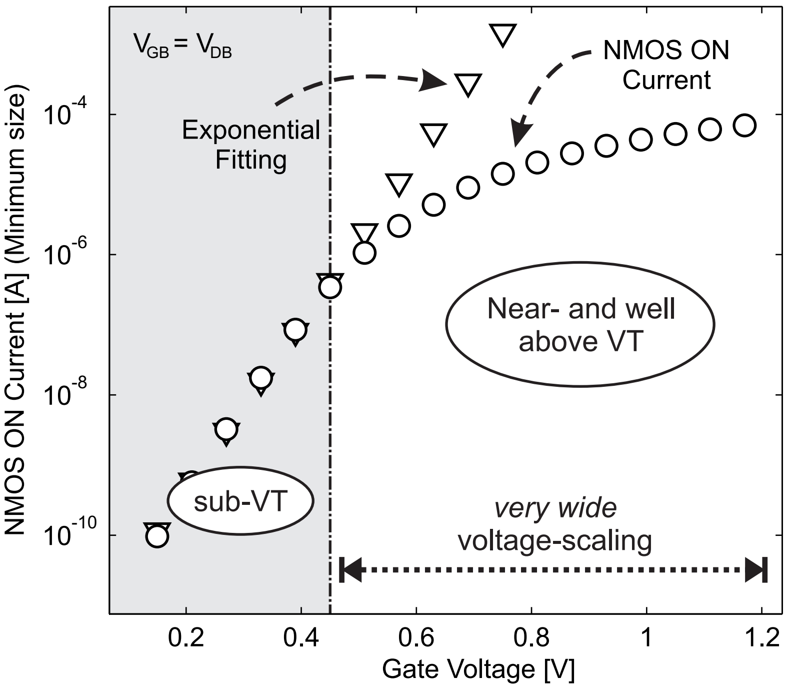
4.1 Cell-design for Near-VT Regime

Traditional dynamic VFS techniques usually does not require cell re-design, since they explore a very narrow voltage range still in strong inversion. For achieving best-in-class energy-savings, we propose to extend the VFS bounds down to the near-VT regime, defining this new range as *very wide*. This very wide voltage range, from nominal voltage down to the near-VT, does require a complete cell re-design for improved robustness at low-voltage, mitigating effects such as reduced SNMs, voltage swings and high static currents. This Section first presents a definition of near-VT regime based on the MOSFET current behavior. Follows our developed cell-library, together with the proposed transistor sizing methodology.

4.1.1 Near-threshold and *very wide* VFS

Near-VT operation is a widely adopted term in the literature; however, it still lacks a solid definition for the boundary between sub- and near-VT. In this thesis, we explore the logarithmic MOSFET sub-VT current to generate an appropriate definition for this boundary, which may be thought as the corresponding voltage point where the sub-VT NMOS current substantially deviates from an exponential fitting. The NMOS current is used due to its slightly higher absolute threshold voltage in our PDK. Figure 4.1 shows the exponential fitting and the boundary between sub- and near-VT for the exercised 65nm PDK and standard-VT devices. The *very wide* dynamic range of VFS is also marked on Figure 4.1, and goes from voltages well-above VT down to the sub-/near-VT boundary located at 450mV for this technology.

Figure 4.1: Sub- and near-VT boundary definition based on the exponential fitting of a minimum sized saturated NMOS ON current (standard-VT at 25°C). For this 65nm PDK, the boundary mark is $V_{GB} = V_{DB}$ at 450mV.

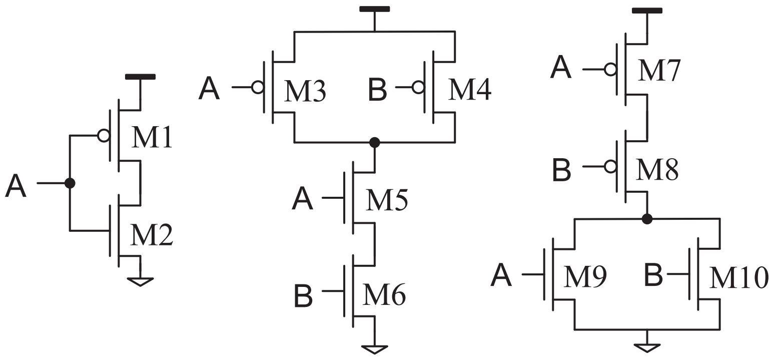


4.1.2 Digital CMOS Sizing at Near-VT

Sutherland (SUTHERLAND; SPROULL, 1991) introduced the well-known logical-effort method to optimize CMOS digital circuits for speed by tuning transistor widths. The logical-effort method actually defines the *logical-effort* quantity for a simple inverter as 1, while all other logic gates have a logical-effort value greater than 1. This value represents how weak the corresponding gate is at driving current than an inverter (given an equivalent amount of input capacitance). From the cell topology point-of-view, the logical-effort method also assumes that the effective width of a n-transistor stack is $1/n$. Regarding accuracy, the logical-effort method was proved to be accurate for hand calculation at supply voltages well-above VT. However, in deeply scaled supply voltages, the MOSFET current is an exponential function of its terminal voltages, thus stacked transistor no longer hold the previous $1/n$ relationship.

Keane (KEANE et al., 2006) derived a closed form solution for the sizing of transistor stacks in sub-VT regime. It was shown that the optimal sizing includes different device widths for each transistors in the same stack, i.e., the transistors closest to the supply rail should be sized up compared to others on the same stack. Even though the theoretical work showed some interesting results, a simplified solution with a simulation-based unique width for all transistors on the same stack is adopted in (KEANE et al., 2006). The sizing method proposed by Keane (KEANE et al., 2006) is based on the DC driving current of CMOS logic gates, where the widths are selected in order to balance the PMOS and NMOS DC currents. Additionally, the inverter ratio is reused to calculate the sizing of other topologies, which does not allow any logic gate to have smaller DC driving current than the reference inverter.

Figure 4.2: Digital CMOS cell-design parameters for near-VT operation. All dimensions refer to transistor width with minimal length. The transistor stack factor $ALFA$ is individually selected for each cell.



Transistor sizing:

M1=K*R_{INV}	M3=K	M7=K*(1+$ALFA_{NR}$)
M2=K	M4=K	M8=K*(1+$ALFA_{NR}$)
	M5=K*(1+$ALFA_{ND}$)	M9=K
	M6=K*(1+$ALFA_{ND}$)	M10=K

Minimum size units; K: Strength factor; R: Ratio Wp/Wn; ALFA: Stack factor

In this work we propose a slightly different approach to size CMOS logic gates for near-VT operation:

- Instead of DC driving currents for NMOS and PMOS, we use directly rise/fall transitions to tune widths and maximize static-noise margins.
- The CMOS logic gates are allowed to have any possible driving strength as long as they keep balanced rise/fall transition times.
- The cell-design should be simply multiplied by a constant K for generating higher strength logic gates.

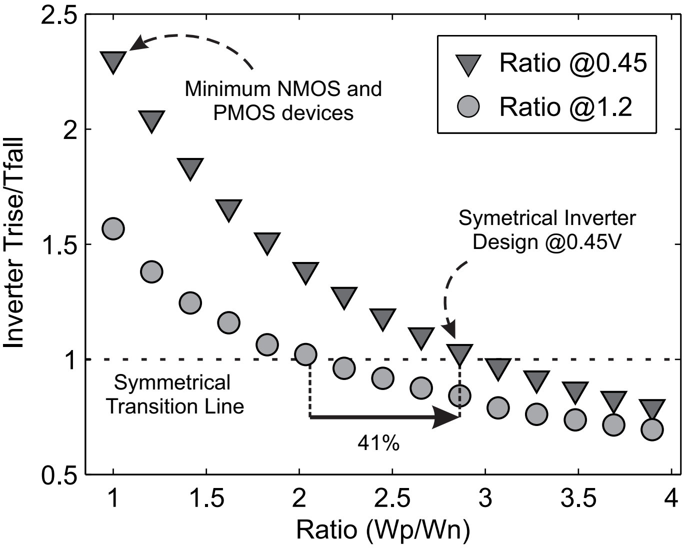
Figure 4.2 shows the sizing rules to the logic-cells that compose the developed near-VT digital cell-library. All values are in minimum size units. The constant K defines cell strength starting from the minimal which is one, and the ratio R_{INV} defines the balanced W_P/W_N for the simple inverter. Regarding stacked transistors, the sizing factor $ALFA$ is defined individually for each cell topology. The complete set of cells that compose our near-VT library is shown in Table 4.1, with a total of 17 cells. The corresponding strengths in which each cell is available are marked with 'X'. Note that the library includes two different registers DFFS, and DFFR – with set and reset inputs, respectively.

The developed library does not target an specific operating voltage, but a very wide range of voltages, that goes from nominal supply voltage down to near-VT operation, as shown in Figure 4.1. Figure 4.3 shows two curves (at 0.45V and 1.2V) of a minimal strength inverter rise/fall transition time versus W_P/W_N ratio, also referred to as R_{INV}. The symmetrical inverter ratio, i.e., equal rise and fall time which maximizes SNM, differs from one supply voltage to another in 41%. The symmetric design, which was certainly

Table 4.1: Cells included on near-VT library.

Cell	X1	X2	X3	X4	X8
INV	X	X	X	X	X
NAND2	X	X		X	
NOR2	X	X		X	
DFFR	X	X		X	
DFFS	X	X		X	

Figure 4.3: Rise/Fall transition times for inverter cell versus ratio R_{INV}.

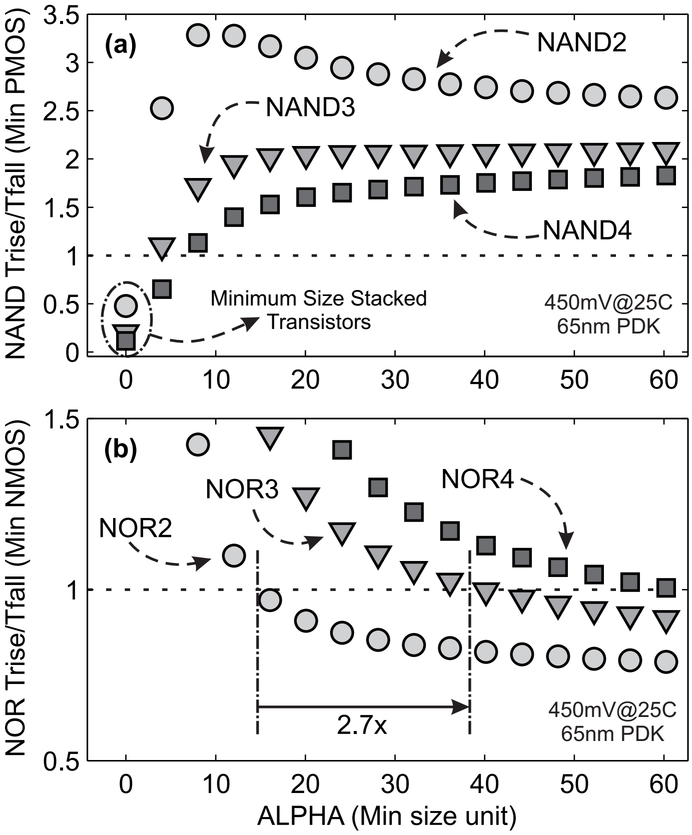


the best choice once it maximizes the SNM, depends on operating voltage, making it impossible to maximize SNM for the whole very wide voltage range. The cell-library has to be inevitably sized for *worst-case voltage conditions*. Thus, our cell-library was sized at 450mV (sub-/near-VT boundary), which ensures the correct operation from 450mV up nominal voltage. Even though the SNM at strong inversion might be non-optimal, it will still present acceptable values, while enabling a very wide VFS down to the near-VT regime which holds much more strict sizing guidelines. Thus, all balancing of rise/fall transition times and associated cell optimizations exercised for this library have been performed at 450mV, 25°C.

Regarding the cells presenting stacked transistors (NAND and NOR), Figure 4.4 (a) shows the simulated rise/fall transition time versus the $ALPHA$ stacking factor (see equations in Figure 4.2). The multiple input NAND cells achieve the desired symmetry with $ALFA$ values smaller than 10. However, Figure 4.4 (b) depicts a 2.7X increase in $ALFA$ from 2-input to 3-input NOR gate, and this number increases even more if 4-input NOR is considered. Such large transistors in the pull-up network, combined with minimum size NMOS devices in the pull-down network may present serious static current issues at high temperature, which can result in the propagation of weak zeros. Since no analysis was performed for different temperature corners, the chosen set of cells limits to a maximum number of two stacked transistors.

One approach for avoiding large transistors at the pull-up network in NOR cells and

Figure 4.4: Rise/Fall transition times for minimal strength gates versus ALFA stacking factor: (a) 2,3,4-input NAND gate; (b) 2,3,4-input NOR gate;

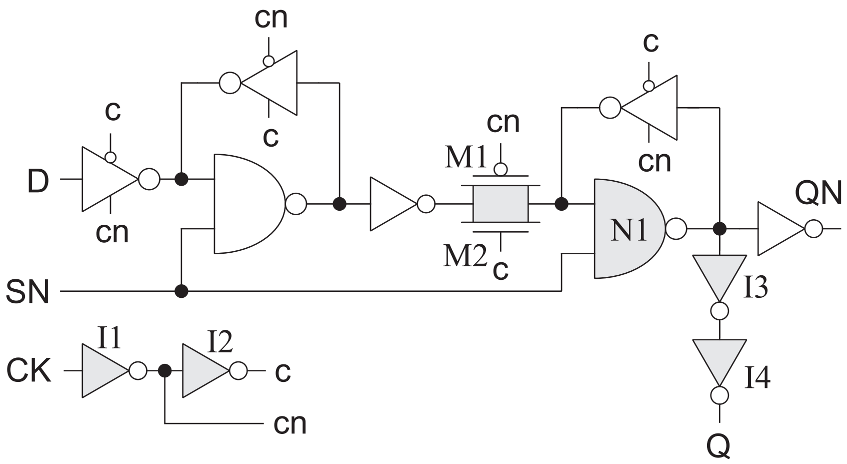


still preserve SNM is to use a larger than minimal channel length at the pull-down network. This will decrease the driving strength of the NMOS transistors, making it easier to match a PMOS width that provides symmetric rise and fall times. Such increase in the NMOS channel length will obviously impact in cell performance, but will ensure its correct functionality. The new PMOS transistor widths for symmetric rise/fall time will be dramatically reduced, resulting in reasonable static current even for multiple input cells at high temperatures.

The library includes two transmission gate master-slave registers, one with active low set (DFFS), and the other with active low reset (DFFR). Figure 4.5 depicts the register micro-architecture, for the DFFS cell – the DFFR is an analogous implementation using a NOR gate instead of NAND. The optimization method used for register sizing is based on (GIACOMOTTO; NEDOVIC; OKLOBDZIJA, 2007), and relies in simulating the design parameters within the allowed range of values. The register shown in Figure 4.5 will have only the gray gates optimized, while all other white gates remain minimum size devices. Note that except for M1 and M2, all gates are logic gates which have been previously optimized. Thus, only the strength factor is being swept during simulation.

The register simulation is focused on Clock-to-Queue delay, assuming the best possible setup time, i.e., the data signal is an initial condition (IC) for the simulation. The energy and delay are averaged for rise and fall output transitions, and the register is always loaded with a fanout-4 (FO4) load. Note that three different optimizations for each register type are required, one for each register strength. Note that I4 is fixed for one of the

Figure 4.5: Master-slave register architecture sizing. White gates are minimum size, while gray gates are optimized through simulation according to the sizing ranges.



Transistor sizing ranges (white gates are minimum size):

M1=[1; I4*3]	I1=FO2	I3=[1;I4]	N1=[1;I4]
M2=[1; I4*1.5]	I2=FO2	I4={1,2,4}	

Units of minimum size; I4 is fixed according to register strength;

target strengths during an entire simulation. The clock drivers I1 and I2 are sized in order to keep a fanout-2 (FO2) delay on the clock signal (ALIOTO; CONSOLI; PALUMBO, 2011).

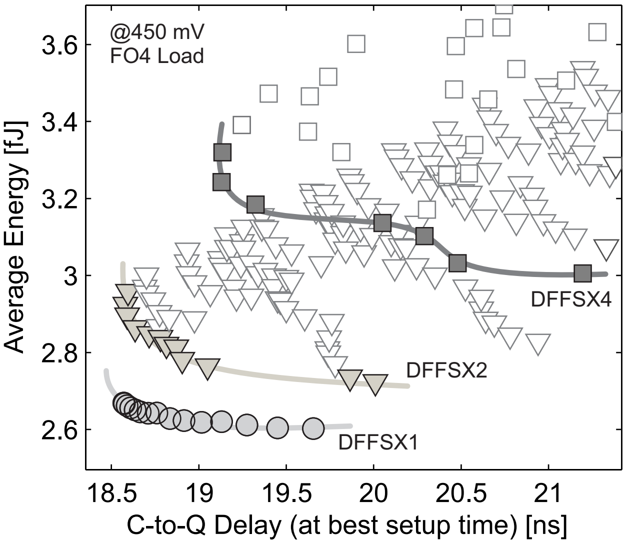
From the near-VT design point-of-view, other works (CALHOUN; CHANDRAKASAN, 2004) have pointed-out that this register micro-architecture fails on sub-VT voltages due to strong leakage currents on M1 and M2 when operating under 3-sigma variability effects. The adopted solution in literature (CALHOUN; CHANDRAKASAN, 2004) is to increase the transmission-gate length, and to make a stronger N1 feedback inverter. Even though adequate, this solution degrades performance at strong inversion. Our experiments however, demonstrate that at 450mV, with our PDK, there is no need for such adjustments. Thus, assuming that the library is not designed to operate on sub-VT supply voltages, no change is required when operating on the very wide range from nominal voltage down to the near-VT regime.

The register optimization results for the DFFS cell at four different strengths are shown in Figure 4.6. The design-space achievable through the proposed sizing ranges is quite large, especially for the X2 and X4 registers. However, only a few set of registers compose the energy-efficient curve, which is marked for each of the register strengths. They represent the best design in terms of energy for the corresponding performance value. Note that every register is loaded with a fan-out of 4, i.e., the higher is the register strength, larger is the load it drives.

4.2 Robustness Analysis

Process and environmental variability is present in every CMOS design. Even at nominal voltage, Figure 2.9 (b) shows that a 3-sigma, 10% supply voltage, and -40ºC results in 50% timing variation for an 11 stages fan-out of 4 ring oscillator. This variability issue

Figure 4.6: Master-slave register design space for DFFS with multiple strengths X1, X2, and X4. The energy-efficient designs for each cell strength are marked.



increases two orders of magnitude when low-voltage operation is required. It is critical thus, for a near-VT designed cell-library, to evaluate the main indicators of cell instability which are SNM and voltage swings. This Section brings such analysis, while discusses the advantages of near-VT over sub-VT operation from cell robustness point of view.

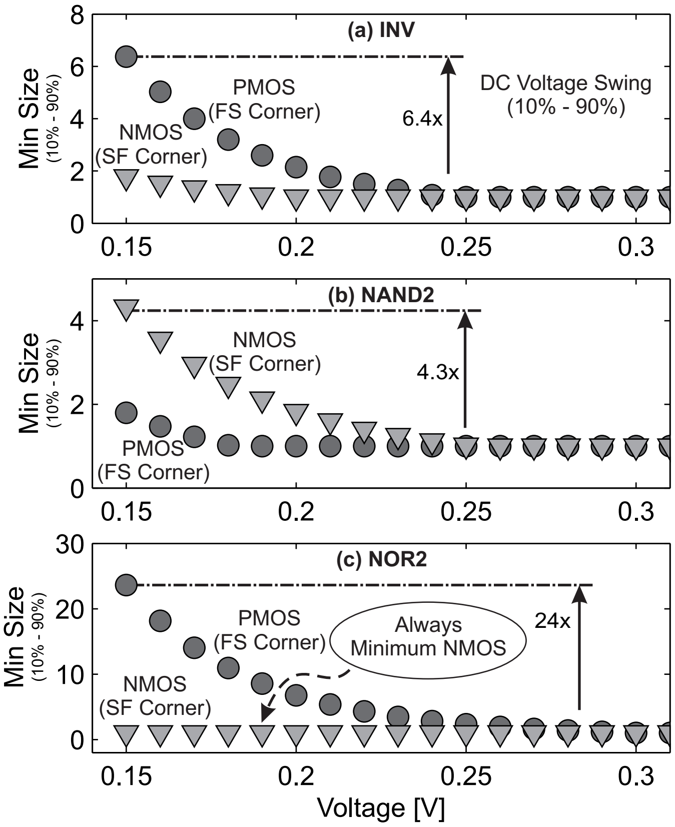
4.2.1 Variability Impact on Voltage Swings

The biggest challenge of the near-VT regime, required for a very wide range of VFS, is to ensure the design robustness at deeply scaled supply voltages. Process variability imposes an enormous challenge for MOSFETs operating in such low-voltages due to the MOSFET exponential dependence on threshold voltage. The smallest threshold deviation, mainly caused by RDF, impacts substantially in transistor driving and static current, affecting the voltage swings and SNM of the cell (ZHAI et al., 2005).

In order to mitigate the voltage swing issues at deeply scaled supply voltages, other works in literature often adopt an universal transistor upsizing (ZHAI et al., 2005; WANG; CHANDRAKASAN, 2005; KWONG; CHANDRAKASAN, 2006; KIM et al., 2007). Large transistor area reduces the VT deviation, according to Equation 2.1, thus reducing the variability effects. Calhoun in (CALHOUN; CHANDRAKASAN, 2004) also shows that the worst corners for digital logic at deeply scaled supply voltages are fast NMOS/slow PMOS (FS) and slow NMOS/fast PMOS (SF). In both cases, the slow transistor needs to be upsized in order to compensate the strong leakage currents from the fast device. As in (WANG; CHANDRAKASAN, 2005), we adopt the DC voltage swing as a measure for evaluating variability effects at different voltage points. Thus, Figure 4.7 shows the minimum sizes (in multiple of W_{MIN}) for PMOS and NMOS transistors to achieve voltage swing at least equal to 10% – 90% for (a) Inverter, (b) 2-input NAND gate, and (c) 2-input NOR gate. Note that when either PMOS or NMOS is varying, the other transistor remains fixed at minimum size.

Figure 4.7 (a) shows that both PMOS and NMOS need to be upsized in order to build a reliable inverter design at voltages below 250mV. The PMOS transistor presents a 6.4X

Figure 4.7: Minimum transistor size (in multiples of W_{MIN}) for achieving 10% to 90% DC voltage swing: (a) Inverter; (b) 2-input NAND; (c) 2-input NOR.



sizing impact at 150mV compared to the minimum transistor size. In (b), the 2-input NAND gate swaps the NMOS and PMOS curve positions. For the NAND design, the NMOS present higher need for upsizing, since it is located into a 2-transistor stack. In (c) the 2-input NOR gate presents a 24X increase in PMOS device size at 150mV. The NOR gate does not require NMOS upsizing due to the low leakage through PMOS transistor stack. Note that for operating in near-threshold regime at 450mV, no universal upsizing is needed. Thus, for very wide range of VFS, minimum width devices achieve adequate voltage swings.

4.2.2 Variability Impact on Static-noise Margins (SNM)

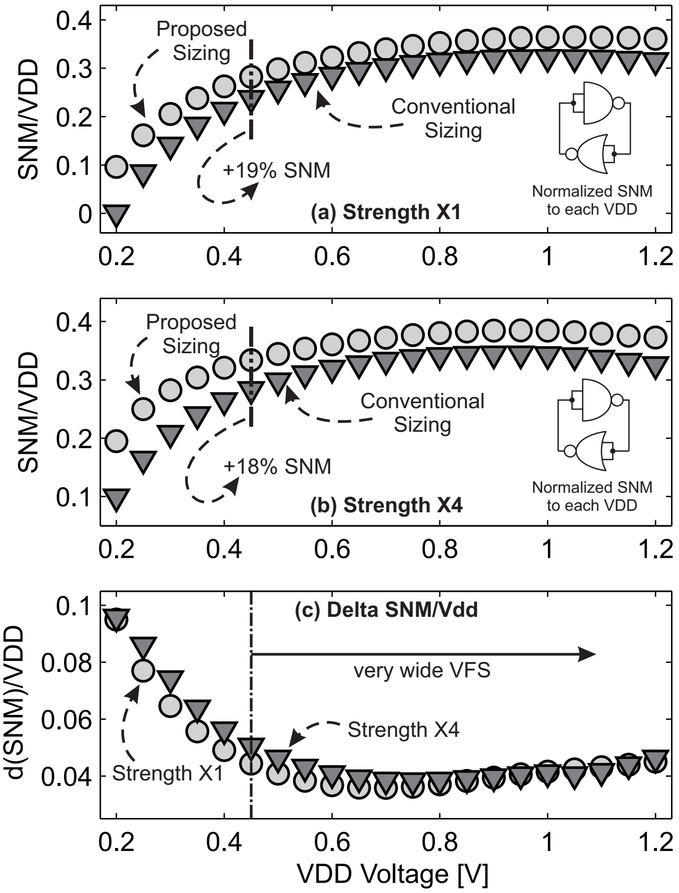
The SRAM cell stability is critical for deeply scaled supply voltage designs. Lohstroh (LOHSTROH; SEEVINCK; GROOT, 1983) showed that two back-to-back gates represent the maximum noise that can be applied before failure to an infinitely long chain composed of the same two gates alternated. In order to measure the SRAM cell stability, we adopt the DC SNM metric based on the side length of the largest inscribed square fitting the worst-case side of the butterfly plot. We exercised a 3-sigma corner-based simulation to the following MOSFET process corners: FF, FS, TT, SF, and SS. Only the worst-case SNM, normalized by the corresponding VDD, is plotted in Figure 4.8. Additionally to our sizing strategy, we also plot a conventional logical-effort based sizing results. The conventional sizing uses W_P/W_N ratio equals 1.5, and adopts the $1/n$ stack correction factor.

Figure 4.8 (a) and (b) both show results for a 2-input NAND/NOR loopback gates with strength X1 and X4, respectively. The SNM shows a dramatically degradation at voltages bellow the sub-/near-VT boundary at 450mV. The adopted sizing strategy shows significant improvements over the conventional sizing, achieving up to 19% improvement at the very wide operating range (from nominal down to the near-threshold). Figure 4.8 (c) presents the relative increment (normalized to VDD) of the SNM when adopting our sizing scheme instead of conventional one, for both strengths X1 and X4. Note that the gains of our sizing strategy increase with voltage reduction in sub-VT range.

4.3 Characterization Methodology

Once all the 17 cells that compose our library have been designed, a characterization process must be run in order to generate an appropriate set of library files. These library files have a table-based form, and list each cell inputs, outputs, timing and power characteristics. The exercised technology in which our library has been developed and characterized is a commercial 65nm PDK (IBM, 2009), using only standard-VT MOSFETs. The layout parasitics are estimated by the SPICE tool, since no cell layout has been made. It is important to mention that characterize a library in such a wide voltage range is not a trivial task. The characterization parameters such as input slopes, output load, simulation end time, and step time need to be set appropriately for each voltage being characterized – since delays and even gate capacitances change according to supply voltage. For generating a complete set of simulated results, the cell library has been characterized from sub-VT voltages starting from 150mV up to 1.2V, in 10mV steps, for three different process conditions, slow (3-sigma SS), fast (3-sigma FF), and typical (3-sigma TT).

Figure 4.8: Static-noise margin (SNM) of a 2-input NAND/NOR SRAM cell (normalized for each VDD). (a) NAND/NOR strength X1; (b) NAND/NOR strength X4; (c) Relative increment of SNM when adopting our sizing instead of the conventional one.



5 ENERGY-SPEED EXPLORATION FOR VERY-WIDE VFS

This chapter presents energy-saving results of our developed cell-library for very-wide range of VFS. The benchmark circuits include a 25kgates notch filter, a 20kgates 8051 compatible core, and 4-combinational/4-sequential ISCAS benchmark circuits (HANSEN; YALCIN; HAYES, 1999; BRGLEZ; BRYAN; KOZMINSKI, 1989). Our proposed approach for dynamic VFS includes a voltage range from well-above VT down to the near-VT regime; however we have extended our cell-library characterization down to 150mV, thus generating comparison results between sub-, near- and well-above VT operation. In addition, a corner based characterization has been performed for the FF, TT, and SS corners, at room temperature (25^{o}C).

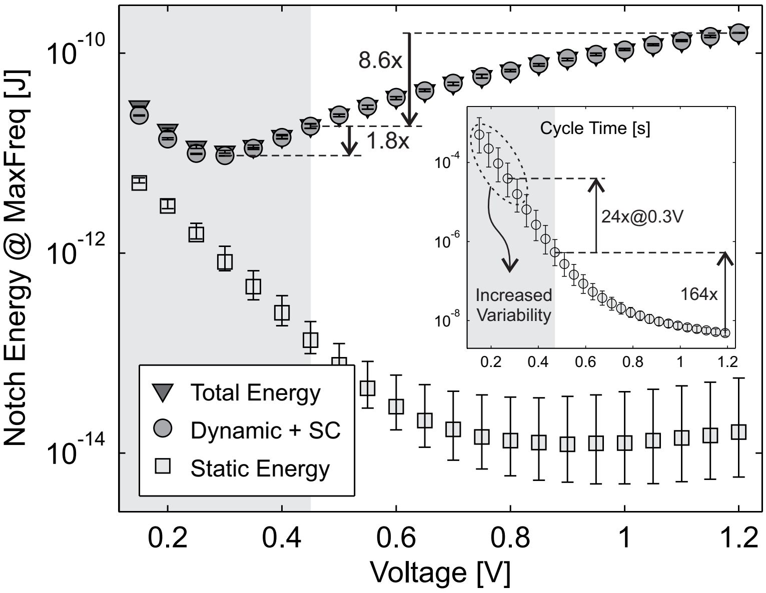
5.1 Power Analysis Methodology

The exercised power analysis methodology was performed using a post-synthesis netlist in a multi-mode multi corner (MMMC) environment, thus the netlist do not change over different voltages. The energy is always divided into two basic categories: static energy and dynamic plus short-circuit (SC) energy. The clock-tree power consumption is not included in the analysis. The notch filter results refer to the energy consumed while computing 2048 data samples from a health-care application. The 8051 core energy data was extracted for 10 loop iterations of the fixed point Dhrystone benchmark. The ISCAS benchmarks are simulated using 4096 random input values. The performance evaluation of the four combinational ISCAS circuits is done through the insertion of output registers into the original netlist. Regarding timing analysis, all constraints are adjusted depending on the voltage being evaluated, thus the input/output slopes and delays refer to actual fanout of 4 and clock-to-queue/setup delays of the corresponding voltage point. Each output pin is constrained with a load equivalent to the load of a simple register, and this load varies according to the voltage point being analyzed due to MOS capacitance effect. With respect to the maximum performance evaluation that is used through the analysis, we simply extrapolate the critical path slack time, reported by the static timing analysis (STA) tool.

5.2 Results on *very wide* range of VFS

The results listed in this Section include energy, performance and variability data for several benchmarks. First a complete analysis of the energy-savings achieved through a notch filter, from maximum frequency, iso-performance, and timing margin. Follows the analysis of an 8051 core, which results are very similar to the notch filter, thus only the

Figure 5.1: Notch filter energy to process 2048 samples at maximum frequency and varying supply voltage. The inset refers to the cycle time applied for each voltage. Upper and lower error bars refers to 3-sigma process variation (SS and FF corners).



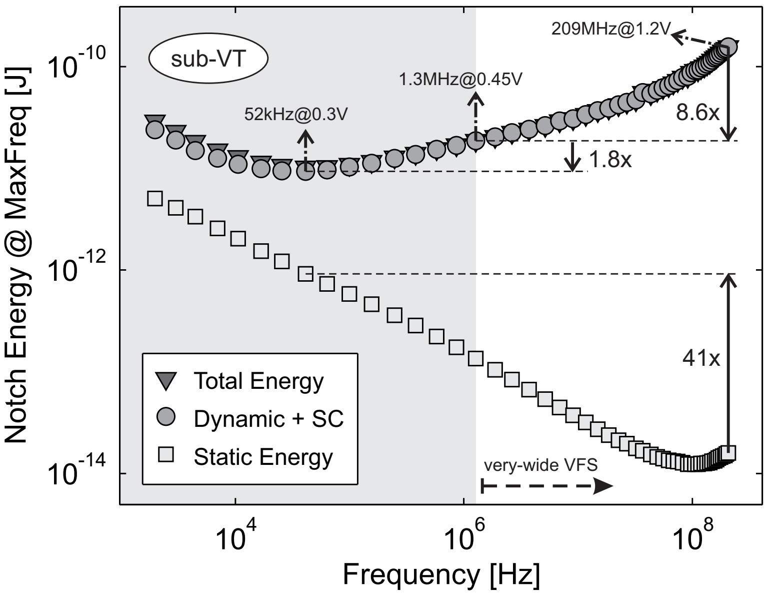
maximum frequency data is presented. Finally, this Section presents the results for a very broad range of benchmarks selected from public domain, known as ISCAS benchmarks (HANSEN; YALCIN; HAYES, 1999; BRGLEZ; BRYAN; KOZMINSKI, 1989). Special care was taken with respect to the simulation environment; for the notch filter, real data was used for filtering (SOARES et al., 2013); on the 8051 core, a synthetic, public domain, Dhrystone benchmark was adopted (WEICKER, 1988); and for the ISCAS circuits, a random input vector was the chosen alternative.

5.2.1 Notch Filter: Maximum Frequency, Iso-Performance, and Timing Margin

The notch filter energy consumed to process 2048 data samples at maximum frequency is shown in Figure 5.1. The energy saving at the lower boundary of near-VT regime, i.e. 0.45V, reaches 8.6X with respect to the nominal voltage at 1.2V. The inset in Figure 5.1 shows the minimum cycle time achievable at each voltage, showing that a 164X performance variation is possible when working on a very-wide range of VFS (at near-VT and well-above VT). Note the increase in static energy at voltages below 0.8V due to the dramatically increased cycle time. The minimum energy point is found at 0.30V, what may be referred as sub-VT operation. The additional energy saving with respect to the near-VT lower-boundary at 0.45V is only 1.8X, with a 24X performance drawback at sub-VT, in addition to the dramatically increased timing variability.

Figure 5.2 presents a clearer view of the energy performance trade-off achievable through voltage-frequency scaling. The energy versus frequency plot shows a large range of performance variation enabled by very-wide range of VFS (from 1.3MHz@0.45V, which is the limit we defined as near-VT operation, to 209MHz@1.2V). The sub-VT operation presents the best energy-efficiency, however the performance degradation is huge compared to the nominal voltage (a 4064X frequency variation for 209MHz at nominal VDD, to 52kHz at the minimum energy point), in addition to the strongly increased

Figure 5.2: Notch filter energy to process 2048 samples with varying VDD and maximum frequency (TT corner). Each point correspond to a different VDD and associates to a different frequency.

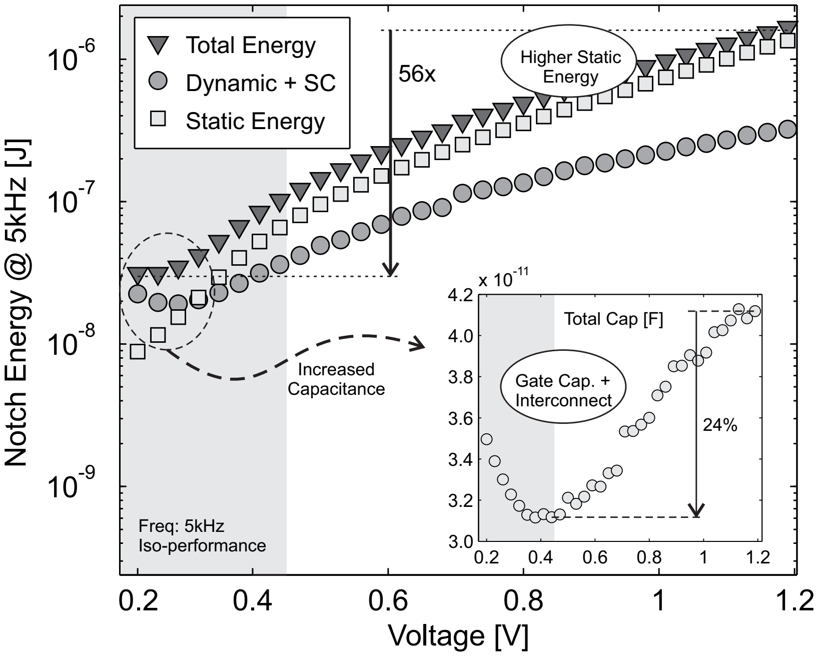


static energy (41X) and variability impact. Note that each point in Figure 5.2 refers to a different supply voltage for circuit operation.

The notch filter energy-efficiency was also evaluated in an iso-performance scenario, with a constant frequency of 5kHz (which is suitable for health-care applications). Figure 5.3 shows the energy consumption of this design operating at 5kHz. Note that voltages below 0.2V are not shown due to timing violation in at least one of the MMMC simulations for the filter operation. The energy curves show a significantly different behavior from the maximum frequency analysis. Static energy is higher than dynamic for every voltage point above 0.3V due to the very low frequency, an advanced 65nm technology has leaky transistors and is not targeted for very long clock periods, or very low performance circuits, at sub-MHz clocks. The energy-saving when operating the notch filter at 0.2V instead of nominal voltage (1.2V) is 56X. Our timing and power analyses performed in this CMOS circuit show the increase in dynamic energy below 0.3V (Figure 5.3). The inset in the same Figure 5.3 explains that a reversal, i.e. increase, occurs in the total net capacitance due to this operation at below 400 mV, with contributions from both junction and active gate capacitances – which increase the net capacitance loads. The total gate capacitance plus the interconnect are estimated by the Encounter® tool, and clearly shows a maximum variation of 24% depending on voltage. The increased sub-VT capacitance, in addition to slower slope delays, contributes to the increase in dynamic and short-circuit energy components.

Process variability is a critical issue for circuits operating at near-VT. The cell-library proposed in Chapter 4 was characterized in the typical (TT), fast (FF), and slow (SS) 3-sigma corners for evaluating the how process variation impact timing and energy of a VLSI circuit. From Figure 5.1 one may note that the static energy presents large coefficient of variation at nominal voltage, but produces a very limited change on total energy due to the dominating dynamic and short circuit components. Regarding timing varia-

Figure 5.3: Notch filter iso-performance energy curve with varying VDD at 5kHz (TT corner). The inset refers to the total circuit capacitance estimated by the multi-mode multi corner (MMMC) analysis tool.

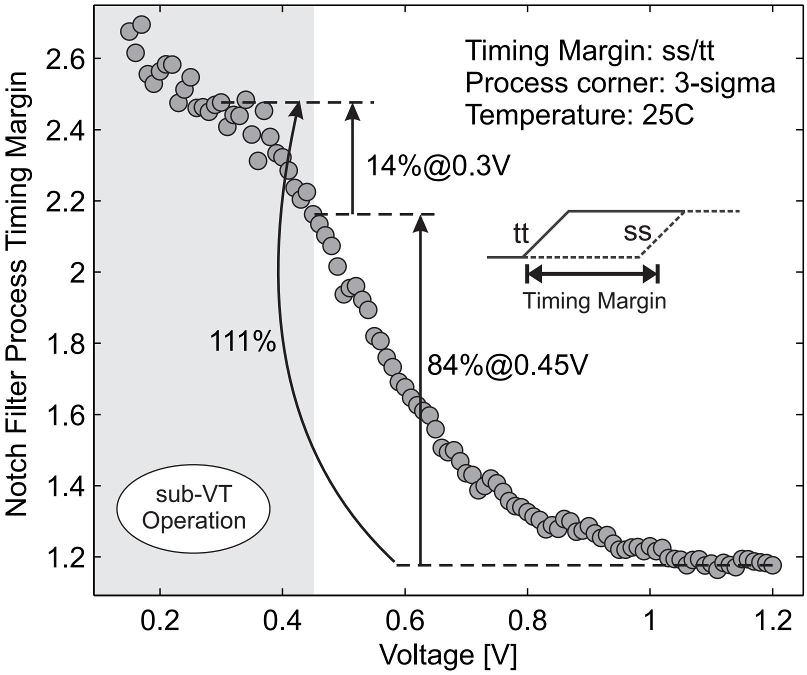


tion, the inset in Figure 5.1 shows the significant increase in timing variability as voltage scales down to near-VT. Figure 5.4 shows the timing margin, i.e. the amount of time a circuit might be slowed down to achieve functionality, with respect to typical conditions. In other words, timing margin refers to the ratio between the slow corner performance and the typical corner performance. The plot in Figure 5.4 shows that even at nominal voltage, nearly 20% of timing margin is required. This value increases in 84% at the bottom of near-VT regime. The minimum energy point presets an increase of 111% in timing margin with respect to nominal voltage.

5.2.2 8051 Core: Maximum Frequency with Dhrystone Benchmark

Among the benchmarks we selected for the energy-efficiency exploration under a very-wide range of VFS is a 20kgates 8051 compatible core. The activity traces of this core were extracted for 10 loop iterations of the fixed point Dhrystone benchmark (WEICKER, 1988). The memory static and dynamic energy consumption are not included in our analysis. The energy evaluation for maximum frequency analysis shows results similar to the notch filter, as shown in Figure 5.5. Operating the 8051 core at the lower boundary of near-VT (0.45V for this 65nm technology), presents 8X energy-savings with respect to nominal voltage, and 1.9X energy-savings at the minimum energy point with respect to the near-VT. The performance variation shows significantly higher frequencies than the notch filter, due to the lower complexity and faster critical path of this design. Figure 5.5 shows the corner cases for the energy consumption of the 8051 core (FF case for higher energy using fast PMOS and fast NMOS models, and SS for lower energy using slow PMOS and slow NMOS models). The maximum frequency at sub-VT operation (minimum energy) is 229kHz@0.28V, while the very-wide range of dynamic VFS goes

Figure 5.4: Notch filter timing margin (TT/SS) for 3-sigma process conditions.



from 7.55MHz@0.45V to 1.18GHz@1.2V.

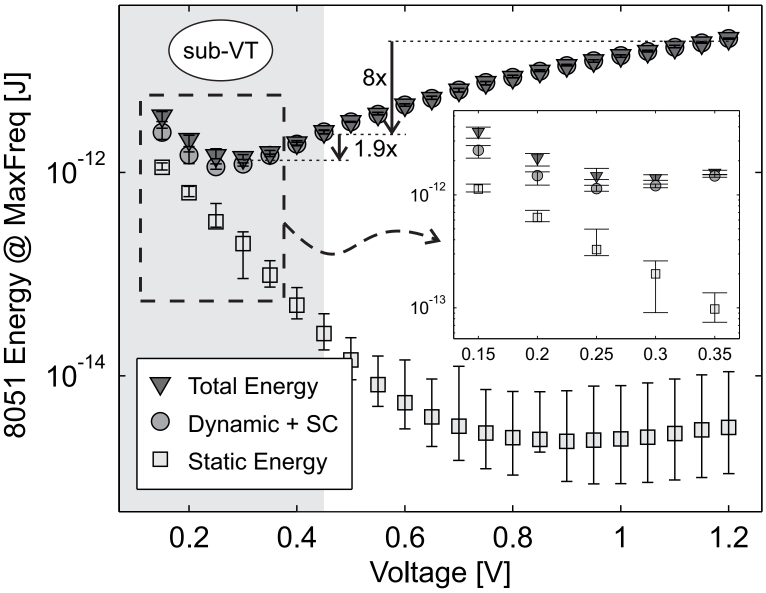
5.2.3 ISCAS Benchmark: Minimum-energy and Near-VT Analysis

In order to provide a more complete set of benchmarks, we adopted 4 combinational and 4 sequential benchmarks from ISCAS (HANSEN; YALCIN; HAYES, 1999; BRGLEZ; BRYAN; KOZMINSKI, 1989). They present a large complexity level variation, going from 420gates to 38kgates. The energy evaluation was exercised for 4096 random input vectors, at maximum frequency of each voltage point. The energy-efficiency and performance curves are similar to the already presented notch filter and 8051 cases, thus the relevant data is shown in Tables 5.1, 5.2, and 5.3. Even though the notch filter and 8051 core are not part of the ISCAS benchmark set, they are listed for comparison purposes.

Table 5.3 shows the analysis for three different voltage points, the minimum energy (at sub-VT), the near-VT lower boundary (at 0.45V for this process technology), and the nominal voltage at 1.2V. The first column shows the name of the benchmark. The ’Volt’ column shows the voltage on which the minimum energy point is located. The ’Energy’ and ’Ratio’ columns report the total energy in Joules and the ratio of static energy over total energy for the analyzed case (sub-, near-, and well-above VT). The ’Freq’ column reports the maximum frequency in Hz achieved at the corresponding voltage point.

Note in Table 5.3, that the minimum energy point voltage is roughly 0.3V for all benchmarks, however this point may suffer a substantial variation when an SRAM memory is included, due to higher static energy consumption. Also note that the ratio of static energy is substantially increased in sub-VT operation, however it reaches a maximum of 25% of total energy when at the minimum energy point, which is in confront to the common assumption that the minimum energy correspond to equal dynamic and static energies. This might be explained by the frequently neglected short-circuit energy com-

Figure 5.5: 8051 compatible core energy versus varying VDD. The simulation environment runs 10 loop iterations of the fixed point Dhrystone benchmark. Upper and lower error bars refer to 3-sigma process variation (SS and FF corners).



ponent and capacitance variation (which is shown to have a 24% variation with voltage in the inset of Figure 5.3.

Tables 5.1 and 5.2 show another point of view for the previously reported energy and performance data. In Table 5.1, the sub-VT total energy in Joules is reported at the second column, while third and fourth columns report the energy increase for near- and well-above VT regimes (@0.45V and @1.2V, respectively). Table 5.2 shows the maximum frequency for the same three cases, and corresponding performance increases. It is clear from Table 5.1 that the energy-savings through very-wide range of VFS may easily reach 8X, with respect to nominal voltages. The sub-VT regime, however, easily reaches an energy saving of 15X, at the cost of strongly reduced performance and increased variability effects.

Table 5.1: Energy increase with respect to the sub-VT minimum energy point, the near-VT @0.45V and the nominal voltage @1.2V.

	SUB-VT ENERGY [J]	NEAR-VT INCREASE	WELL-ABOVE VT INCREASE
Notch	10.27p	1.8X	15.5X
8051	1.37p	1.9X	15.2X
C432	43.16f	2.2X	18.3X
C1355	159.59f	1.9X	16.9X
C3540	288.91f	2.1X	17.4X
C6288	2.69p	1.9X	15.9X
S420	16.37f	2.8X	25.2X
S1423	141.39f	2.2X	18.8X
S9234	276.68f	2.4X	20.4X
S38584	1.37p	1.9X	16.6X

Table 5.2: Frequency increase with respect to the sub-VT minimum energy point performance, the near-VT @0.45V and the nominal voltage @1.2V.

	SUB-VT FREQ [HZ]	NEAR-VT INCREASE	WELL-ABOVE VT INCREASE
Notch	51.3k	24X	4064X
8051	229k	33X	5182X
C432	76.1k	61X	10562X
C1355	163k	31X	5508X
C3540	100k	40X	6663X
C6288	72.4k	31X	5331X
S420	178k	31X	5339X
S1423	90.1k	29X	4900X
S9234	184k	31X	5132X
S38584	272k	20X	1880X

Table 5.3: Energy-saving results for the Notch filter, 8051 core and ISCAS benchmarks. Energy in Joules, frequency in Hz, and voltage in Volts. The ’Ratio’ refers to static energy ratio over total energy.

	Minimum Energy Point				Near-VT (0.45V)			Well-above VT (1.2V)		
	Volt	Energy	Ratio	Freq	Energy	Ratio	Freq	Energy	Ratio	Freq
Notch	0.30	10.27p	7.95%	51.26k	18.79p	0.72%	1.27M	159.62p	0.01%	208.33M
8051	0.28	1.37p	15.95%	228.62k	2.53p	1.03%	7.55M	20.80p	0.01%	1.18G
C432	0.26	43.16f	15.15%	76.10k	93.48f	0.54%	4.69M	788.25f	0.01%	803.86M
C1355	0.29	159.59f	7.12%	162.68k	307.10f	0.53%	5.06M	2.70p	0.01%	896.06M
C3540	0.28	288.91f	11.19%	100.12k	596.20f	0.56%	4.05M	5.02p	0.02%	667.11M
C6288	0.29	2.69p	2.81%	72.36k	5.03p	0.22%	2.30M	42.75p	0.00%	385.80M
S420	0.29	16.37f	18.59%	177.53k	46.60f	0.85%	5.58M	411.78f	0.01%	947.87M
S1423	0.29	141.39f	25.25%	90.09k	313.94f	1.30%	2.69M	2.66p	0.02%	441.50M
S9234	0.29	276.68f	16.48%	184.16k	667.52f	0.78%	5.85M	5.64p	0.01%	945.18M
S38584	0.31	1.37p	19.95%	271.89k	2.63p	1.49%	5.61M	22.74p	0.04%	511.25M

6 CONCLUSION

This work presented results on the very-wide range of dynamic voltage-frequency scaling (VFS) of digital CMOS, from nominal voltage down to the lower boundary of near-VT operation. For this purpose, a sizing methodology has been used to mitigate variability effects, enhance static-noise margins and reduce the area and power overhead when operating the design at strong inversion. The developed cell-library allows maximum of 2-stacked transistor, and provides sizing rules for CMOS gates and master-slave registers. The evaluated benchmark circuits include a 25kgates notch filter design, a 20kgates 8051 compatible core, and 4-combinational/4-sequential ISCAS circuits. The results show that roughly 8X energy savings are possible through very-wide range of VFS, with respect to nominal voltages. The sub-VT operation however, presents an even higher energy-saving of roughly 2X with respect to near-VT, but at the cost of increased variability and a huge performance degradation. The iso-performance results shows an interesting increase in the dynamic plus short-circuit energy, which is probably due to the increased gate capacitance on sub-VT voltages (variation of 24% is reported), and increased short-circuit energy due to slower slope delays.

The results also show that at the minimum energy point, for the exercised benchmarks, the static energy is at most 25% of the total energy. This finding is in confront to the common assumption that at the minimum energy point, dynamic and static energies are roughly the same. This result is probably related to the commonly neglected short-circuit energy term, and also to the varying gate capacitance factor, which were both taken into account for this analysis.

The minimum energy point is shown to be located at roughly 0.29V for all evaluated designs. However, none of the exercised analysis take into account an important part of a SoC energy, the SRAM memory. The static energy consumption substantially increases when such a low activity circuit is included in the design, which may significantly deviate the minimum energy point to higher voltages depending on the SRAM size and topology. Another condition that may increase the minimum energy point, and was not taken into account in this work, is the temperature variation. The industrial range includes temperatures much higher than 25^{o}C, which substantially affects the device's threshold voltage, causing an exponential increase in sub-VT currents.

REFERENCES

ALIOTO, M.; CONSOLI, E.; PALUMBO, G. Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops: part i - methodology and design strategies. **Very Large Scale Integration (VLSI) Systems, IEEE Transactions on**, [S.l.], v.19, n.5, p.725–736, May 2011.

ASENOV, A. et al. Integrated atomistic process and device simulation of decananometre MOSFETs. In: SIMULATION OF SEMICONDUCTOR PROCESSES AND DEVICES, 2002. SISPAD 2002. INTERNATIONAL CONFERENCE ON. **Proceedings...** [S.l.: s.n.], 2002. p.87–90.

BHUNIA, S.; MUKHOPADHYAY, S. **Low-Power Variation-Tolerant Design in Nanometer Silicon**. 1st.ed. [S.l.]: Springer, 2011.

BISDOUNIS, L.; NIKOLAIDIS, S.; LOUFOPAVLOU, O. Propagation delay and short-circuit power dissipation modeling of the CMOS inverter. **Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on**, [S.l.], v.45, n.3, p.259–270, 1998.

BRGLEZ, F.; BRYAN, D.; KOZMINSKI, K. Combinational profiles of sequential benchmark circuits. In: CIRCUITS AND SYSTEMS, 1989., IEEE INTERNATIONAL SYMPOSIUM ON. **Proceedings...** [S.l.: s.n.], 1989. p.1929–1934.

CADENCE. Cadence Design Systems: product manual. **RC910, EDI101**, [S.l.], 2011.

CALHOUN, B.; CHANDRAKASAN, A. Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits. In: LOW POWER ELECTRONICS AND DESIGN (ISLPED), INTERNATIONAL SYMPOSIUM ON. **Proceedings...** [S.l.: s.n.], 2004. p.90–95.

CALHOUN, B.; CHANDRAKASAN, A. A 256-kb 65-nm Sub-threshold SRAM Design for Ultra-Low-Voltage Operation. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.42, n.3, p.680–688, Mar 2007.

CALHOUN, B.; WANG, A.; CHANDRAKASAN, A. Modeling and sizing for minimum energy operation in subthreshold circuits. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.40, n.9, p.1778–1786, Sep 2005.

CHANDRAKASAN, A.; SHENG, S.; BRODERSEN, R. Low-power CMOS digital design. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.27, n.4, p.473–484, apr 1992.

DEAL, B. et al. Characteristics of the surface-state charge (Q) of thermally oxidized silicon. **Journal of the Electrochemical Society**, [S.l.], v.114, n.3, p.266–274, Mar 1967.

DENNARD, R. et al. Design Of Ion-implanted MOSFET's with Very Small Physical Dimensions. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.9, n.5, p.256–268, Oct 1974.

ESMAEILZADEH, H. et al. Dark silicon and the end of multicore scaling. **SIGARCH Comput. Archit. News**, [S.l.], v.39, n.3, p.365–376, Jun 2011.

FRANCH, R. et al. On-chip Timing Uncertainty Measurements on IBM Microprocessors. In: TEST CONFERENCE, 2008. ITC 2008. IEEE INTERNATIONAL. **Proceedings...** [S.l.: s.n.], 2008. p.1–7.

GIACOMOTTO, C.; NEDOVIC, N.; OKLOBDZIJA, V. The Effect of the System Specification on the Optimal Selection of Clocked Storage Elements. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.42, n.6, p.1392–1404, Jun 2007.

HANSEN, M.; YALCIN, H.; HAYES, J. Unveiling the ISCAS-85 benchmarks: a case study in reverse engineering. **Design Test of Computers, IEEE**, [S.l.], v.16, n.3, p.72–80, 1999.

HE, Y. et al. Xetal-Pro: an ultra-low energy and high throughput simd processor. In: DESIGN AUTOMATION CONFERENCE (DAC), 2010 47TH ACM/IEEE. **Proceedings...** [S.l.: s.n.], 2010. p.543–548.

HSU, S. et al. A 280mV-to-1.1V 256b reconfigurable SIMD vector permutation engine with 2-dimensional shuffle in 22nm CMOS. In: SOLID-STATE CIRCUITS CONFERENCE DIGEST OF TECHNICAL PAPERS (ISSCC), 2012 IEEE INTERNATIONAL. **Proceedings...** [S.l.: s.n.], 2012. p.178–180.

IBM. Industrial Business Machines: product manual. **CMOS10LPE Bulk**, [S.l.], 2009.

JAIN, S. et al. A 280mV-to-1.2V wide-operating-range IA-32 processor in 32nm CMOS. In: SOLID-STATE CIRCUITS CONFERENCE DIGEST OF TECHNICAL PAPERS (ISSCC), 2012 IEEE INTERNATIONAL. **Proceedings...** [S.l.: s.n.], 2012. p.66–68.

KEANE, J. et al. Subthreshold logical effort: a systematic framework for optimal subthreshold device sizing. In: DESIGN AUTOMATION CONFERENCE (DAC), PROCEEDINGS OF THE 43RD ACM/IEEE. **Proceedings...** [S.l.: s.n.], 2006. p.425–428.

KIM, T. et al. Utilizing Reverse Short-Channel Effect for Optimal Subthreshold Circuit Design. **Very Large Scale Integration (VLSI) Systems, IEEE Transactions on**, [S.l.], v.15, n.7, p.821–829, Jul 2007.

KIM, T. et al. A 0.2 V, 480 kb Subthreshold SRAM With 1 k Cells Per Bitline for Ultra-Low-Voltage Computing. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.43, n.2, p.518–529, 2008.

KWONG, J.; CHANDRAKASAN, A. Variation-driven device sizing for minimum energy sub-threshold circuits. In: LOW POWER ELECTRONICS AND DESIGN (ISLPED), PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM ON. **Proceedings...** [S.l.: s.n.], 2006. p.8–13.

LEFURGY, C. et al. Active management of timing guardband to save energy in POWER7. In: ANNUAL IEEE/ACM INTERNATIONAL SYMPOSIUM ON MICROARCHITECTURE, 44. **Proceedings. . .** [S.l.: s.n.], 2011. p.1–11.

LIU, T.-T.; RABAEY, J. A 0.25 V 460 nW Asynchronous Neural Signal Processor With Inherent Leakage Suppression. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.48, n.4, p.897–906, 2013.

LOHSTROH, J.; SEEVINCK, E.; GROOT, J. de. Worst-case static noise margin criteria for logic circuits and their mathematical equivalence. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.18, n.6, p.803–807, 1983.

NOSE, K.; SAKURAI, T. Analysis and future trend of short-circuit power. **Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on**, [S.l.], v.19, n.9, p.1023–1030, 2000.

RABAEY, J.; CHANDRAKASAN, A.; NIKOLIC, B. **Digital Integrated Circuits**: a design perspective. 2nd.ed. [S.l.]: Prentice Hall, 2003.

ROTEM, E. et al. Temperature measurement in the Intel CoreTM Duo Processor. In: INTERNATIONAL WORKSHOP ON THERMAL INVESTIGATIONS OF ICS, 12. **Proceedings. . .** [S.l.: s.n.], 2006. p.8–13.

SEEVINCK, E.; LIST, F.; LOHSTROH, J. Static-noise margin analysis of MOS SRAM cells. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.22, n.5, p.748–754, Oct 1987.

SOARES, L. et al. 61 pJ/sample Near-Threshold Notch Filter with Pole- Radius Variation. In: IEEE 4TH LATIN AMERICAN SYMPOSIUM ON CIRCUITS AND SYSTEMS (LASCAS), 2013. **Proceedings. . .** [S.l.: s.n.], 2013. p.746.

SUTHERLAND, I.; SPROULL, R. Logical effort: designing for speed on the back of an envelope. In: UNIVERSITY OF CALIFORNIA/SANTA CRUZ CONFERENCE ON ADVANCED RESEARCH IN VLSI, 1991. **Proceedings. . .** [S.l.: s.n.], 1991. p.1–16.

VEMURU, S. R.; SCHEINBERG, N. Short-circuit power dissipation estimation for CMOS logic gates. **Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on**, [S.l.], v.41, n.11, p.762–765, 1994.

VERMA, N.; CHANDRAKASAN, A. A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.43, n.1, p.141–149, 2008.

WANG, A.; CHANDRAKASAN, A. A 180-mV subthreshold FFT processor using a minimum energy design methodology. **Solid-State Circuits, IEEE Journal of**, [S.l.], v.40, n.1, p.310–319, Jan 2005.

WEICKER, R. Dhrystone benchmark: rationale for version 2 and measurement rules. **SIGPLAN Not.**, [S.l.], v.23, n.8, p.49–62, Aug 1988.

ZHAI, B. et al. Analysis and mitigation of variability in subthreshold design. In: LOW POWER ELECTRONICS AND DESIGN, 2005. ISLPED ’05. PROCEEDINGS OF THE 2005 INTERNATIONAL SYMPOSIUM ON. **Proceedings. . .** [S.l.: s.n.], 2005. p.20–25.

APPENDIX A INTRODUÇÃO

O projeto de circuitos CMOS foi historicamente focado em performance. Processadores de alto desempenho tiveram sua frequência aumentada até o ano de 2004, quando alcançaram os 4.0GHz – dissipando 120W de potência em um único chip. As abordagens de projeto orientadas somente a performance não mais sustentavam a demanda por performance com um custo razoável. A mudança no cenário de potência foi de fato causada pela falha nas regras de miniaturização preditas por Dennard (DENNARD et al., 1974), onde parâmetros intrínsecos dos materiais como potencial de junção e tensão de limiar impuseram uma barreira para a diminuição da tensão de alimentação de um processo para outro, portanto aumentando a densidade de potência a cada nova tecnologia. Para superar esta barreira, uma nova metodologia de projeto foi desenvolvida. Técnicas para economia de potência em projetos VLSI variam desde otimizações no processo, layout, circuito, arquitetura, até a pilha de software e aplicação.

Do ponto de vista arquitetural, uma mudança significativa precisou ocorrer para continuar aumentando a performance: a adoção de multi-cores. Mesmo com a mudança de paradigma na programação, projetos multi-core possibilitaram um aumento contínuo em performance, sem comprometer o projeto em termos de potência dissipada. Entretanto, pesquisas recentes apontam que independente da organização do chip ou topologia, o aumento no número de cores também é limitado por potência; no nodo de 22nm, 21% dos transistores no chip devem ser desligados para dissipar o calor gerado, e este número aumenta para mais de 50% no nodo de 8nm (ESMAEILZADEH et al., 2011). Além desta questão da densidade de potência, um número crescente de aplicações requerem soluções de ultra-baixo consumo, capazes de maximizar a vida útil da bateria para sistemas autônomos como rede de sensores, dispositivos vestíveis, sensores leves, monitoramento ubíquo de ambientes, e assim por diante.

Para minimizar a potência e a energia de circuitos digitais CMOS, diversas técnicas vêm sendo apresentadas, como: *clock gating*, para reduzir chaveamentos desnecessários no clock em um grupo de registradores; isolamento de operadores, para evitar a propagação de *glitches* desnecessários pelo fluxo de dados; *power gating*, para reduzir o consumo de energia estática; dispositivos multi-limiar, para a diminuir o consumo estático em caminhos não críticos; e finalmente, ajuste de tensão e frequência (VFS), que visa reduzir a tensão e a frequência, procurando explorar a característica de rajada da maioria das cargas de trabalho. Estas técnicas possibilitam o projeto de processadores que atendem a um amplo espectro de performance e dissipação de energia, fornecendo soluções altamente configuráveis para o compromisso entre energia e performance – sem comprometer o tempo de chegada ao mercado.

Este trabalho dedica atenção especial a técnica de VFS, que possibilita uma mudança dinâmica em aspectos chaves do chip, como tensão e frequência. Portando, fornecendo uma solução integrada de tempo real para explorar o compromisso de potência e performance de acordo com a carga de trabalho. Notavelmente, na maioria das aplicações, a carga de trabalho varia desde a máxima performance (como nos modos de transmissão e recepção), para o mínimo consumo em modos de economia de energia. Neste cenário de cargas de trabalho variáveis, VFS dinâmico surge como uma técnica proeminente para economia de energia, causando uma enorme e monotônica redução de potência, além de um aumento em eficiência energética de uma ordem de grandeza. A indústria no entanto, explora uma estreita faixa de VFS dinâmico em processadores de alto desempenho, que geralmente se traduz em 2X – 3X variação em frequência.

A adoção limitada da técnica de VFS na indústria ocorre devido a variabilidade oriunda de variações no processo e no ambiente. Os efeitos da variabilidade também estão presentes em tensão nominal, porém seu impacto aumenta drasticamente em baixas tensões, que por sua vez, apresenta um coeficiente de variabilidade significativo em tecnologias mais avançadas (abaixo de 100nm) (ZHAI et al., 2005). Para ilustrar, um projeto que opera em 250mV, deve levar em conta 2.5X de aumento no tempo de ciclo devido a variação de 3-sigma no processo, 1.7X devido a 10% de variação na tensão de alimentação, e 23X para o *corner* de -40^{o}C – resultando em uma variação total de 99X. O mesmo projeto em tensão nominal, sob as mesmas condições, está sujeito a uma variação total de somente 1.5X no tempo de ciclo.

Esta dissertação de mestrado visa demonstrar os desafios e vantagens ao operar circuitos digitais CMOS robustos em tensões de alimentação próximas ao limiar do transistor, assumindo a utilização de VFS dinâmico. Para superar os efeitos da variabilidade nestas tensões ultra-baixas, e garantir a operação robusta com respeito a varições de processos e ambiente, circuitos que operam abaixo ou próximo ao limiar do transistor requerem o projeto cuidadoso de uma biblioteca de células digitais robusta e capaz de tolerar os efeitos da variabilidade. As células projetadas devem levar em consideração todos os efeitos de baixa tensão que surgem quando os circuitos operam em faixas de tensão tão reduzidas, por exemplo, amplitudes reduzidas de tensão, degradação das margens estáticas de ruído e maior relação entre corrente dinâmica e estática. Esta nova biblioteca de células projetada para tensões ultra-baixas pode apresentar perdas em performance e área quando operando em tensão nominal se comparada a bibliotecas comerciais, no entanto a eficiência energética atingida em baixas tensões é substancial, e oferece uma nova perspectiva para o projeto de circuitos CMOS de baixa potência.

Trabalhos relacionados em (CHANDRAKASAN; SHENG; BRODERSEN, 1992), procuram pela menor tensão de alimentação possível, enquanto compensam a degradação no rendimento com paralelismo arquitetural. Em (CALHOUN; WANG; CHANDRAKASAN, 2005), os autores derivam equações para o ponto de energia mínima e mostram que, teoricamente, transistores de tamanho mínimo são ideias para redução do consumo. O trabalho em (WANG; CHANDRAKASAN, 2005) demonstra, utilizando um processador FFT, a metodologia de projeto para uma biblioteca de células e uma memória SRAM que funcionam em tensões abaixo do limiar do transistor. O processador FFT é fabricado em uma tecnologia 0.18u, atingindo, para um projeto 16-bit/1024 pontos, o ponto de energia mínimo em 180mV com frequência de 10kHz. Em (KWONG; CHANDRAKASAN, 2006), os autores propõem uma nova metodologia de dimensionamento de transistores, baseada em simulações Monte Carlo e medidas de margem estática de ruído. Os resultados mostraram que o aumento do tamanho dos transistores é necessário para operação robusta quando abaixo do limiar do transistor. O método de dimensionamento baseado em simulação proposto em (KEANE et al., 2006) utiliza simulações SPICE DC para encontrar o melhor dimensionamento de transistores de modo a melhorar as margens estáticas de ruído. O autor também deriva uma solução fechada para o dimensionamento de dispositivos em série. O dimensionamento destes circuitos de ultra-baixa tensão pode também depender da tecnologia de processo, ou seja, em (KIM et al., 2007) o dimensionamento ótimo de dispositivos abaixo do limiar é mostrado requerer uma largura de canal maior do que o valor mínimo devido ao efeito de canal curto reverso (RSCE).

No que diz respeito ao projeto de SRAMs de ultra-baixa tensão, uma célula SRAM de 10T é proposta em (CALHOUN; CHANDRAKASAN, 2007). A memória SRAM é demonstrada funcionando apropriadamente em 380mV CMOS 65nm. Em (KIM et al., 2008), a célula 10T é mostrada funcionando em 0.2V CMOS 130nm. Do ponto de vista de circuito topológico, a célula SRAM 10T separa as linhas de escrita e leitura, e insere um buffer de adicional para melhorar a leitura da *bitline*. Embora a célula de 10T seja capaz de operar em tensões bastante reduzidas, o aumento em área devido ao maior número de transistores em cada célula ainda impõe uma forte limitação para sua ampla adoção. Mesmo a célula melhorada 8T proposta em (VERMA; CHANDRAKASAN, 2008) não é tão densa quanto a célula tradicional 6T. Em (HE et al., 2010), um processador SIMD que utiliza uma memória SRAM de 10Mbit é apresentado; questões relativas a densidade de memória levaram os projetistas a usar uma memória SRAM 6T, que está situada em uma ilha de tensão a parte, cuja alimentação não reduz.

Outras estratégias tem sido recentemente desenvolvidas para diminuir os efeitos da variabilidade em nível de circuito. Em (LIU; RABAEY, 2013) um processador de sinais neurais é proposto utilizando técnicas de projeto assíncronas, portanto o atraso do caminho crítico pode teoricamente ser o mínimo para qualquer processo e condições ambientais. Outra demonstração de técnicas em nível de circuito para mitigar a variabilidade é apresentada em (LEFURGY et al., 2011), onde um monitor de caminho crítico, acoplado com circuitos de geração de clock, otimiza o tempo de guarda do processador, que depende de muitas fontes de variações dos atrasos, por exemplo, processo e condições ambientais, envelhecimento, carga de trabalho, etc.

Trabalhos recentes tem demonstrado processadores operando com VFS dinâmico. Em (JAIN et al., 2012), um processador IA-32 fabricado em tecnologia CMOS 32nm, opera de 280mV até 1.2V. A performance varia três ordens de magnitude, de 3MHz até 915MHz, possibilitando uma variação de 4.7X em eficiência energética. Em (HSU et al., 2012), uma máquina de permutação vetorial SIMD reconfigurável de 4 até 32 fluxos é demonstrada em CMOS 22nm operando de 240mV até 1.1V. A performance varia de 15MHz até 2.5GHz, enquanto que o ponto de energia mínima oferece uma eficiência energética 9X maior do que a obtida em tensão nominal.

Esta dissertação de mestrado está organizada como segue: o Capítulo 2 discute as maiores fontes de variabilidade, incluindo processo, ambiente e envelhecimento. Este capítulo também inclui uma análise quantitativa da variabilidade para circuitos de ultra-baixa potência a partir de um ponto de vista de projeto. O Capítulo 3 traz uma breve revisão das principais técnicas para projeto de circuitos CMOS de ultra-baixa potência, incluindo *clock gating*, isolamento de operadores, *power gating*, dispositivos multi-limiar e VFS dinâmico. O Capítulo 4 mostra a metodologia de projeto utilizada para desenvolver esta biblioteca de células lógicas para operação próxima ao limiar do transistor. O Capítulo 5 reporta a metodologia utilizada em simulação, e traz os resultados em eficiência energética obtidos com nossa biblioteca de células para operação próximo ao limiar do transistor, em VFS dinâmico. Finalmente, o Capítulo 6 apresenta as conclusões deste trabalho.

APPENDIX B RESUMO DA DISSERTAÇÃO

Processos avançados impõem desafios cada vez maiores para projetistas de circuitos no que diz respeito a variabilidade. O Capítulo 2 discute as principais fontes de variação em circuitos CMOS, incluindo fontes sistemáticas por meio de litografia, e também estatísticas como as originadas por variações de dopantes (RDF, *random dopant fluctuations*). As variações ambientais são também discutidas e focam nas variações devido a temperatura e tensão de alimentação. No que diz respeito a confiabilidade de circuitos CMOS, variações temporais também são brevemente exploradas por meio de BTI (*bias temperature instability*), HCI (*hot carrier injection*), e TDDB (*time-dependent dielectric breakdown*). O final do Capítulo 2 apresenta algumas análises quantitativas de processo e condições ambientais utilizando um estudo de caso de um oscilador em anel.

O projeto de circuitos de ultra-baixa potência requer otimizações em todos os níveis, ou seja, processo, circuito, micro-arquitetural, e nível de sistema. Otimizações de alto nível apresentam um forte impacto em energia, porém são bastante específicas para um certo caso ou aplicação. O Capítulo 3 visa fornecer informações em técnicas para projeto de circuitos de baixa potência, que não são dependentes de aplicação e podem ser aplicadas a um grande número de projetos. O Capítulo 3 começa com uma breve revisão dos principais componentes da potência, e segue com diversas técnicas ao nível de circuito para projetos CMOS de baixa potência, incluindo *clock gating*, *power gating*, isolamento de operadores, dispositivos multi-limiares, e finalmente, ajuste de tensão e frequência.

O estado da arte em projeto de chips em CMOS decananometro pode facilmente atingir centenas de milhões de transistores em um único chip. O projeto destes circuitos VLSI altamente integrados requer diversos níveis de abstração. Em um fluxo de projeto digital tradicional, uma biblioteca de células padrão pode ser considerada a ponte entre a lógica e o layout, separando o domínio digital discreto do analógico contínuo. Bibliotecas de células encapsulam funções lógicas em pequenos blocos de layout. Estes layouts regulares implementam uma função lógica com entradas e saídas, portanto fornecendo abstração de projeto necessária para as camadas mais altas. Outra vantagem desta abordagem baseada em blocos associa-se ao posicionamento e roteamento automáticos, que são enormemente simplificados devido a regularidade no layout das células lógicas. O Capítulo 4 cobre tópicos a respeito da biblioteca de células lógicas para operação próximo ao limiar do transistor, no que diz respeito ao dimensionamento de transistores para tensões de alimentação ultra-baixas. Questões de variabilidade que afetam as margens estáticas de ruído (SNM, *static noise margins*) e as amplitudes de tensão também são avaliadas.

O Capítulo 5 apresenta os resultados referentes a eficiência energética da nossa biblioteca de células lógicas para ampla ajuste de tensão e frequência. Os circuitos de avaliação incluem um filtro notch de 25kgates, um microcontrolador de 20kgates e 4 circuitos ISCAS combinacionais/sequenciais (HANSEN; YALCIN; HAYES, 1999; BRGLEZ; BRYAN; KOZMINSKI, 1989). Nossa abordagem para ajuste dinâmico de tensão e frequência inclui uma faixa de tensão desde o valor nominal até o limite inferior da operação próxima ao limiar do transistor; entretanto nós estendemos a caracterização de nossa biblioteca de células até 150mV, gerando portanto resultados de comparação entre abaixo-, próximo e muito acima do limiar do transistor. Adicionalmente, foi realizada uma caracterização nos *corners* FF, TT e SS, todos em temperatura ambiente (25ºC).

APPENDIX C CONCLUSÃO

Este trabalho apresentou resultados para operação de circuitos digitais CMOS em uma ampla faixa de frequências e tensões, desde a tensão nominal até o limite inferior da operação próxima ao limiar do transistor. Com este propósito, a metodologia de dimensionamento dos transistores visa diminuir os efeitos da variabilidade, aumentar as margens de sinal ruído, e reduzir o custo em área e potência quando operando o circuito em inversão forte. A biblioteca de células desenvolvida permite um número máximo de dois transistores em série, e define regras de *sizing* para portas lógicas digitais CMOS, assim como registradores mestre-escravo. Os circuitos avaliados incluem um filtro notch de 25kgates, um microcontrolador 8051 de 20kgates, e 4 circuitos combinacionais/sequencias ISCAS de avaliação. Os resultados mostram que ganhos de aproximadamente 8X em eficiência energética, comparando com tensão nominal, são possíveis explorando uma ampla faixa de tensão e frequência. A operação abaixo do limiar do transistor, no entanto, apresenta um ganho em eficiência energética ainda maior, de 2X em relação a operação próxima ao limiar do transistor, porém com o custo de um aumento na variabilidade e enorme diminuição em performance. Os resultados para performance constante mostram um aumento interessante na energia dinâmica mais curto-circuito, que provavelmente deve-se ao aumento na capacitância de porta para transistores que operam em tensões abaixo do limiar (uma variação de até 24% é reportada). Este aumento da energia dinâmica mais curto-circuito também deve-se as transições mais lentas quando operando em baixas tensões, causando um aumento na componente de curto-circuito.

Os resultados também mostram que no ponto de energia mínima, para os circuitos testados, a energia estática é no máximo 25% da energia total. Esta descoberta está em confronto com a suposição de que no ponto de energia mínima, as energias dinâmica e estática são aproximadamente as mesmas. Este resultado está provavelmente relacionado à comumente negligenciada energia de curto-circuito, e também à variação da capacitância de porta; este trabalho levou ambos em consideração.

O ponto de energia mínima é demonstrado estar aproximadamente em 0.29V, para todos os projetos avaliados. Entretanto, nenhuma das análises demonstradas leva em consideração uma importante parte da energia de um SoC, a memória SRAM. O consumo de energia estática aumenta substancialmente quando um circuito de baixa atividade de chaveamento é incluído no projeto, podendo significativamente desviar o ponto de energia mínima para tensões maiores dependendo do tamanho da SRAM e de sua topologia. Outra condição que pode aumentar o ponto de energia mínima, e não foi levada em consideração neste trabalho, é a variação em temperatura. A faixa de operação industrial inclui temperaturas muito maiores do que 25^{o}C, afetando substancialmente o limiar dos transistores, e portanto causando um aumento exponencial nas correntes do dispositivo quando operando abaixo do limiar.