Energy Efficient Design for Deep Sub-micron CMOS VLSIs

by

Mohamed Elgebaly

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2005

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Over the past decade, low power, energy efficient VLSI design has been the focal point of active research and development. The rapid technology scaling, the growing integration capacity, and the mounting active and leakage power dissipation are contributing to the growing complexity of modern VLSI design. Careful power planning on all design levels is required. This dissertation tackles the low-power, low-energy challenges in deep sub-micron technologies on the architecture and circuit levels.

Voltage scaling is one of the most efficient ways for reducing power and energy. For ultra-low voltage operation, a new circuit technique which allows bulk CMOS circuits to work in the sub-0.5V supply territory is presented. The threshold voltage of the slow PMOS transistor is controlled dynamically to get a lower threshold voltage during the active mode. Due to the reduced threshold voltage, switching speed becomes faster while active leakage current is increased. A technique to dynamically manage active leakage current is presented. Energy reduction resulting from using the proposed structure is demonstrated through simulations of different circuits with different levels of complexity.

As technology scales, the mounting leakage current and degraded noise immunity impact performance especially that of high performance dynamic circuits. Dual threshold technology shows a good potential for leakage reduction while meeting performance goals. A model for optimally selecting threshold voltages and transistor sizes in wide fan-in dynamic circuits is presented. On the circuit level, a novel circuit level technique which handles the trade-off between noise immunity and energy dissipation for wide fan-in dynamic circuits is presented. Energy efficiency of the proposed wide fan-in dynamic circuit is further enhanced through efficient low voltage operation.

Another direct consequence of technology scaling is the growing impact of interconnect parasitics and process variations on performance. Traditionally, worst case process, parasitics, and environmental conditions are considered. Designing for worst case guarantees a fail-safe operation but requires a large delay and voltage margins. This large margin can be recovered if the design can adapt to the actual silicon conditions. Dynamic voltage scaling is considered a key enabler in reducing such margin. An on-chip process identifier to recover the margin required due to process variations is described. The proposed architecture adjusts supply voltage using a hybrid between the one-time voltage setting and the

continuous monitoring modes of operation. The interconnect impact on delay is minimized through a novel adaptive voltage scaling architecture. The proposed system recovers the large delay and voltage margins required by conventional systems by closely tracking the actual critical path at anytime. By tracking the actual critical path, the proposed system is robust and more energy efficient compared to both the conventional open-loop and closed-loop systems.

Acknowledgements

This dissertation is dedicated to the memory of my father, Maher Elgebaly.

Looking back on my days at Waterloo, I realize that I am tremendously blessed by God Almighty to have the ability to withstand the pressure of graduate studies and to be able to complete this dissertation. I have been also blessed to have had teachers, friends, colleagues, and a family who have been providing me with guidance and support.

I am fortunate to have had Prof. Manoj Sachdev as my research advisor. This thesis would have been difficult without his encouragement, support, and listening to my ideas even when they did not make sense. In Waterloo, Manoj and many other professors inspired me on both the academic and the personal levels. In particular, I would like to acknowledge Mark Aagaard, Anwar Hassan, and Gord Agnew.

It would have been difficult for me to continue the course of graduate school without an exceptional group of friends. I still remember when Amr Wassal and Ayman Alsayed picked me up from Pearson Airport when my feet first touched the Canadian land. Nayer Wanas and Mohamed Mohsen were my housemates for a year. We had a wonderful time together, from painting our rooms in the house, the submarine, to the amount of food that we used to cook. I will always remember that whatever the amount of food left after dinner would not see the morning light. Has Khalid Hammouda joined the club, I have had a good time listening to an enormous number of funny jokes. We enjoyed our dinners together, the movie nights, and the road trips to Montreal and Toronto. My memories of Algonquin park camping trips will never fade away. With Muhammed Nummer, I drafted the second idea on dynamic voltage scaling. I learned from Mohamed Kamal how objectivity is an essential part of success. Ehab Elsaadaany and Yasser Ibrahim are like big brothers to me with their soft and open hearts. With the support of these friends and many others I managed my way through the years I spent in Waterloo.

Having an internship at Qualcomm in San Diego has had a profound impact on my research track. Amr Fahim, Tauseef Kazi, and Inyap Kang have introduced me to the second part of this dissertation, dynamic voltage scaling, and guided me through. Amr was of tremendous help during my stay at Qualcomm and in San Diego in general. I also would like to acknowledge Lew Chua and Devin Kelley who helped me testing my second test chip.

In the department of Electrical and Computer Engineering, Wendy Boles and Phil Regier have been very helpful on the administrative and the computing related issues. The Natural Science and Engineering Research in Canada (NSERC) provided the financial support for my research.

I am at a loss of words to express my gratitude to my mother and my brother for their continuous love and support. Spending years away from home was hard for me and even harder for my mother who patiently has encouraged and supported me. Mahmoud, my wonderful brother, has been there for me whenever I needed him. I feel fortunate to have them always standing by my side.

I am grateful to God who blessed me with my wife, Amira. Her love, support, and encouragement throughout the last and most important period of my Ph.D. has made it possible for me to reach the stage of writing these lines. She put my dreams ahead of hers. For this I am very grateful, my princess.

Contents

List of Tables

List of Figures

Nomenclature

V_{DD}	Supply voltage.
V_{TH}	Transistor threshold voltage.
v_{TH}	Normalized Transistor threshold voltage.
V_{TH0}	Transistor threshold voltage at zero bias.
γ	Body effect coefficient.
η	Drain-induced barrier lowering effect.
V_T	Thermal Voltage (26mV at 25^oC).
$2\Phi_F$	Surface potential at strong inversion.
λ	Channel length modulation coefficient.
α	Alpha-law power model coefficient.
I_{DSAT}	Transistor saturation Current.
V_{DSAT}	Transistor saturation voltage.
I_{sub}	Subthreshold leakage current.
I_{w}	Weak inversion leakage current.
n	Subthreshold swing coefficient.
V_{off}	BSIM3 subthreshold current fitting parameter.
N_{CH}	Channel carrier concentration.
μ_{eff}	Effective mobility.
tox	Oxide thickness.
C_{gd}	Gate-Drain capacitance.
C_{diff}	Diffusion capacitance.
C_D	Dynamic node capacitance.

C_L	Load capacitance.
τ_r	Signal rising time.
t_{HL}	Output delay time during a high to low transition.
t_{LH}	Output delay time during a low to high transition.
I_{ratio}	Interconnect to logic delay ratio of a path delay.
t_{dl}	Logic delay potion of a path delay.
t_{di}	Interconnect delay potion of a path delay.

List of Abbreviations

DTPMOS	Dynamic threshold PMOS.
MTCMOS	Multi-threshold MOS.
VTCMOS	Variable-threshold MOS.
DVT	Dual-threshold logic.
SOI	Silicon On Insulator (SOI).
FA	Full adder.
SD	Split-domino logic.
CKP	Conditional-keeper domino logic.
UGDN	Unity-gain dynamic noise.
DVS	Dynamic voltage scaling.
AVS	Adaptive voltage scaling.
PWM	Pulse-width modulated signal.
CPE	Critical Path Emulator
LUT	Look-up table.

Chapter 1

Introduction

The tremendous success of the semiconductor industry over the last 50 years has simply caused a significant change in our lifestyle. Integrated circuits are everywhere from computers to automobiles, from cell phones to home appliances. The growth of the semiconductor industry is predicted to continue even at a faster pace. Since the first integrated circuit was invented in the labs of Texas Instruments in 1958, the integration capacity of the transistors on a single chip is doubling every two to three years. In 1965, Gordon Moore showed that for any MOS transistor technology there exists a minimum cost that maximizes the number of components per integrated circuit. He also predicted that as transistor dimensions are shrunk from one technology generation to the next, the minimal cost point allows doubling the number of transistors every two to three years. This trend has been sustained and is expected to be maintained well into the first 20 years of this century [2].

Historically, technology scaling resulted in scaling of the transistor's dimensions by 0.7X each generation. Gate oxide also has been scaled to gain a better control over transistor characteristics. Supply voltage was kept constant is the so called "constant voltage scaling".

Not until reliability emerged to become an issue due to the continuous scaling of gate oxide, that the industry shifted into a different law of scaling. Constant field scaling has emerged in the early years of the last decade in order to keep a constant electric field inside the device. By then, electronics designers have started to face new challenges to keep scaling transistor dimensions as Moore predicted in his historic law.

1.1 Motivation for Low Voltage and Low Energy Design

As number of transistor is doubled every technology generation, chips grow in functionality and switching frequencies. The millions of parasitic capacitances charging and discharging at an ever increasing rate has led to a soaring amount of power dissipation. It has been shown that the usual scaling trend of transistors is facing three main challenges going forward [2]. The first and the most challenging is power dissipation. With clock speeds exceeding 4 GHz and switching millions of transistors, chip temperature has reached unprecedented levels requiring expensive packaging and heat dissipation techniques. Figure 1.1 shows that heat dissipation of modern processors is reaching the level of a hot plate. Serious reliability issues arise when working at such high temperatures [2].

Not until the last decade that power has started to become an issue that low power design has emerged to play an important role in modern VLSI design. Sakurai [3] showed that the trend of power dissipation of recent published microprocessors and digital signal processors (DSPs) is tapering off due to the limitation on power dissipation imposed by physical limits. Figure 1.2 shows that the early scaling trend for power dissipation was 4× every 3 years. The rate of power dissipation has changed to 1.4× every 3 years since heat is approaching the limit that current packaging technology is able to handle. The

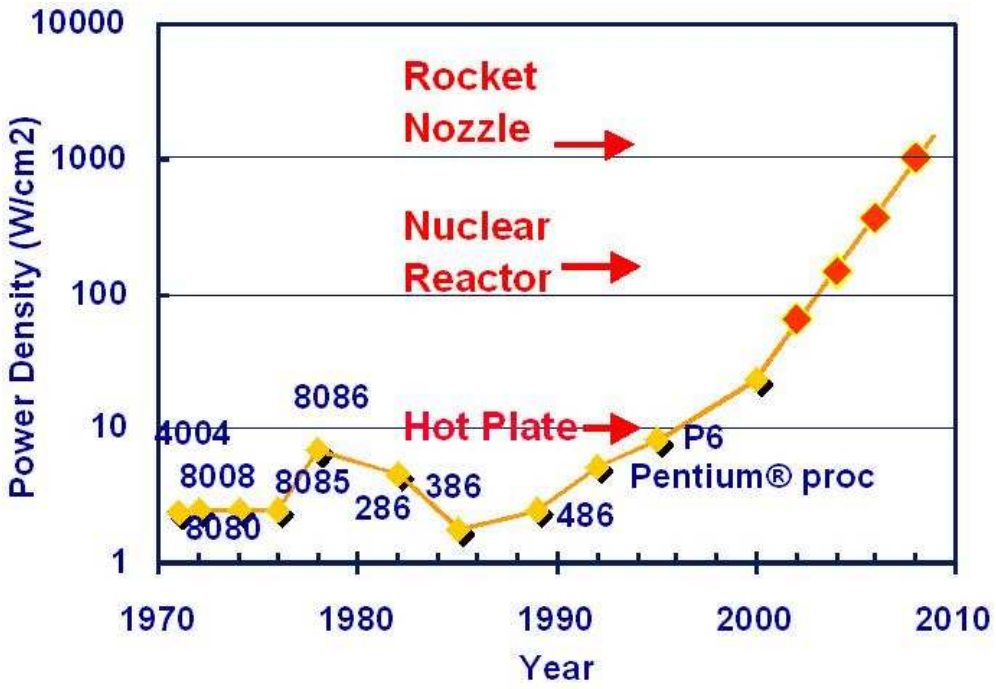


Figure 1.1: Power Density of modern microprocessors approaches that of the Hot plate.

International Technology Roadmap for Semiconductors (ITRS) predicts an even slower rate of power dissipation increase moving forward.

Over the last few years, there has been a growing interest in low power processors and DSPs as shown in Figure 1.2. Since the early 1990s, the increasing demand for portable devices such as cellular phones has driven the semiconductor industry into a new low power and low energy frontier. A limited amount of energy stored in a small battery requires extensive power management techniques to lengthen battery lifetime for as long as possible. On the other hand, battery capacity have grown at the very modest rate (2 to 3 times over the last 30 years) [4]. Keeping performance enhancements with a limited energy source is

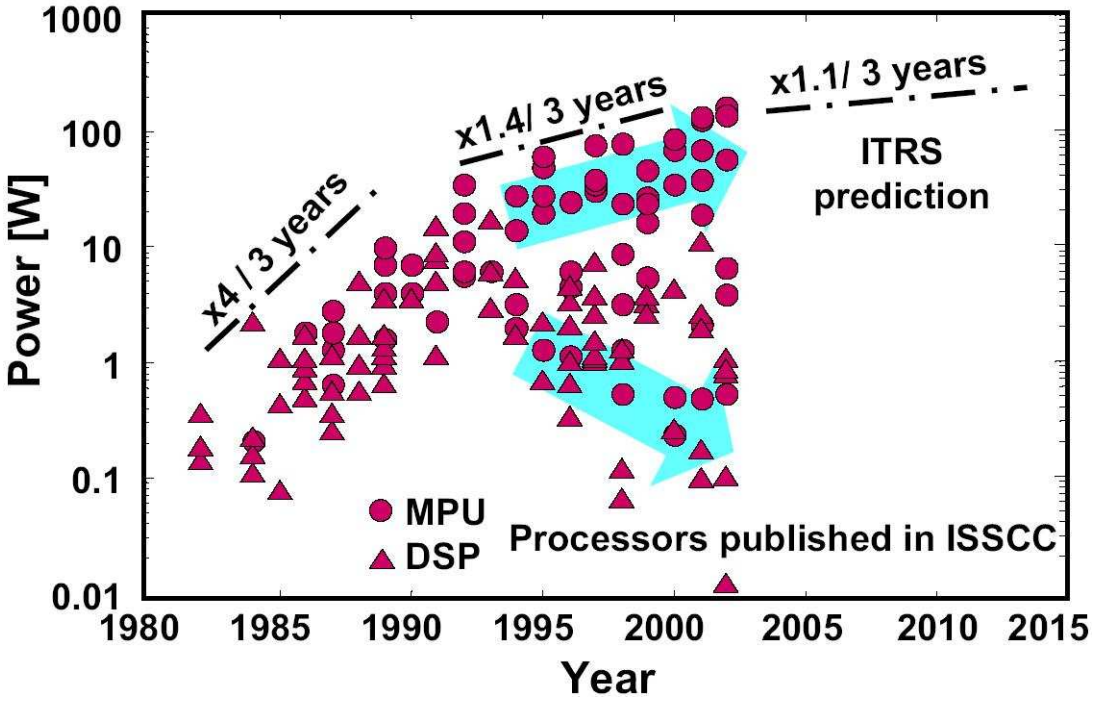


Figure 1.2: Power Dissipation increase is bounded with scaling.

a great challenge that faces low power designers.

Beside the above elements, another motive for energy efficient design is related to the environment. The information technology industry council estimated that electricity consumption of computers in the U.S. was about 13% of the total power in 1998 with an annual growth of $2-3\%$ [5]. That means in about a decade, power consumed by the IT industry would be 25% of the total power consumed in the U.S. As more microelectronics are being used in everyday's life, the demand on energy will sharply increase. Therefore, the lower the power consumption, the lesser the heat generated and so the lower the cost required for extra cooling systems in offices and homes. In this respect, energy efficient design facilitates competitive cost-to-performance ratio of the electronic equipment.

1.2 Thesis Organization

Energy efficient design often requires optimizations on all fronts and design levels. In this dissertation, new techniques to achieve energy efficient design on the architecture and on the circuit levels are presented. Chapter 2 presents an overview for the low power and low energy design aspects. The main power and energy reduction techniques are described. The concepts presented in Chapter 2 serve as a background and motivate the need for the work presented in later chapters. Moreover, low voltage, low energy circuit design is demonstrated through the DTPMOS technique. Low voltage is achieved by reducing the threshold voltage of the device dynamically during the active mode in order to increase current drive and speed. During the inactive mode, the threshold voltage is restored back to normal. Supply voltage applied to DTPMOS circuits is limited to 0.5V in order to limit the current resulting from the forward-biased drian/source to well junctions. The DTPMOS technique extends the concept of connecting the gate to the well usually used in Silicon On Insulator (SOI) technologies and applies it to bulk CMOS. Such a connection is possible in PMOS devices in the bulk technology. Shorting the well to the gate of the PMOS transistor helps improving its driving capability. However, since NMOS devices are connected to a common substrate, connecting the gate to the well is not possible.

Energy optimization of high speed circuits is addressed in Chapter 3. A new circuit technique suitable for scaled supply voltages in high-speed applications is presented. The Split Domino (SD) technique is a dynamic logic circuit technique. The high-speed advantage of dynamic logic is preserved by the SD technique while energy dissipation is reduced. The SD circuit technique offers reduced dynamic node capacitance and reduced contention at the start of the evaluation phase yielding better energy efficiency. In addition, a delay model for wide domino gates is presented. Model accuracy is close to HSPICE simulation. The model is used to examine different design tradeoffs early in the design stage to further

improve energy efficiency.

Chapter 4 focuses on supply scaling reduction as a mean for power and energy reduction. Two different architectures to control supply voltage dynamically based on performance requirements are presented. Dynamic Voltage Scaling (DVS) systems are often categorized into an open-loop and a closed-loop system. The open-loop system is based on a one-time voltage setting that accommodates worst case delay scenario. The closed-loop system relies on continuous monitoring of the actual system performance through on-chip structures. A hybrid between the one time voltage setting (open-loop) and the closed-loop system is presented. The hybrid system saves energy by detecting the actual silicon conditions and adjusting supply voltage at the closest point required to achieve the required performance.

The impact of interconnect delay is increasing as the feature size is continuously being shrunk. Selecting a single critical path for a system and monitoring its actual performance is growing in complexity. It is becoming common in modern VLSIs to see several paths that have close delays with different mixtures of logic and interconnect delay. These paths have different voltage scaling characteristics due to the difference between voltage scaling behavior of logic and interconnect delay. The traditional DVS approach is to select a certain path and add enough margin to it to guarantee that it remains the most critical at all times. Otherwise, the dynamic voltage scaling system would fail. Such a margin is growing as technology is scaled down due to the increasing impact of the interconnect delay. Chapter 4 presents a technique to mitigate the impact of interconnect delay on deep sub-micron dynamic voltage scaling systems. The proposed critical path emulator (CPE) system closely tracks the actual critical path of the system whose supply voltage is dynamically scaled. The CPE system reduces the margin required by conventional systems and therefore, is more energy efficient.

Experimental results for the open-loop and the closed-loop DVS systems described in Chapter 4 are presented in Chapter 5. Chapter 6 summaries the conclusions and thesis contributions.

Chapter 2

Low-Power, Low-Energy CMOS Design

2.1 Introduction

Low power and low energy have captivated circuit designers for the past few years in the quest for enhancing performance and extending battery lifetime. The increasing demand for integrating more functions with faster speeds is met by a slow increase in the capacity of batteries. For example, the third generation (3G) wireless protocol provides real-time streaming video at a high data rate on a 3G-enabled cellular phone. Such a computation intensive application can impact the battery life of the portable device. Therefore, the demand for increased battery life will require designers to seek out new technologies and circuit techniques to maintain high performance with longer battery lifetime.

Portable devices, however, are not the sole motive behind the low power and low energy design efforts. The increasing power dissipation for fixed supply devices is almost equally challenging as for portable devices. As technology feature size is reduced, the number of

transistors on the chip is increased and more power is dissipated. According to Moore's law, the number of transistors quadruples every two to three years. One hundred billion transistors on a single chip are projected before 2020 [6]. Expensive packing techniques are essential for dissipating such extensive power generated from that large number of transistors. Also, increased power dissipation has a negative impact on device's reliability.

Several methods for power and energy reduction have been proposed. Voltage supply, V_{DD}, scaling is considered one of the most effective elements in the process of reducing power dissipation in CMOS circuits. Threshold voltage, V_{TH}, has also to be reduced to maintain the required current drive. Reducing V_{TH} results in an exponential increase in leakage power. In order to keep leakage power under control, the ratio V_{DD}/V_{TH} tend to decrease with technology scaling.

The terms *low power* and *low energy*, although have different definitions, both serve to achieve the same objective. *Power* is defined as the average power supplied to a chip from the power supply and is measured in watts. Meanwhile, the term *energy* refers to the energy dissipated per operation and is measured in joules. In fact, energy can be expressed in terms of the *Power-Delay Product* (PDP), which is the product of power consumption and delay [7].

Table 2.1 shows the different strategies in converting a high-performance chip to a low-power chip using various power reduction methods [1]. The DEC Alpha 21064 chip operating at a supply voltage of 3.45V and consumes 26W of power at 200 MHz has been used as the starting point. As shown in the table, voltage supply reduction is the most effective among all other power reduction. When the supply voltage is scaled from 3.45V to 1.5V, power dissipation is reduced by a 5.3 $\times$. Function reduction comes in second with 3$\times$ reduction.

Shrinking device geometries introduces non-ideal device behavior in the form of short

Table 2.1: Strategies for converting a high-performance chip to a low-power chip [1].

Strategy	Power Reduction
V_{DD} reduction (3.45V→1.5V)	5.3×
Function reduction (Architectural level)	3×
Scale process (0.75μm→ 0.35μm)	2×
Clock Load reduction (Latches→Single edge-triggered FF)	1.3×
Clock frequency reduction (200MHz→160MHz)	1.25×

and narrow channel effects, drain-induced barrier lowering (DIBL), threshold voltage roll off. Producing low power, high performance, manufacturable transistors at low cost in deep-submicron (DSM) technology generations is growing in complexity. Further technology scaling problems arise due to inter and intra-die process variations.

2.2 Power Dissipation Components in Digital CMOS Circuits

Power consumption in CMOS circuits can be divided into three main components: short-circuit power, switching power, and leakage power. Short-circuit power arises when a conducting path between supply and ground is formed. The pull-up and pull-down devices should to be sized properly to achieve approximately equal rise and fall time. This component of power consumption can be significant in precharge and evaluate circuits, e.g. dynamic circuits. Careful design is required to keep this component of power dissipation

small enough to be ignored [8].

Switching power is a result of the power consumed in charging and discharging internal capacitances in the circuit. Leakage power is the power dissipated while the device is turned off. Leakage power has started to form a significant portion of the total power consumption as a result of the low threshold devices normally used in advanced DSM technologies. Figure 2.1 shows the increase in static (leakage) power for different technology generations [9]. It is apparent that static power is dramatically increasing with technology scaling. The ratio of leakage to total power is expected to exceed 50% in 45nm designs from about 10% in 90nm designs. Since switching and leakage power are the dominant components of power consumption, they are discussed in detail below.

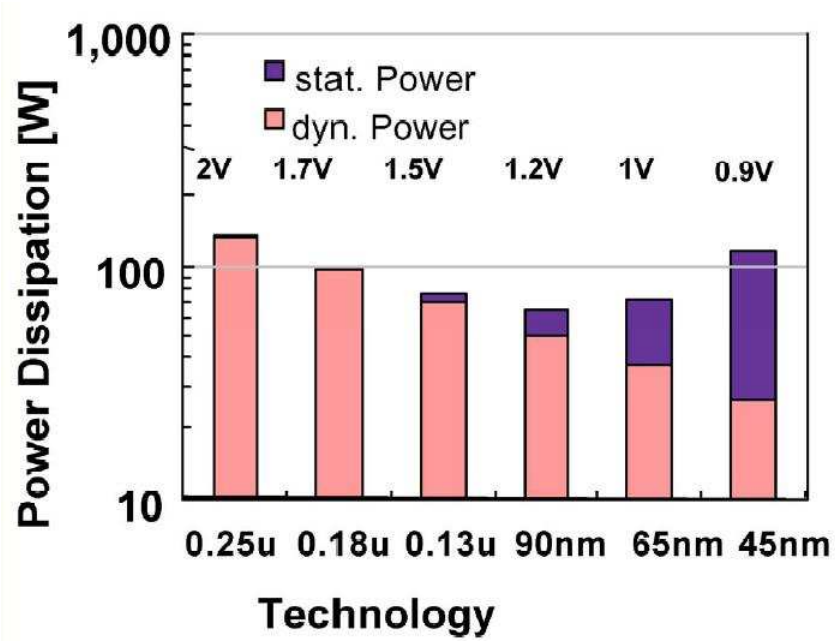


Figure 2.1: Static (Leakage) and Dynamic (Switching) Power for different technology generations.

2.2.1 Switching power

Switching power is the largest contributor to the total power dissipation in conventional CMOS technologies. It is a result of switching the junction, diffusion, and interconnect capacitances. Consider the CMOS inverter circuit in Figure 2.2. The parasitic capacitances are lumped into the output capacitor C. Consider the behavior of the circuit over one full clock cycle with the input going from V_{DD} to zero and back to V_{DD}. As the input switches from high to low, the NMOS pull-down transistor is turned OFF while the PMOS pull-up transistor is ON and capacitor C is charged. This charging process draws an energy equal to CV_{DD}^2 from the power supply. Half of this energy is dissipated immediately in the PMOS transistor, while the other half is stored on C. When the input switches from zero back to V_{DD}, the NMOS pull-down turns ON and the capacitance C discharges through it. If the rise time of the input signal is slow, both PMOS and NMOS are simultaneously ON causing a short circuit current to flow. This slow rise/fall time should be avoided through proper transistors sizing.

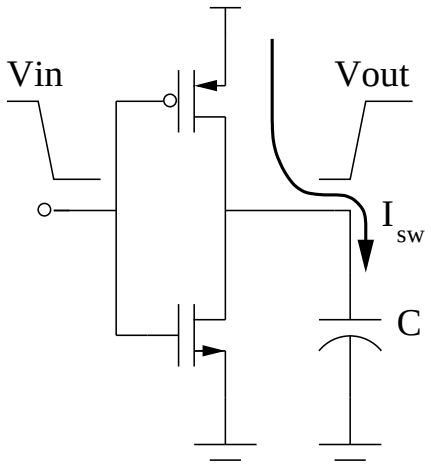


Figure 2.2: Switching Power in a CMOS Inverter.

For any logic gate, if inputs to the gate are assumed to switch at a rate of f times per second, then the average switching power for that gate is given by

$$P_{sw} = \alpha.C.\Delta V^2.f_{clk} \tag{2.1}$$

where α is the switching activity factor which represents the probability of the output switching from 0 to 1, C is the switching capacitance, ΔV is the voltage swing, and f_{clk} is the switching frequency.

Generally, α is less than one. As an example for activity factor computation, consider a 2-input NOR gate with equal probability of 0 and 1 at its inputs. The probability that the output becomes 0, is (3/4). While the probability the output is 1 would be (1/4). Therefore, the activity factor of the CMOS gate is given by the probability that the output is at 0 state (=3/4) multiplied by the probability the next state is 1 (=1/4). For the NOR gate, this translates to

$$\alpha = p(0)p(1) = p(0)(1 - p(0)) = \frac{3}{4}(1 - \frac{3}{4}) = \frac{3}{16} \tag{2.2}$$

Similar probabilities can be derived for other CMOS gates. In case of a logic network of several levels of gates, the activity factor of the gate becomes a function of its inputs probabilities.

For certain logic styles, however, glitching can form a non trivial part of the overall consumption. Glitching often arises when paths with unbalanced proportional delays converge at the same node in the circuit. If glitching due to signal races is to be accounted for, α might be greater than one [10]. Calculations of this activity in a circuit is very difficult and requires careful logic and/or circuit level characterization of the gates in a library as well as detailed knowledge of the circuit structure [4].

Obviously, reducing any term in (2.1) will result in a reduction in switching power.

However, low power techniques needs to address power reduction without affecting performance or device functionality. For example, frequency reduction is beneficial in terms of power consumption but it affects the overall system speed. Therefore, it is often a challenge to reduce power dissipation while maintaining the system performance.

2.2.2 Leakage Power

Leakage power forms a significant potion of the total power dissipation in DSM technologies. The different leakage current components are shown in Figure 2.3 [11]. I_1 is the reverse-bias p-n junction leakage caused by barrier emission and minority carrier diffusion and band-to-band tunneling. I_2 is subthreshold conduction current. I_3 results from the drain-induced barrier lowering (DIBL) effect. I_4 is gate-induced drain leakage (GIDL). I_5 is channel punchthrough. I_6 is hot carrier injection current. I_7 is oxide leakage. I_8 is gate current due to hot carrier injection. I_1 through I_6 are OFF currents while I_7 and I_8 are ON and switching currents. Here, the main concern is the OFF leakage current and therefore, the focus is on the current components I_1 through I_6 which are explained below [12].

- Junction Reverse Bias Current (I_1): I_1 has two components: One is minority carrier diffusion/drift near the edge of the depletion region, and the other is due to electron-hole pair generation in the depletion region of the reverse biased junction [13]. Heavily doped junctions are also prone to Zener and band-to-band tunneling. The p-n reverse bias leakage is a function of junction area and doping concentration. I_1 is normally a minimal contributor to total OFF current.

- Subthreshold Conduction Current (I_2): Subthreshold conduction or weak inversion current between source and drain when supply voltage is below V_{TH}. The subthreshold current occurs due to carrier diffusion when the gate-source voltage, V_{GS}, has

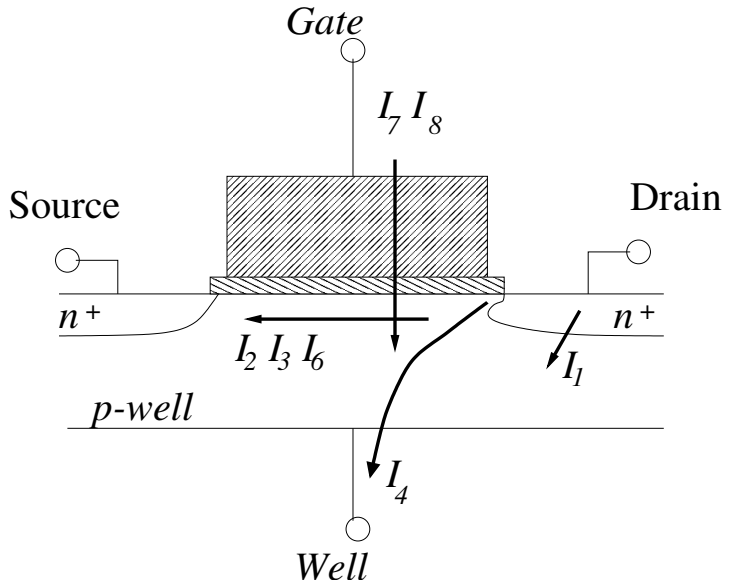


Figure 2.3: Leakage Current Components.

exceeded the weak inversion point, but still below the threshold voltage, where carrier drift is dominant. Subthreshold conduction typically dominates modern device off-state leakage due to the low threshold devices.

- Drain-Induced Barrier Lowering, DIBL (I_3): DIBL is the effect of lowering the source potential barrier near the channel surface as a result of the applied drain voltage. Ideally, DIBL does not change the subthreshold slope but does lower V_{TH}. Higher surface and channel doping, and shallow source/drain junction depths work to reduce the DIBL mechanism.

- Gate-Induced Drain Leakage, GIDL (I_4): GIDL current arises in the high electric field under gate/drain overlap region causing a thinner depletion region of drain to well junction. GIDL results in an increase in leakage current when applying a negative

voltage to the gate (NMOS case). GIDL is small for normal supply voltage but its effect rises at higher supply voltages (near burn-in).

- Punchthrough (I_5): Punchthrough occurs when source and drain depletion regions approach each other and the gate voltage loses control over the channel current in the subgate region. Punchthrough current varies quadratically with drain voltage. Punchthrough is often regarded as a subsurface version of DIBL.

- Narrow width effect (I_6): Threshold voltage tends to decrease in trench-isolated small effective channel width devices. The narrow width effect causes the threshold voltage to decrease in trench isolated technologies for channel widths on the order of $W \leq 0.5\mu$m. It can be ignored for device sizes $>> 0.5\mu$m.

Subthreshold leakage current is the largest leakage current component. It increases exponentially as a result of threshold voltage reduction. In a simple form, subthreshold leakage current, I_{sub}, is given by

$$I_{sub} = I_0 e^{\frac{V_G - V_S - V_{TH0} - \gamma V_s + \eta V_{DS}}{nV_T}} \left(1 - e^{\frac{V_{DS}}{V_T}}\right) \tag{2.3}$$

where V_{TH0} is the zero-bias threshold voltage, γ is the linearized body effect coefficient, η is the DIBL coefficient, V_T is the thermal voltage (26 mV at room temperature), and I_0 is a constant proportional to V_T and transistor dimensions.

Various techniques have been developed to keep both active and leakage power under control. In the next section, some of the effective power and energy reduction methodologies are described. The intent is to focus on these particular methodologies since the work presented in this thesis builds on these methodologies.

2.3 Power and Energy Reduction Techniques

Since switching power is the major source of power dissipation in CMOS technologies, various techniques have been proposed on a variety of design levels to achieve switching power reduction. Considering a top-down design paradigm, power and energy reduction can be achieved on the architecture, circuit, and device levels. Starting at the top level, the architecture is modified to lower power dissipation by introducing or adding parallelism or pipelining. When such modifications are implemented, power can be reduced via supply or frequency scaling. Moving down the design paradigm, both circuit and device level optimizations are required to enable energy efficient operation.

2.3.1 Supply Voltage Reduction

Many designers have focused on power supply reduction as a mean for low power operation. By noting the three parameters that appear in (2.1), it is obvious that reducing frequency or switching capacitance provides a linear reduction in switching power. However, supply voltage reduction leads to quadrable savings. Moreover, subthreshold leakage current can be reduced exponentially with supply voltage reduction. As can be seen from (2.3), both V_G and V_{DS} are reduced when supply is scaled yielding an exponential scaling of subthreshold leakage. In [3], it was shown that both dynamic and leakage power can be effectively reduced through supply scaling .

Voltage reduction enables architectural level power optimizations. Parallelism or pipelining can be employed to reduce power dissipation [10] [14]. Consider the *multiply and accumulate* (MAC) structure shown in Figure 2.4 (a). Assume that the clock period for maximum throughput at normal supply voltage is T. Using a duplicated MAC unit in parallel with the original one, the clock frequency can reduced by half (doing the computa-

tions in parallel) as shown in Figure 2.4 (b). Slashing the operating frequency by half can allow for a 40% reduction in the supply voltage (this reduction might vary from design to design and from one technology to another). Due to the parallelism used, the capacitance increases by a factor of 2 as a result of using a duplicated MAC. In addition, capacitance increases by another 20% due to the extra routing required. Therefore, the resulting reduction in power consumption of the parallel architecture compared to the original one is given by

$$\begin{aligned} P_{\text{parallel}} &= CfV^2 = (2.2C_{\text{org}})(0.6V_{\text{org}})^2(0.5f_{\text{org}}) \\ &= 0.4P_{\text{org}} \end{aligned} \tag{2.4}$$

where C_{org} is the original effective capacitance being switched per clock cycle. Apparently, the main restriction on using parallelism to reduce overall power is the area. A considerable part of the extra area required for parallelism is the extra routing area. Wiring capacitance represents a significant part of the total capacitance of a chip. In addition, wiring capacitance does not scale as much as the feature size. Therefore, careful optimization and sophisticated routing techniques have to be utilized to fully exploit the advantage of parallelism and minimize its side effects.

For area-constrained designs, pipelining is a viable option with much less area overhead compared to parallelism but yet a comparable throughput. By adding two extra latches at the adder inputs as shown in Figure 2.4 (c), the minimum clock period is reduced to that of the multiplier (assuming that the adder delay is less than that of the multiplier). Assuming that the clock frequency can be reduced by only 20% instead of 40% in case of parallelism for the static CMOS MAC architecture, this reduction in the clock frequency would leave a room for supply voltage reduction to get the same throughput of the original structure. Supply voltage can then be reduced by approximately 15%. The area overhead represented by the extra latches results in a switching capacitance increase of 15% instead of 220% in

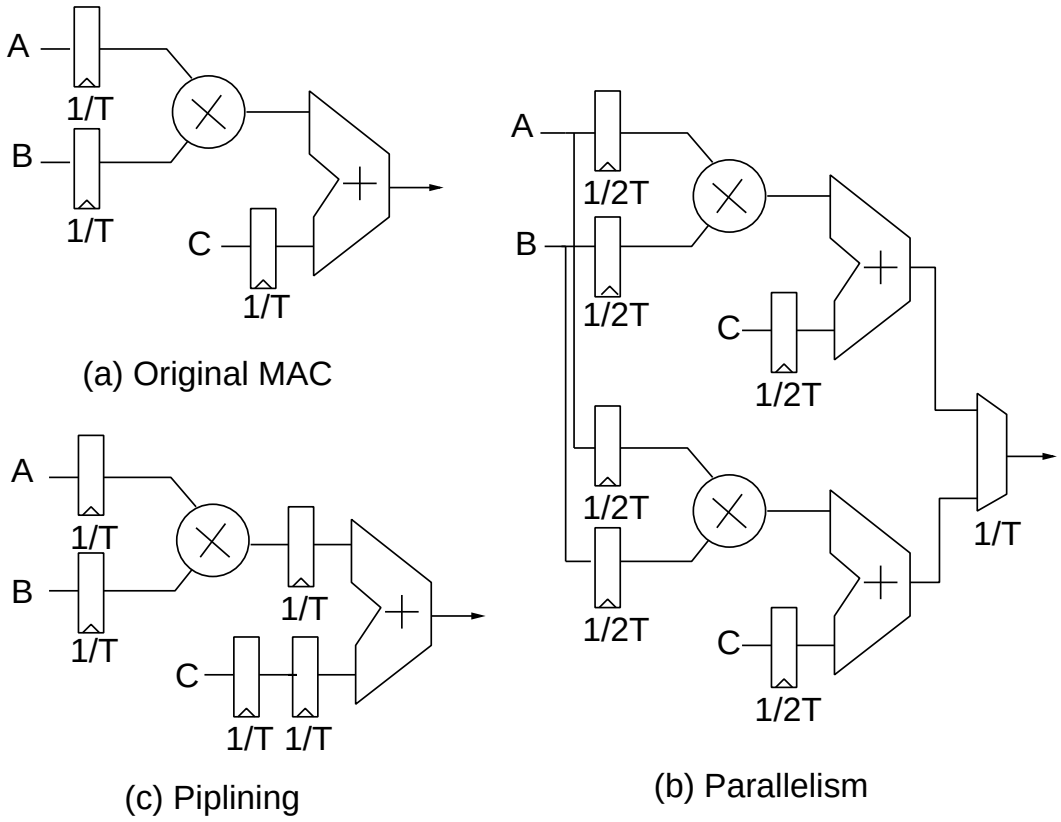


Figure 2.4: Parallelism vs. Pipelining.

the case of using parallelism. Therefore, the power reduction in case of pipelining would be

$$\begin{aligned} P_{\text{pipeline}} &= CfV^2 = (1.15C_{\text{org}})(0.85V_{\text{org}})^2(0.8f_{\text{org}}) \\ &= 0.65P_{\text{org}} \end{aligned} \tag{2.5}$$

The power reduction is less than that of the parallel structure. Balancing the delay of all the pipelined stages is extremely important to achieve further reduction in power. That would allow for more supply voltage reduction and hence more power savings. In addition, increasing the level of pipelining also reduces the logic depth and hence the power

contributed by hazards and critical races.

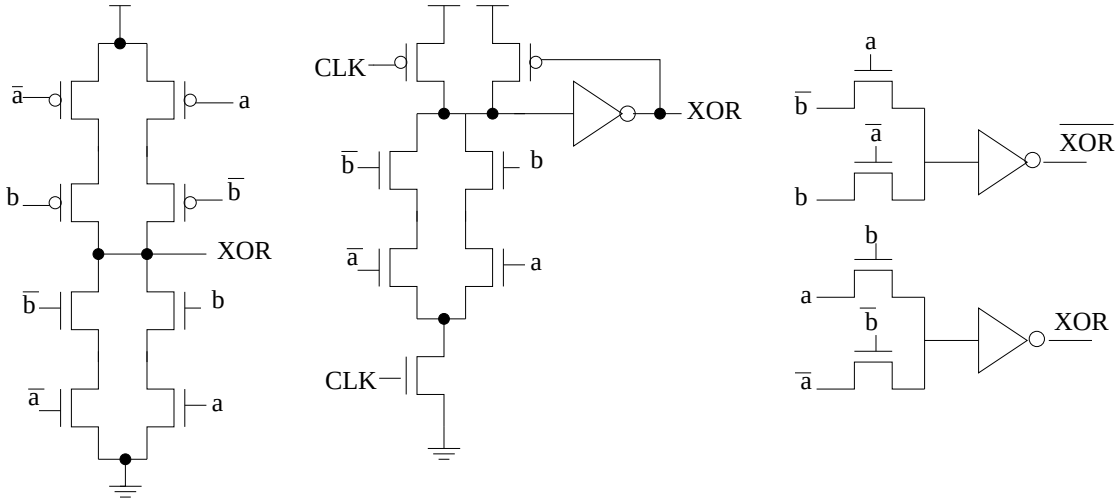
Furthermore, exploiting both pipelining and parallelism is more attractive. This architectural choice results in further speedup and more room for supply voltage reduction. This combination, given no restriction on area, would allow for more power savings.

2.3.2 Circuit Level Techniques

Different static and dynamic logic styles have been introduced for the sole aim of reducing power. It is also a common design practice to combine both static and dynamic logic styles to optimize for delay and power at the same time. The merits of each logic style are explained below.

- Conventional CMOS logic style: Static CMOS logic refers to conventional CMOS circuits which are constructed using an NMOS pull-down network and a complementary PMOS pull-up network as shown in Figure 2.5 (a). Due to the complementary nature of the circuit, conventional CMOS logic style is inherently able to reject noise. Therefore, static CMOS is robust against voltage scaling and transistor sizing. Input signals are connected to the gate terminals, which facilitates the usage and characterization of logic cells. The layout of CMOS gates is simple and regular due to the similar, yet complementary, pull-up and pull-down network structure.

  On the other hand, conventional CMOS suffers from inherent disadvantages due to the pull-up PMOS network. One of the main disadvantages is the increased gate capacitance resulting from the large size PMOS transistors. Furthermore, the PMOS transistor is usually made larger to compensate for the speed difference with respect to the NMOS due to the lower hole mobility compared to electron mobility. However, this disadvantage is diminishing as technology feature size is shrunk. The carrier drift



(a) CMOS (b) Dynamic (c) CPL

Figure 2.5: Logic Styles

velocities of both PMOS and NMOS approach the saturation velocity and therefore the size ratio between PMOS and NMOS devices is quickly approaching one [15]. Another drawback of static CMOS logic is the relatively weak driving current. By adding output buffers, the driving current can be enhanced.

- Dynamic Logic Style: Dynamic logic operates in two phases: precharge and evaluation. During the precharge phase, the CLK signal charges up the dynamic node (shown in Figure 2.5 (b)). During the evaluation phase, the CLK signal switches High. Depending upon input values, the dynamic node is discharged or remains charged. Dynamic logic is usually faster than static CMOS due to less capacitance (PMOS network is eliminated). However, dynamic logic consumes more power. Many dynamic logic styles with improved delay and power compared to the conventional dynamic style shown in Figure 2.5 (b) have been reported. Some of these design

styles will be discussed in more details in Chapter 3.

- Pass-Transistor Logic Style: Unlike static and dynamic logic, pass-transistor logic provides complementary output. Moreover, inputs are connected to both the gates and the sources of transistors. Pass-transistor gates have two input categories: pass inputs and control inputs. Pass inputs are connected to the sources of the devices while control inputs are connected to the gates. The strongest advantage of pass-transistor implementation is that it can use just one network, usually NMOS, to build the logic. Also, the dual rail nature of the logic style can be used efficiently to implement multiplexing functions. However, connecting some inputs to the source causes a V_{TH} drop. As a result, the voltage swing is reduced and it requires restoration at the output stage to increase noise margin and to minimize short circuit currents. As a consequence, two NMOS networks would be used in addition to the output buffering circuitry. This overhead annihilates the advantage of the low transistor count and small input capacitance. Moreover, pass-transistor logic is sensitive to voltage scaling and transistor sizing. Finally, the layout of pass-transistor logic is complicated due to the extra wiring normally required. One example of pass-transistor logic is the complementary pass-transistor logic (CPL) shown in Figure 2.5 (c). CPL has two NMOS networks, one for each rail, and two inverters for level restoration [16]. CPL has small input capacitance, a fast differential output stage, and a high driving current. However, CPL, as a member of the pass-transistor logic family, suffers from short circuit currents at the output and wiring complexity due to the dual rail. Other pass-transistor logic styles have been proposed. A good comparison between the different styles can be found in [17]. In [17], static CMOS has been shown to have superior performance over pass-transistor logic. Therefore, static and dynamic logic usually occupy a larger share of the circuit design space.

From a low power perspective, static logic dissipates less power compared to dynamic logic due to the following reasons:

1. Spurious Transitions: Static designs are prone to spurious transitions more than dynamic circuits due to critical races and dynamic hazards in static logic. The magnitude and the number of those undesirable transitions in a logic structure is a function of the logic design, delay skew, and logic depth. For example, an 8-bit ripple carry adder consumes an extra 30% of power due to spurious transitions [10]. Dynamic logic, however, intrinsically does not suffer from spurious transitions, since any node can undergo at most one power-consuming transition per clock cycle.

2. Switching Capacitance: Dynamic logic has fewer devices, typically $N+2$ for N-input gate compared to $2N$ in case of CMOS. This is reflected directly on the switching capacitance and thus has a direct impact on delay and power dissipation.

3. Switching Activity: Dynamic logic is notorious for its high switching activity. The dynamic node has to be precharged every clock cycle even if it going to be discharged immediately after evaluation starts. For example, for a 2-input dynamic NOR gate, the switching activity is (3/4) compared to just (3/16) in case of static logic implementation. If spurious transitions are neglected, then using dynamic logic would result in a 4 times increase in power. But if reduction in capacitance and spurious transitions are taken into account, the resulting power increase would not be that dramatic.

With fewer transistors required to implement a certain dynamic logic function compared to static logic, standby leakage current of dynamic logic can be less than its static logic counterpart. In some applications where fast evaluation time is followed by a long idle period, dynamic logic can be more attractive than static logic for its low standby leakage.

2.3.3 Device Level Optimizations

As mentioned before, CMOS is regarded as the technology of choice for low power and low energy applications. It offers a good performance and a considerable stability. However, as supply voltage is reduced, threshold voltage has to be reduced to maintain the required performance. A reduced threshold voltage directly results in an exponential increase in subthreshold current.

Some technologies have been offering a solution for the increase in subthreshold current resulting from the reduced threshold voltage. Silicon on insulator (SOI) technology has emerged with a good potential in low power and low voltage applications. A simple SOI device structure is shown in Figure 2.6 (a). In SOI technology, the thin film is totally isolated from the body by a thick film oxide. The thick oxide serves to suppress the radiation induced current. Also, due to the thick oxide layer, the gate to source/drain capacitance is greatly reduced. As a consequence, SOI devices are faster and consume less dynamic power compared to CMOS. In terms of integration and technology down scaling, the depletion regions in bulk CMOS which are used for isolation put a lower limit on feature size in bulk CMOS. The buried thick oxide in SOI makes it easier to down scale device dimensions.

Figure 2.6 shows two additional SOI structures. DTMOS SOI and DGSOI are shown in Figure 2.6 (b) and (c) respectively. DTMOS refers to the Dynamic Threshold MOS structure proposed in [18]. In the DTMOS structure, the gate is tied to the body of the SOI device. This type of connection allows for low threshold during the ON state and high threshold during the OFF state. The DGSOI is a Double-Gate SOI device in which there is a back gate separated from the body of the device by the back oxide [19]. The DGSOI has a higher current drive for high output load in addition to an excellent ability of leakage control [20].

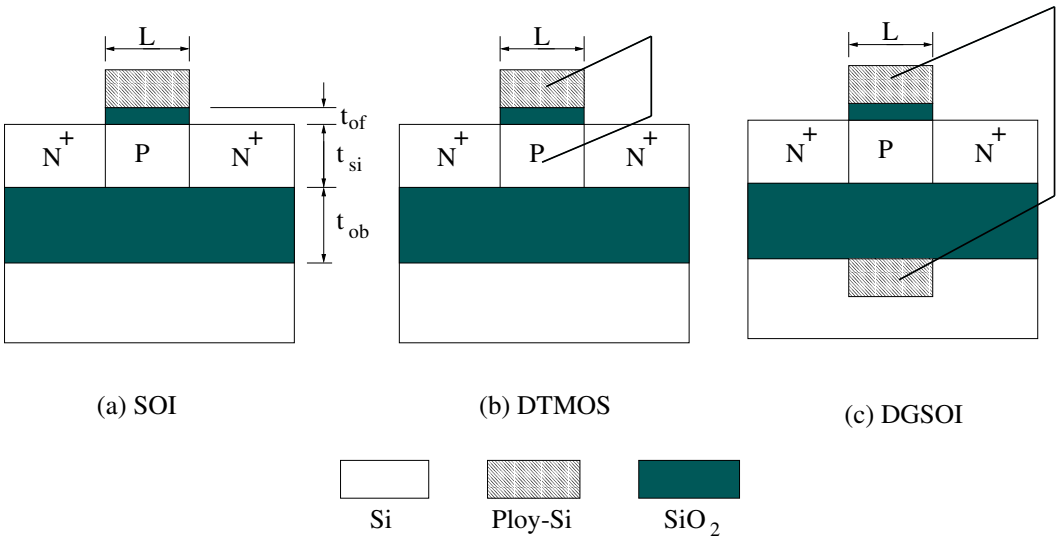


Figure 2.6: Silicon On Insulator (SOI) devices.

However, the history-effect of SOI devices, and low thermal conductivity of the buried oxide which results in an increase in temperature are among the drawbacks of using SOI technology. Further development and innovations are required to enable a cost-effective and efficient SOI solution.

In addition to switching power, leakage power is forming an increasing portion of the total power dissipated in modern technologies, several techniques have been developed to reduce its impact. Some of these techniques are summarized in the next section.

2.4 Leakage Reduction Techniques

Modern DSM technologies are suffering from a dramatic increase in leakage current. Constant field scaling dictates that the supply voltage has to be reduced when downsizing the technology feature size. Low threshold voltage devices are used to maintain the required

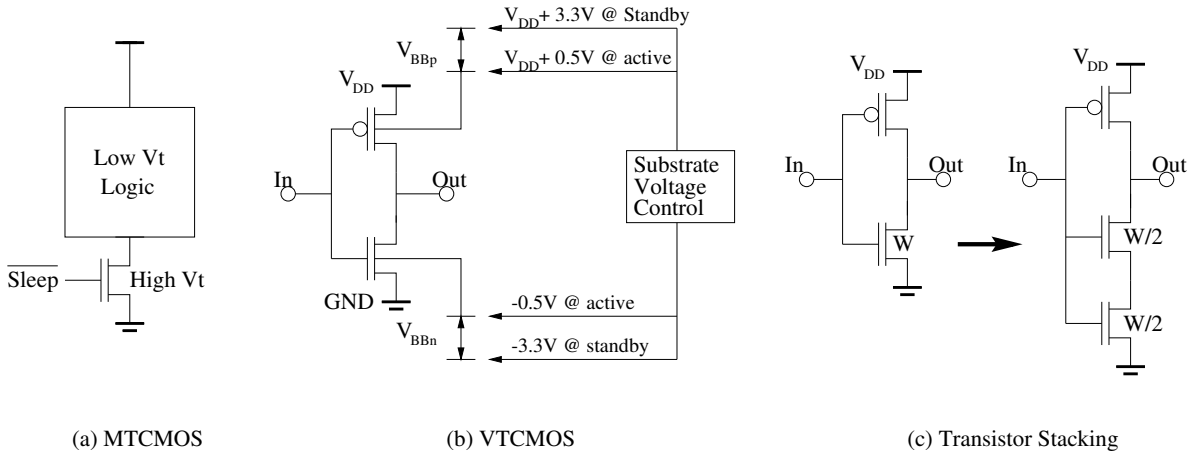
current drive and to satisfy performance specifications. Low threshold devices have caused a dramatic increase in leakage current. A direct and effective solution for that is to utilize low threshold devices in the critical path and high threshold devices elsewhere. The threshold voltage can be controlled utilizing the well bias of the device in the so called *Variable Threshold CMOS* (VTCMOS) [21].

Dual threshold technology is another way to address the increasing active and leakage power problem. The technology is a CMOS process with two types of devices, low threshold and high threshold device. Performance is enhanced by placing the low threshold devices on the critical path to increase performance and place the high threshold devices on the non-critical paths to decrease leakage. Several mechanisms have been developed to optimize the process of placing the low/high threshold devices on the gate level such as in [22] or on the transistor level such as in [23] and [24]. This method was presented in [25] and referred to as *Multi-Threshold CMOS* (MTCMOS). These two methodology are discussed in detail below. Some recent enhancements and design considerations are also summarized.

2.4.1 Multi-Threshold CMOS (MTCMOS)

The leakage current can be dynamically controlled using multi-threshold devices as was proposed in [25] and is shown in Figure 2.7 (a). In this scheme, low V_{TH} logic is used for faster evaluation while a high V_{TH} NMOS device, *Sleep* device, is used to disconnect the logic from the supply during standby. A *Sleep* control signal is used to turn the high V_{TH} NMOS device ON and OFF depending upon the mode of operation. A clear drawback of this technique is the impact of the *Sleep* device sizing on performance. Increasing the *Sleep* transistor size more than necessary would add to the circuit capacitance and power dissipation while sizing it too small would result in a supply current limitation and speed degradation. Another potential problem in the MTCMOS scheme is the bounce of virtual

ground line bouncing. In fact, the capacitance of the virtual ground line is much larger than that of the real ground resulting in a ground bounce. This bounce adversely affects both noise margin and delay. A methodology for properly sizing the *Sleep* device to minimize the delay based on mutual exclusive discharge patterns was proposed in [24].



(a) MTCMOS (b) VTCMOS (c) Transistor Stacking

Figure 2.7: Leakage Reduction Techniques

The advantage of low leakage during standby mode is stressed by back biasing the sleep transistors to more than V_{DD} [26] [27]. By reverse biasing the body of the sleep transistor, threshold voltage is increased and leakage current is decreased. Therefore, a low threshold voltage device can be used without an increase in leakage current during standby. The low threshold sleep device limits voltage drop during the active mode and provides more current drive. Improving the current drive during the active mode is highly desirable in order to achieve more speed. By increasing the voltage swing of the gate of the sleep transistor, the gate-to-source voltage becomes greater than zero and boosts the current drive [28].

2.4.2 Variable-Threshold CMOS (VTCMOS)

VTCMOS technique uses all low threshold devices [21]. However, the threshold voltage is controlled using the well bias of the devices in a triple-well CMOS process. During the ON state, the well bias is $V_{DD} + 0.5V$ for the p-well and $-0.5V$ for the n-well allowing for low threshold voltage realization as shown in Figure 2.7 (b). During standby, the source-body junction is strongly reverse biased to increase the threshold voltage and to reduce leakage current. The p-well bias is set to $V_{DD} + 3.3V$ while the n-well bias is set to $-3.3V$. Consequently, V_{TH} is adjusted to be 0.77V during the active mode and greater than 0.5V during the standby mode. One potential problem with this approach is that the threshold voltage varies as the square root of the body-source voltages. Therefore, the body-source voltage has to significantly increase to change the threshold voltage to a relatively higher value. VTCMOS is even more efficient in leakage current suppression for series connected transistors due to the increased body-effect [29].

VTCMOS scheme depends on a high body-effect to control the threshold voltage. With technology scaling, the body-effect is reduced from one technology generation to the next. The body effect is primarily reduced due to the short channel effects. Techniques such as well doping can be applied to enhance the short channel effects. However, well doping causes the doping levels in the vicinity of source-body and drain-body junctions to increase significantly. As the doping limit approaches the tunneling limit, the junction current increases exponentially, and becomes the dominant leakage component. Therefore, body-effect is reduced and limits the effectiveness of the VTCMOS scheme [30].

SOI technology can also be used in the implementation of VTCMOS. In [31], a silicon-on-insulator-with-active-substrate (SOIAS) was used to dynamically control the threshold voltage. The dynamic threshold MOS (DTMOS) scheme is another mean to provide low threshold during the ON state and high threshold during the OFF state [18].

A summary of the different features of the MTCMOS and VTCMOS techniques is presented in Table 2.2 [32]. Moving towards smaller feature size, the MTCMOS technique seems to be a better choice. However, VTCMOS is more effective in reducing process variations which are increasing with technology scaling. Therefore, the choice between MTCMOS and VTCMOS is application dependent.

Table 2.2: VTCMOS vs. MTCMOS techniques

	MTCMOS	VTCMOS
Principle	Sleep-mode switch	Well-bias threshold control
Low leakage in standby	√	√
Products are already rolled out	√	√
Scalability	√	×
V_{TH} fluctuations compensation	×	√
I_{DDQ} testing	×	√
Serial Sleep Device	√	×
	slower, lower yield..	
Sleep-mode storage	Dual V_{TH} FF's	Conventional FF's
Process	Dual threshold	Triple Well or SOI

2.4.3 Transistor Stacking

Narendra *et.al.* [33] examined the effect of transistor stacking on subthreshold leakage current reduction. It has been shown that the stacking effect increases as the technology

scales. Therefore, by forcing transistor stacking as shown in Figure 2.7 (c), speed can be traded for leakage reduction. The stack effect can reduce leakage current by 2 orders of magnitude for low V_{TH} devices and 3 orders of magnitude for high V_{TH} devices [33]. Stacking of transistors during the standby mode can be accomplished by controlling the sleep mode input vector to maximize the number of transistors is the stack during sleep. Such a technique has a negligible speed penalty. However, the minimum leakage state is difficult to achieve by using a specific vector that maximizes the use of stacking since it is not a default feature in all logic gates (e.g. Inverter, Nor, etc.). By combining the use of sleep vector control and forcing stacks in stackless structures, a 30-90% reduction is leakage can be achieved [34]. In addition, transistor stacking has been shown to effectively reduce gate leakage [35].

2.4.4 Gate level leakage reduction

Wei *et.al.* [22] proposed a gate level optimization method for leakage reduction. In their work, gates are divided into groups, one is low threshold and the other is high threshold. Gates in the critical path are low V_{TH} for faster evaluation while non-critical path gates are high V_{TH} to reduce leakage. The optimization method is run iteratively to find the optimum gate assignment for the a given V_{TH} value. Leakage reduction through the use of multiple supply voltages can be also achieved. The normal supply voltage is assigned to the gates on the critical path while reduced voltages are applied to gates not on the critical path [36]. An earlier work was proposed in [37] where optimization of supply and threshold voltages are performed to achieve low power implementations.

2.5 Ultra-Low Voltage Circuit Techniques

Voltage supply, V_{DD}, reduction has been utilized to achieve low power, energy efficient operation due to the quadratic relationship between power and V_{DD}. The transistor's threshold voltage, V_{TH}, is often reduced to maintain a decent performance. In a limiting case, fully static CMOS logic works when V_{DD} is slightly greater than max $\{\mid V_{TH_p} \mid, V_{TH_n}\}$ where V_{TH_p} and V_{TH_n} are the threshold voltages of the PMOS and NMOS transistors respectively. However, when V_{DD} is reduced below that value, the switching delay increases appreciably.

Reducing supply voltage is often accompanied by threshold voltage reduction in order to prevent current drive degradation and the resulting delay increase. Assaderaghi *et.al.* [18] introduced the concept of dynamically controlling the threshold voltage (DTMOS) by connecting the gate of the MOS transistor to its substrate in silicon on insulator (SOI) technology. Threshold voltage is reduced during the active mode and is restored back to normal during the standby mode. This results in a significant speedup during the active mode and normal standby leakage current. However, an exponential increase in active mode leakage current is observed due to threshold voltage reduction. A potential remedy to this increase in leakage current is to use one of the leakage reduction techniques described earlier.

The DTPMOS technique extends the concept of dynamic threshold to bulk CMOS technologies. However, only the gate of the PMOS transistor is connected to the well. This type of connection can be implemented in bulk CMOS since each PMOS transistor is implemented in a separate well isolated from other PMOS transistors. This technique allows for energy efficient realizations of digital blocks working at sub-0.5V. The DTPMOS technique is described below. A technique to mitigate the active mode leakage current is also introduced.

2.5.1 Dynamic Threshold PMOS (DTPMOS) Scheme

The proposed concept relies on the connection between the gate and the well of PMOS transistors to reduce V_{TH_p} during the on-state and maintain a high V_{TH_p} during the off-state. For simplicity, the threshold voltage of the DTPMOS transistor will be denoted V_{TH}. The dynamic nature of the DTPMOS threshold voltage can be explained using the expression

$$V_{TH} = V_{TH_0} - \gamma(\sqrt{|\ -2\Phi_F\ |} - \sqrt{|\ -2\Phi_F + V_{BS}\ |}) \tag{2.6}$$

Here V_{TH_0} is the threshold voltage at zero body bias, γ is the body effect coefficient, $2\Phi_F$ is the surface potential at strong inversion, and V_{BS} is the body-source voltage. The minus sign of the body effect coefficient in (2.6) is due to the forward biased body-source junction [18]. During the on-state and assuming that V_{DD} is 0.5 V, V_{BS} for conventional CMOS is zero while it is -0.5 V for DTPMOS. Assuming that V_{TH_0} is -0.435 V, γ is 0.5667 and $2\Phi_F$ is 0.6 V, V_{TH} is reduced to -0.28 V (36% reduction) compared to its value at zero body bias. During the off-state, however, V_{BS} is set back to zero and V_{TH} returns to its original value at zero body bias, V_{TH_0}. The low threshold voltage in the on-state leads to a significant reduction in delay at a low voltage supply.

Compared to conventional CMOS, DTPMOS results in a higher PMOS current drive and consequently a higher operating speed at a very low voltage. This is mainly due to a larger inversion charge and a lower effective normal field in the channel. The lower effective normal field leads to higher mobility and consequently higher current drive [18]. The main features of the DTPMOS scheme are discussed below by applying the scheme to the different building blocks of a parallel multiplier. Performance, standby power, and energy comparisons of DTPMOS and conventional CMOS are also discussed.

2.5.2 DTPMOS Implementation of Parallel Multiplication Building Blocks

The parallel multiplier is one of the most analyzed structures in digital VLSIs. Several multiplier architectures and implementations have been proposed for low power applications [38, 39, 40, 41]. In general, the parallel multiplier architecture can be divided into three blocks: the partial product (PP) generator, the summation network, and the final adder. Modified Booth algorithm (MBA) is used for PP generation. MBA is implemented using Booth encoders and Booth selectors. Full adders (FAs) are used to implement a carry save addition tree in the summation network. Finally, a carry skip adder is used to produce the final product. The multiplier circuit is designed using minimum size transistors in most of the instances to minimize the power consumed. Non-minimum size transistors are used where the load and/or the fanout is high. Furthermore, the pass-transistor logic has been extensively used to further achieve lower power operation.

In order to explore the characteristics of operation of the DTPMOS scheme, a pass-transistor full adder (FA) circuit is implemented in both DTPMOS and conventional CMOS. Figure 2.8 shows the DTPMOS implementation of the FA circuit.

HSPICE simulations for the FA circuit for both the DTPMOS and conventional CMOS schemes are carried out in the 0.18 μm CMOS technology. The input frequency is 10 MHz. This speed is adequate for certain applications specially hearing aids where energy is a very critical design constraint and the typical operating speed is 1-2 MHz [42][43]. Another area of application is sensor networks where battery life is expected to last for years [44] [45]. The simulation setup is to connect all the outputs of the FA circuit (*simulated* circuit) to inputs of a similar FA circuit (load circuit). The outputs of the load circuit are connected to 10 fF loads.

A comparison between the DTPMOS FA and the conventional CMOS FA is illustrated

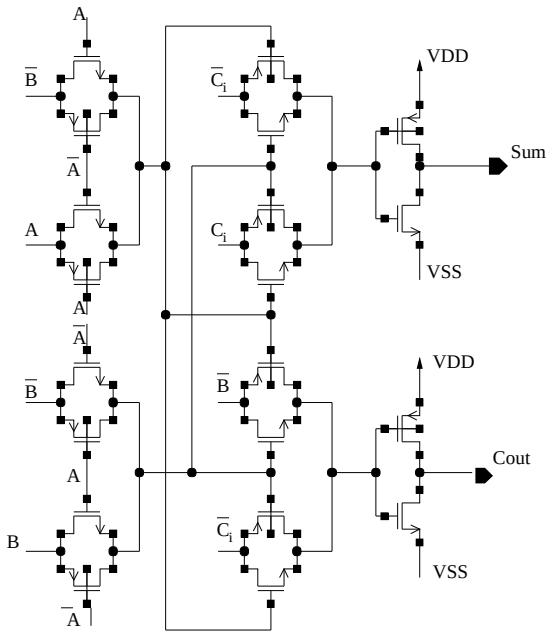
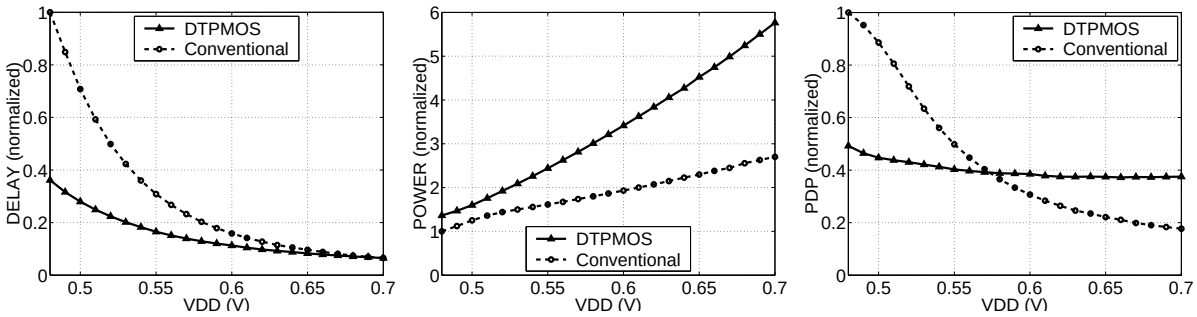


Figure 2.8: DTPMOS Full Adder circuit.

in Figure 2.9. Figure 2.9 (a) shows that using DTPMOS is beneficial below the 0.5V supply voltage. Delay of DTPMOS is 60% less than that of conventional CMOS at a supply voltage of 0.48 V. The DTPMOS delay advantage decreases as supply voltage goes higher than 0.6 V. This is a result of a significant increase in the active leakage current due to the forward biased source and drain junctions. As supply voltage exceeds the built-in junction potential, the excessive current flowing through the forward biased junctions has a small positive impact on delay while causing a large dissipation of power. Consequently an increase in power and power-delay product (PDP) is expected. This trend is shown in Figure 2.9 (b) and (c). Power dissipation of DTPMOS is almost double that of the conventional CMOS at 0.7 V. With a small delay enhancement of the DTPMOS scheme a 0.7 V, PDP of conventional CMOS is approximately half that of DTPMOS. This trend

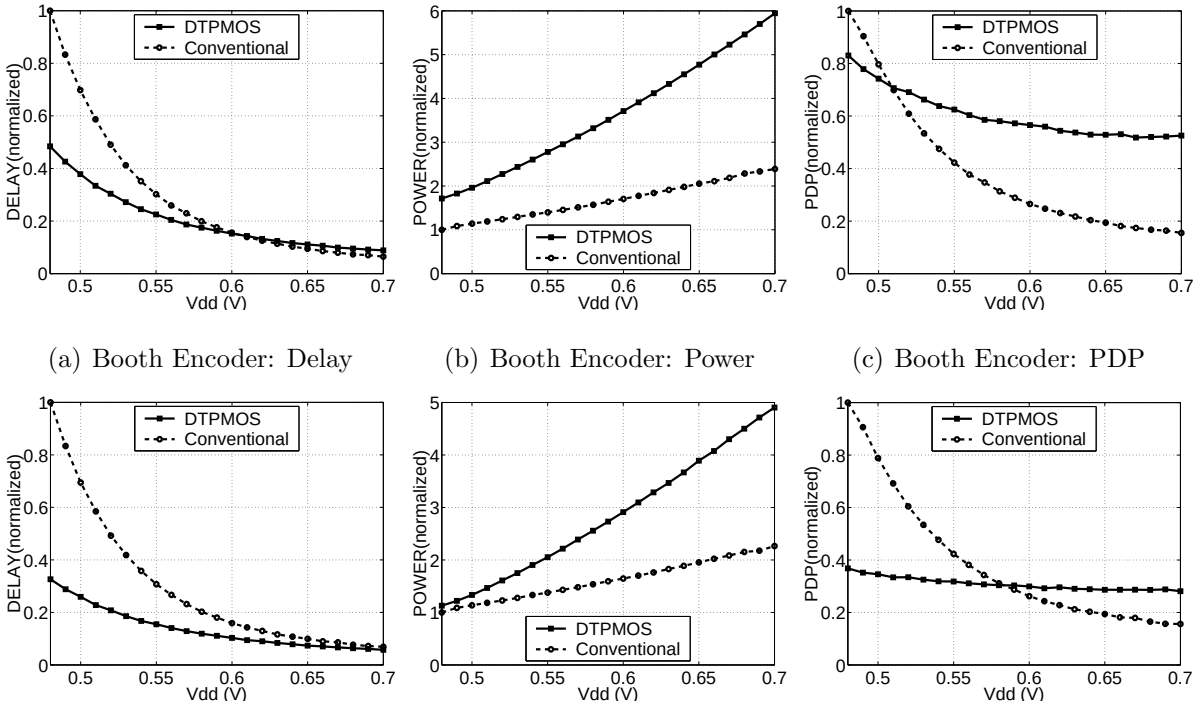
is reversed when supply voltage is lowered to sub-0.5 V. At 0.48 V, the small difference in power dissipation of both schemes in addition to a large delay reduction of DTPMOS leads to reducing PDP of DTPMOS to approximately half that of the conventional CMOS implementation. Therefore, the DTPMOS scheme is attractive for sub-0.5 V operation.



(a) Delay (b) Power (c) PDP

Figure 2.9: Simulation results for the DTPMOS and Conventional CMOS implementations of the FA circuit at different supply voltages.

In addition to the FA, the DTPMOS scheme is utilized in two of the main multiplier building blocks, the Booth encoder and the Booth selector. Simulation results show similar characteristics to that shown in Figure 2.9. At a supply voltage of 0.48 V, DTPMOS results in reducing delay by 50% and 65% compared to conventional CMOS in the Booth encoder and the Booth selector respectively. This is shown in Figure 2.10 (a) and (d) respectively. As supply voltage increases, delay enhancement due to using DTPMOS is reduced. Using DTPMOS in the implementation of the Booth encoder circuit results in approximately 80% increase in power dissipation at 0.48 V as shown in Figure 2.10 (b). Power dissipation increases by only 10% in the Booth selector circuit when using DTPMOS. However the increase in power dissipation becomes dramatic as supply voltage is increased above 0.5 V.



(a) Booth Encoder: Delay (b) Booth Encoder: Power (c) Booth Encoder: PDP

(d) Booth Selector: Delay (e) Booth Selector: Power (f) Booth Selector: PDP

Figure 2.10: Simulation results for the different implementations of the Booth encoder and the Booth selector circuits at different supply voltages.

For example, power dissipation of DTPMOS is approximately 2.5× that of the conventional implementation as can be seen from Figure 2.10 (b) and (e). For Booth selector, PDP enhancement if approximately 60 % at 0.48V. Due to the larger power dissipation of the DTPMOS Booth encoder, PDP enhancement of DTPMOS is only 18% at 0.48V compared to conventional CMOS.

The increased power dissipation of the Booth encoder is primarily due to the increased complexity and the increased number of DTPMOS transistors. For low data activity applications, the turned OFF DTPMOS transistors virtually have the same leakage current

as conventional CMOS. However, the turned ON DTPMOS transistors suffer from higher active state leakage power. The lower the data activity and the higher the number of turned ON DTPMOS transistors the higher the active leakage power. This increase in active leakage power results in a reduction the PDP advantage of DTPMOS. In the next section, a static and a dynamic techniques for reducing active leakage power are described.

2.5.3 Active Leakage Power Management Techniques

The main drawback of connecting the gate to the well in the DTPMOS scheme is the resulting increase in the active leakage current. The source/drain-body junction becomes forward biased when the supply voltage is increased above the diode cut-in voltage[46]. Above the 0.5 V supply, this leakage current increases exponentially.

Two approaches for active leakage power reduction are proposed. The static approach utilizes a single cut-off device to turn OFF all transistors when computation is done. The dynamic approach divides the computational blocks into several stages. Each stage is enabled through its own cut-off device. The enable signals of these devices are sequentially turned ON then OFF to allow each individual stage to finish computation and immediately turns OFF. The details of both techniques are described below.

The static active power management technique was proposed by Kawaguchi *et.al.* [27]. In this scheme a reduction in the standby current is achieved by adding a low-V_{TH} PMOS transistor to power the DTPMOS circuit down during standby mode. In this work, the DTPMOS serves as the low-V_{TH} cut-off device. The cut-off DTPMOS with its high ON current drive is advantageous since it can be implemented in normal bulk CMOS without the need for a multi-threshold technology. The OFF state leakage current of DTPMOS transistors is virtually the same at conventional PMOS transistors. This approach is illustrated in Figure 2.11.

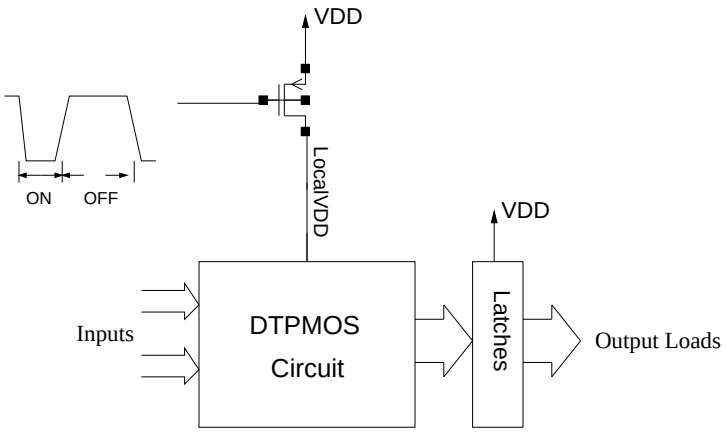
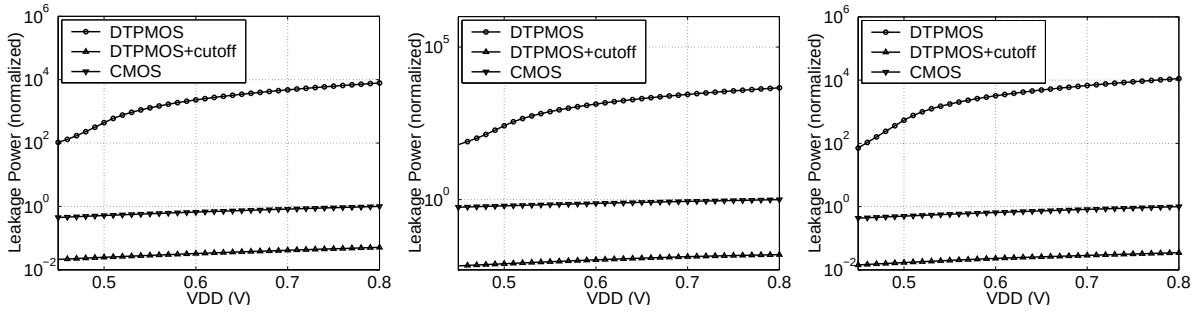


Figure 2.11: Static active leakage power management scheme.

When $LocalVDD$ is powered down, the output signals have to be stored until the supply voltage is powered up. A simple latch in the form of two cross-coupled inverters is used to store the output value of each signal. Those latches are never shut OFF by directly connecting them to the main supply voltage, V_{DD}, as shown in Figure 2.11. Since the latches are always powered up, conventional CMOS transistors are used in the latch's structure to minimize standby current.

Simulation results of active leakage power dissipation of the static approach is shown in Figure 2.12. The technique is applied to the full adder, the Booth encoder, and the Booth selector. Figure 2.12 (a) indicates that leakage of the DTPMOS technique is 2 orders of magnitude compared to the conventional CMOS. As mentioned earlier, this is due to the active leakage current flowing through the forward biased soruce/drain-junctions. When using a cut-off DTPMOS device, the active leakage is reduced by more than one order of magnitude than that of conventional CMOS. A similar reduction in active leakage power is achieved in the Booth encoder and the Booth selector as shown in Figure 2.12 (a) and (b) respectively.



(a) Full Adder (b) Booth Encoder (c) Booth Selector

Figure 2.12: Simulation results for the static active leakage reduction technique when applied to (a) Full Adder, (b) Booth Encoder, and (c) Booth Selector.

The second approach is a dynamic active leakage power management technique. In this scheme, the circuit is divided into consecutive stages. Each stage is controlled individually through a DTPMOS cut-off transistor. Figure 2.13 (a) shows the control unit which generates three control signals, S1-EN, S2-EN and S3-EN. Each control signal is connected to the gate a DTPMOS transistor to power each stage up or down at the appropriate time. The overlap between the control signals shown in Figure 2.13 (b) is to sustain the output levels of one stage till the next stage is powered up and starts processing. Each output from the final stage, Stage 3, is connected to a latch which is always powered up. Simulation results of the dynamic power management technique and a comparison to other schemes are presented below.

2.5.4 Simulation Results and Comparison

Four different versions of 16x16-bit multiplier are implemented using DTPMOS, conventional CMOS, DTPMOS with static active leakage power management, and DTPMOS

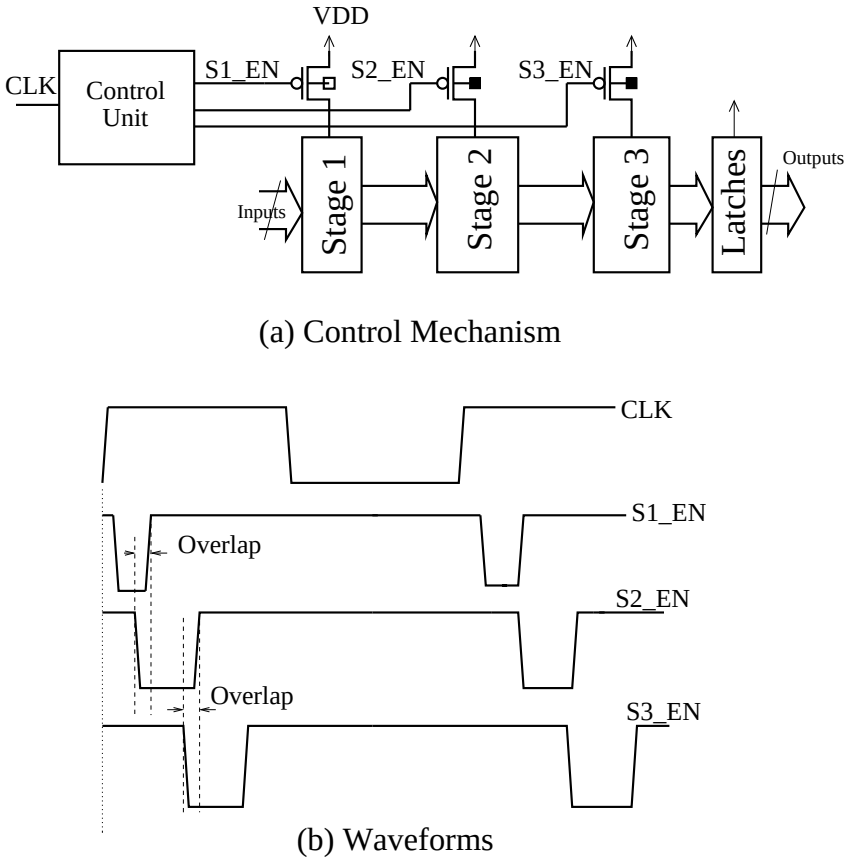


Figure 2.13: Dynamic active leakage power management scheme.

with dynamic power management technique. Simulation results at 2 MHz are indicated in Table 2.3.

At 0.48 V, the conventional CMOS multiplier has failed to work at 2 MHz input frequency. Utilizing the DTPMOS scheme, the multiplier has approximately double the speed of conventional CMOS. Unlike the conventional multiplier which fails to finish computation during one full clock cycle, the DTPMOS multiplier can be shut down after the computation is done. In the static power management scheme, the cut-off transistor is turned OFF

Table 2.3: Simulation results for the 16x16-bit multiplier architectures at 2MHz and 0.48V.

Structure	Delay (ns)	Power (μW)	Energy (pJ)
DTPMOS	220	7.8	1.73
Static power management	250	7.26	1.82
Dynamic power management	300	5.62	1.68
CMOS	fails	-	-

Table 2.4: 16x16-bit Multiplier Architectures Comparison

Source	Tech	V_{dd}(V)	Delay(ns)	Power(W)	Energy (pJ)
This design	0.18 μm (CMOS)	0.48	300	5.62 μ @ 2 MHz	1.68
Fuse *et.al.* [40]	0.4 μm (SOI)	0.5	18	4 m	70
Law *et.al.* [39]	0.8 μm (BiCMOS)	3.3	10.4	38 m @ 10 MHz	395
Shetti *et.al.* [38]	0.6 μm (CMOS)	2.5	Conventional: 6.07 Leapfrog: 3.06	2.27 m 1.48 m	13.8 4.5

after 250 ns (50% of the clock period). The size of the cut-off transistor is optimized to minimize energy consumption. Simulation results shown in Table 2.3 indicate that power is decreased by 7% with a 12% increase in delay and a 5% increase in energy compared to the DTPMOS scheme. The increase in delay and energy is mainly due to switching the capacitance of the large cut-off transistor. Using the dynamic power management scheme results in a 23% and 3% reduction in power and energy respectively with a 27% increase in delay.

A comparison between the proposed design with dynamic power management and some of the other 16x16-bit multiplier designs reported in the literature is shown in Table 2.4. It is evident that the proposed design has the lowest energy consumption amongst the other designs. Utilizing the the proposed technique results in a significant energy reduction

compared to the design reported in [38]. However, careful technology scaling is required to make a fair comparison across the different technology generations indicated in Table 2.4.

It is important to note that the well to gate capacitance increases the input capacitance of the DTPMOS technique. The well capacitance is significant and would affect the overall delay and power of the DTPMOS scheme. Such capacitance was not taken into account due to the lack physical process information regarding the area of the well and well separation and models describing such capacitance. Such data would have affected the results of the DTPMOS scheme and should be carefully considered early in the design stage.

2.6 Summary

Designing for power and energy efficient designs has become a necessity for modern VLSI technologies. With doubling integration capacity every two to three years, power dissipation presents a real threat for reliability and even functionality of the devices. As a result, tremendous effort has been devoted to achieve lower power dissipation without affecting performance. The main components of power dissipation are switching power and leakage power. Switching power, being the dominant power component, has caught special attention in recent years. Many techniques have been introduced to control this ever increasing power component on all levels of design abstraction. Increased leakage current due to technology feature downsizing is another challenge that faces circuit designers in the deep sub-micron era. System, circuit, and device levels are all examined for potential solutions for overall power reduction.

A dynamic threshold PMOS (DTPMOS) scheme has been presented. By connecting the gate and the well of the PMOS transistor, the DTPMOS demonstrates a low threshold voltage in the on-state and a high threshold voltage in the off-state. The new scheme

allows for sub-0.5 V operation in bulk CMOS technologies with a significant improvement in performance and a reasonable reduction in energy compared to conventional CMOS. A 16x16-bit multiplier was designed utilizing the DTPMOS scheme in the 0.18 μm bulk CMOS technology. Simulation results show that the energy consumed by the multiplier is only 1.68 pJ at 0.48 V and a frequency of 2 MHz. The DTPMOS scheme is mostly suitable for sub-0.5 V operation. Above the 0.5 V, the efficiency of the DTPMOS scheme is reduced due to the increase in static power dissipation. The well capacitance adds a significant input loading to the DTPMOS scheme and should be considered carefully early in the design stage to accurately assess the DTPMOS advantages/disadvantages.

Chapter 3

Analysis and Design of Energy Efficient Dynamic Circuits

3.1 Introduction

Advances in dynamic circuits are driven by the need to meet high performance targets in modern VLSI designs. Compared to static CMOS logic, dynamic logic leads to up to 30% performance gain [47]. Speed critical paths often deploy dynamic logic to meet speed requirements. Performance gain over static logic becomes even larger as the number of inputs to the logic grow. Wide fan-in dynamic logic such as domino are often used in performance critical paths, e.g. fast lookahead adders and RAM decoders, to achieve high speeds where static CMOS fails to meet performance objectives.

As the VLSI industry is steering towards more integration, supply voltage has to be reduced in order to keep a constant electric field inside the device. With constant field scaling, maximum device performance for each technology generation can be achieved while maintaining adequate device reliability [48]. However, the resultant degradation in

performance due to supply voltage scaling often forces designers to reduce threshold voltage of the device to meet performance goals. An exponential increase in subthreshold leakage current is a direct consequence of threshold voltage reduction. Subthreshold leakage power is expected to increase by a factor of 5× each technology generation [47]. Furthermore, gate leakage is expected to increase for future technology generations due to thinner gate oxide and scaled geometries. This mounting leakage current severely degrades noise immunity for DSM VLSIs.

Dynamic circuits are more susceptible to noise compared to static CMOS. Unlike static logic, the dynamic node of dynamic logic is not always driven. This problem is further compounded by increased fan-in and elevated temperature resulting in increased leakage current and potential false evaluation. Crosstalk, charge sharing, and ground bounce also can alter the behavior of dynamic logic [49]. Therefore, it is often a challenge to maintain stability of dynamic logic while achieving the performance target. This challenge is quite evident in the design of dynamic gates. The high performance advantage of dynamic logic is often traded off for improved noise immunity and leakage tolerance.

In addition to noise immunity, power dissipation of dynamic logic has limited the utilization of dynamic logic in low power applications. Switching of the Clock every cycle irrespective of the logical result and the corresponding power dissipated in the clock network leaves dynamic power at levels far above those that low power applications can afford. With the speed advantage of dynamic logic over slower logic families, e.g. static CMOS, supply voltage can be reduced while meeting the target performance. This allows for energy savings and help reduce clock power quadratically. However, noise immunity can be negatively impacted by a reduction in supply voltage. As a result, an undesirable false evaluation can occur. Therefore, a great deal of time and effort is spent on designing dynamic logic in order to meet performance, noise tolerance, and power targets.

3.2 Leakage Tolerant Wide Domino Logic

Dual-threshold (DVT) dynamic logic implementations have been introduced to address the trade-off between performance and stability. Low-threshold (LVT) transistors are deployed in speed critical paths while high-threshold (HVT) transistors are used elsewhere to keep leakage current within limits. Most of DVT implementations have been applied to static CMOS circuits [25] [22] to minimize OFF state leakage while maintaining the required performance. Recently, Kao and Chandrakasan [23] proposed a DVT technique for domino logic. Figure 3.1 shows a conventional DVT n-input wide domino gate [23]. In this configuration, the pulldown NMOS evaluation transistors are all LVT for high performance operation, while all other transistors are HVT to minimize OFF state leakage current.

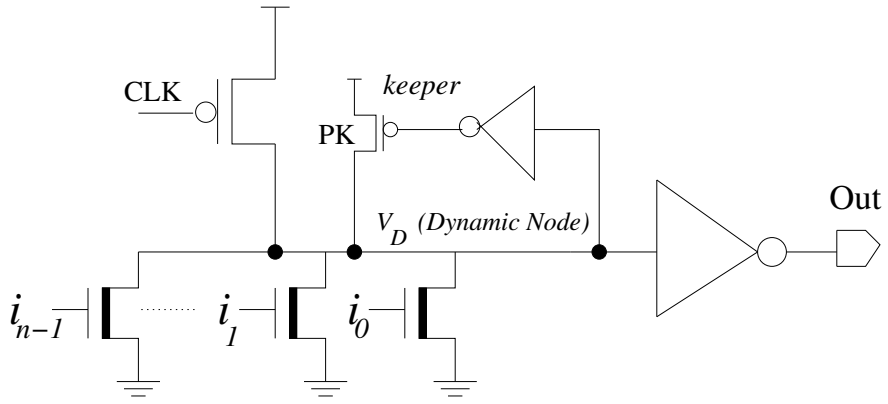


Figure 3.1: Conventional DVT Wide Fan-In Domino n-input OR gate.

A DVT domino can be realized with or without a footer transistor. The footer transistor (connected between sources of the pulldown transistors and ground) is often avoided to maximize performance. As shown in Figure 3.1, removing the footer transistor implicitly restricts the CLK signal to arrive before data to avoid DC conduction when both the precharge and pulldown devices are conducting. A footed domino has a 10% better noise

immunity while resulting in a 30% performance loss compared to a footless domino [50]. However, a footled domino has a significantly lower leakage current compared to a footless domino due to the stacking effect of two or more series transistors [51].

In order to compensate for the OFF state charge loss, a keeper transistor is utilized. In conventional domino circuits, the keeper contends with evaluation transistor(s) since the keeper is already ON at the onset of evaluation. Therefore, upsizing the keeper in order to compensate for charge leakage results in a performance degradation. Moreover, increased contention due to the upsized keeper can result in a false evaluation when a single pulldown transistor fails to discharge the dynamic node. However, as leakage currents are increasing with technology scaling keeper upsizing is becoming a necessary requirement. Some experts speculate that the conventional domino logic may become nonfunctional when the keeper becomes large enough in the 70 nm technology generation [52]. Therefore, significant attention has been given to the design of leakage tolerant domino circuit techniques [53] [54] [55] [56] [57].

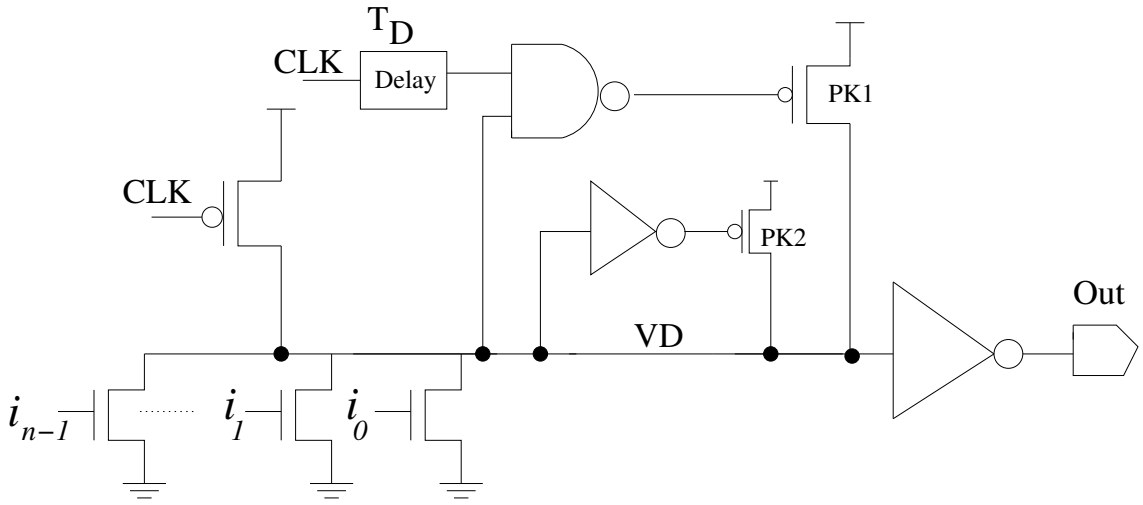


Figure 3.2: Conditional Keeper technique.

The increasing contention between the keeper and the evaluation transistors as leakage increases with technology scaling has spurred research to mitigate the effect of increased keeper size on performance without compromising robustness. Recently, a conditional keeper (CKP) technique, shown in Figure 3.2, for noise tolerant wide fan in gates was proposed in [58] [53]. In this technique, the keeper device (PK) in conventional domino (see Figure 3.1) is divided into two smaller ones, $PK1$ and $PK2$. The keeper sizes are chosen such that $PK = PK1 + PK2$. Such sizing guarantees the same level of leakage tolerance as the conventional gate but yet allows for faster evaluation. A typical ratio for $PK1/PK2$ is 9/1. The large keeper ($PK1$) in Figure 3.2 is deployed after a certain delay T_D, to prevent erroneous discharge of the dynamic node (V_D) when all inputs remain LOW. The small keeper ($PK2$), however, remains ON to compensate for charge leakage until $PK1$ is activated. Deploying a larger portion of the keeper device after the delay T_D, depending upon the condition of the dynamic node, reduces contention power and hence enhances performance. The timing after which the large keeper $PK2$ is enabled is critical in trading off speedup and noise immunity. A detailed timing analysis of the conditional keepers is given in [59].

A delayed keeper technique with gate biasing was proposed in [60]. A delayed clock is used to disable the keeper for a period of time in which most of the contention occurs. The gate of the keeper is controlled to provide a weak keeper at the start of evaluation and a full keeper when the dynamic node does not evaluate.

Kursun *et.al.* proposed a conditional keeper technique through back biasing [61] of the keeper transistor. In this scheme, the well of the keeper is biased at a voltage higher than the normal supply voltage at the beginning of the evaluation phase. The source to body junction which has a higher voltage than the supply results in a higher threshold voltage and less current drive for the keeper. Contention at the beginning of the evaluation phase

is, therefore, reduced. After a certain delay the body voltage of the keeper is restored to the normal supply (source voltage) and the strength of the keeper is restored.

In this chapter, a novel DVT circuit technique to mitigate the impact of the increased subthreshold leakage current in wide OR gates is presented in section 3.3. The new circuit technique mitigates the impact of leakage current in wide domino circuits by splitting the number of evaluation devices into two sections. Such splitting results in a smaller dynamic node capacitance and consequently a faster evaluation. Reducing the number of transistors in each pulldown network also allows for the use of smaller keeper devices and hence a reduction in contention power.

Furthermore, the speed and power advantage of the proposed technique is enhanced as supply voltage is scaled down for low power applications. Low voltage operation of dynamic circuits inherently poses potential energy gains compared to static CMOS through fast evaluation followed by clock gating. However, noise immunity becomes an issue when less charge is stored on the dynamic node as supply voltage is reduced. The speed and power enhancement of the proposed circuit technique is a result of the reduction in diffusion capacitance and contention current.

Design of DVT wide OR gates is optimized through the development of an accurate delay model for conventional DVT wide domino logic. The objective is to analyze the stability of DVT domino logic when subjected to DC-noise. This model allows us to investigate various design and technology trade-offs in order to achieve performance, leakage, and stability objectives. Performance is examined as V_{TH} is reduced and the fan-in number is increased while maintaining the same level of robustness. The effect of keeper upsizing to maintain leakage within bounds is also considered. Section 3.6 describes the basic MOSFET model used. The delay model for conventional DVT domino is described section 3.7. First, the optimal keeper size which accommodates for worst case leakage is obtained. Subsequently,

implications of keeper upsizing on delay are estimated. In addition to threshold voltage, fan-in size of the gate is also considered in the analysis as a design parameter. The model is then extended to complex DVT domino implementations in section 3.8. A comparison between estimated delay using the proposed model and HSPICE simulations is presented in section 3.9. The model allows us to examine the impact of threshold voltage reduction on stability and performance for different circuit techniques and for a given fan-in.

3.3 Split Domino (SD) Circuit Technique

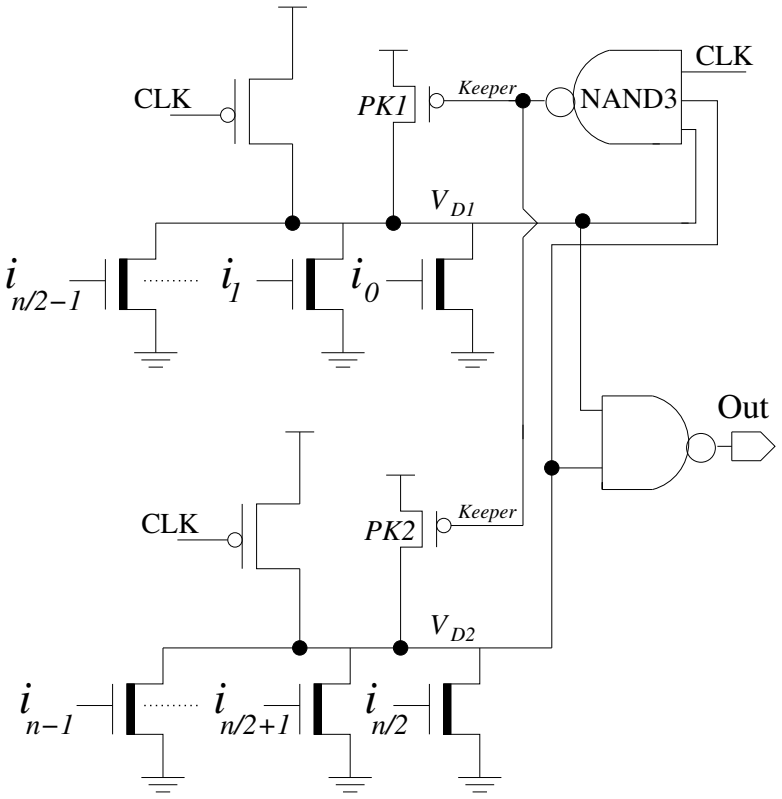


Figure 3.3: n-input split domino (SD) OR gate.

The SD gate shown in Figure 3.3 achieves higher performance of operation through splitting the pulldown devices into two networks. A logical 2-input NAND operation is then utilized to generate the output. The output inverter shown in Figures 3.1 and 3.2 is no longer required for the SD circuit. Also, the keeper device is split equally between the two networks. The main advantage of splitting the pull-down network into two sections is to reduce the dynamic node capacitance and consequently faster evaluation. Also, the large keeper transistor in the conventional case is replaced by another transistor which is nearly half the original keeper size leading to less contention.

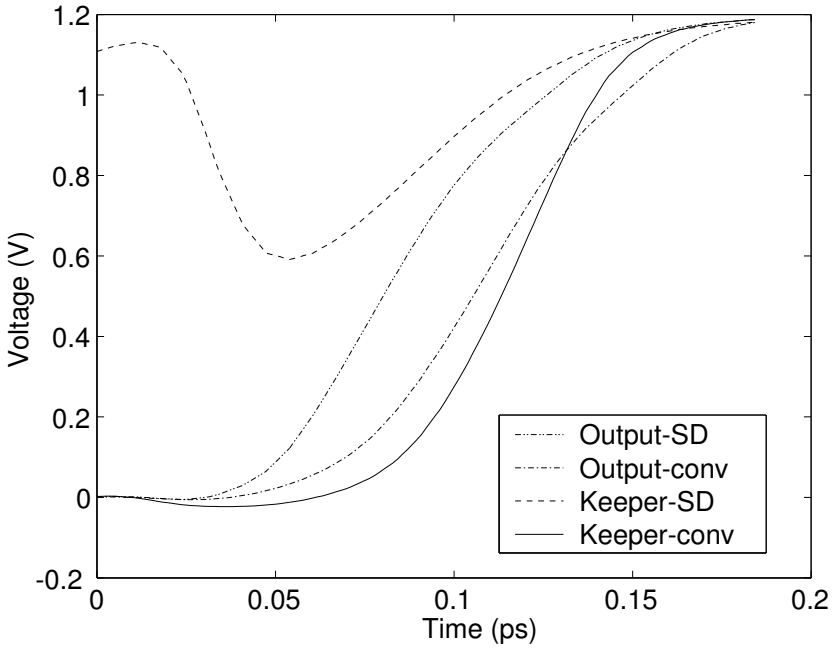


Figure 3.4: Keeper and output waveforms for SD and conventional 32-input OR gate

The operation of the SD circuit is described as follows. During precharge, CLK is LOW, the keeper devices are OFF and the output is LOW. At the onset of evaluation,

contention is eliminated since keeper devices remain OFF. There are two different cases that need to be considered during the evaluation phase. When all inputs remain LOW and leakage current is at its maximum, the keeper devices controlled by the 3-input NAND gate are quickly activated to prevent the dynamic node from drooping and to keep output noise within the required limit. The 3-input NAND gate is skewed in such a way to allow for a very fast discharge of the keeper control signal in case all inputs remain LOW. The other case is when the gate evaluates, where at least one input turns HIGH. In this case, the dynamic node discharges very quickly due to decreased capacitance and nearly keeps the keeper devices in the OFF state and contention is therefore minimized. Figure 3.4 shows the different waveforms for both SD and conventional techniques. Clearly, the SD keeper is OFF at the onset of evaluation while the conventional keeper is ON. The keeper control signal can be seen to droop and quickly recovers to V_{DD}q, as shown in Figure 3.4, maintaining keeper devices virtually OFF.

The design overhead of the SD gate is represented by the power dissipation of the 2-input and 3-input NAND gates in addition to more CLOCK power due to the extra loading by the 3-input NAND gate. This overhead can be fairly justified, as shown below, by the resultant performance improvement making the SD circuit technique a good candidate for low energy applications.

As the number of inputs grows, the number of splits can be increased to gain further speed up. The limitation on the number of splits is speed degradation resulting from the output and the feedback NAND gates. The speedup results from using n splits can be absorbed by the speed degradation resulting from using n-input output NAND gate and $n + 1$-input NAND gate for the feedback. Such trade off needs to be considered when deciding the optimal number of splits.

3.4 Simulation Results and Comparison of the SD Circuit Technique

The SD circuit technique is compared to both conventional domino as a reference design and to the CKP technique. The comparison is based upon simulations of the three techniques for 16 and 32-bit OR gates implemented in 0.13μm dual threshold technology. Low threshold devices are used for the NMOS evaluation network. High threshold devices are used otherwise (e.g. keepers, inverters, etc.). The output load is set to a fan out of 4. The clock frequency is kept at 2 GHz. Keeper sizing is performed at worst case leakage condition, i.e. at temperature of 110^{o}C, V_{DD} of 1.2 V, and the Fast-Slow process corner. The keeper transistors are sized such that the noise level at the output node does not exceed that applied at the inputs. The *Unity-Gain DC Noise* (UGDN) as defined in [50] is used as the leakage tolerance criterion. As the input DC noise level increases, leakage current increases exponentially. The DC input noise is limited to 12% of V_{DD} such that the input noise is always below the low threshold voltage for the pulldown devices. This level of noise is based on the assumption that noise from a previous stage is approximately 12% of V_{DD}. Crosstalk and ground bounce are neglected in this analysis Crosstalk is ignored based on the assumption that all input signal wires are properly shielded. Power grid is assumed to supply enough current with enough decoupling capacitors to minimize ground bounce. After keeper sizing is performed, the performance metrics (delay, power, and power-delay-product (PDP)) are measured at typical process corner, normal operating temperature (27^{o}C), and zero DC input noise.

Figure 3.5 shows the delay enhancement of the SD technique compared to CKP technique both normalized to the delay of the corresponding conventional gates. Delays are normalized to the delay of the corresponding delay of a conventional domino gate. The

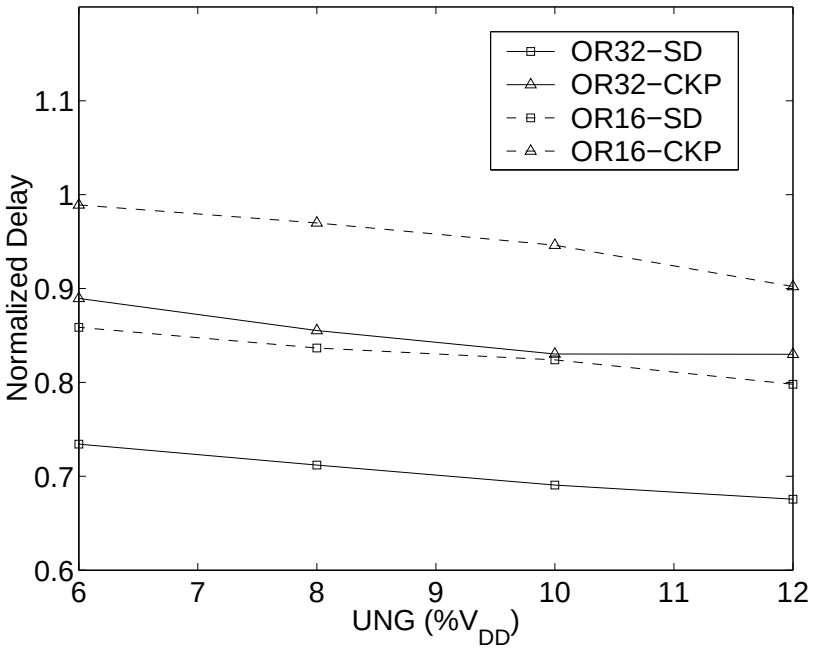


Figure 3.5: Delay of the SD and the CKP gates relative to the conventional technique.

delay metric is defined here as the Data-to-Output delay since data should arrive after the CLOCK in footless structures as mentioned in Section 3.2. As the DC input noise level is increased, the performance improvement of the SD gate becomes evident. For the OR16 and OR32 and at a noise level of 12% UGDN, the SD gate offers 20% and 32% performance improvement respectively over the corresponding conventional gate delay. The delay reduction is 7% and 12% compared to the CKP technique for the OR16 and OR32 respectively.

Power dissipation of SD gates compared to that of CKP technique is shown in Figure 3.6. The results are normalized to power dissipation of the corresponding conventional gates. Power dissipation of SD gates are 6 to 8 % higher than that of the corresponding conventional gates due to the overhead of the two NAND gates used in the SD logic.

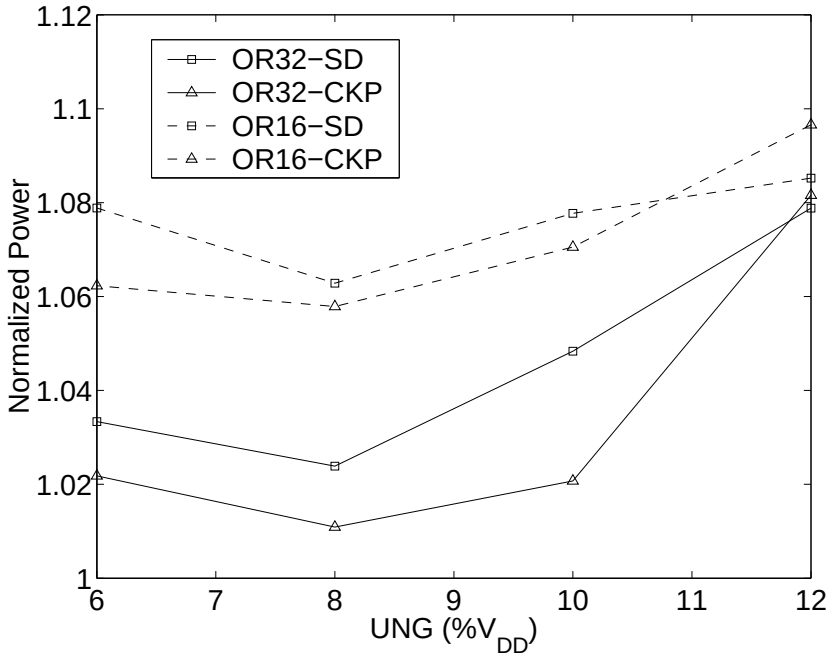


Figure 3.6: Power dissipation of the SD and the CKP gates relative to the conventional technique.

Since keeper devices and the 3-input NAND gate have to be upsized to sustain higher leakage, power dissipation reaches a maximum at 12% UGDN compared to conventional logic. Power dissipation of the CKP gates is slightly lower than that of the SD gates at low noise levels. However, both techniques tend to dissipate the same amount of power at high noise levels. The reason is that the small keeper in the CKP technique has to be upsized to maintain leakage with limits at the onset of evaluation. Therefore, contention power increases and the total power also increases to reach the level dissipated by the SD technique.

Figure 3.7 shows that the proposed technique is more energy (PDP) efficient compared to the other wide domino techniques. The SD technique can achieve up to 28% PDP reduction for the OR32 case compared to the conventional technique at 12% UGDN .

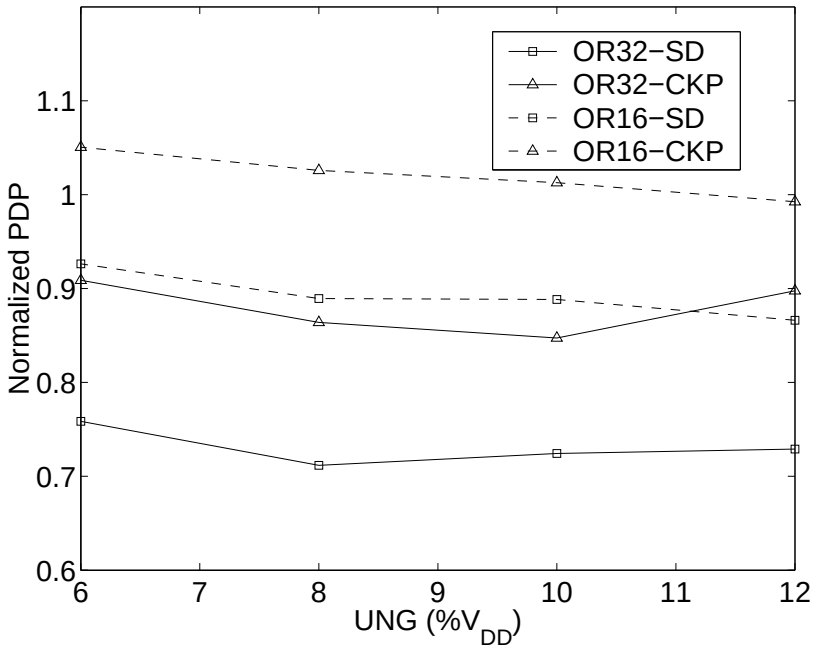


Figure 3.7: Energy(PDP) of the SD and the CKP gates relative to the conventional technique.

At the same noise level, PDP saving is 12% for the OR16 case. The PDP savings of SD technique are 14% and 17% for OR16 and OR32 respectively compared to the CKP technique. A comparison between delay, power, and PDP at 12% UGDN for OR16 and OR32 using the three different techniques is summarized in Table 3.1.

Dynamic power in the clock network of dynamic circuits can be reduced when supply voltage is reduced. However, careful examination of the impact of supply voltage reduction on noise immunity of dynamic circuits is necessary. Low voltage operation of both the conventional and SD domino logic is examined in the following section.

Table 3.1: Simulation results at 12% UGDN

Gate	Technique	Delay	Power	PDP
OR16	Conventional	1	1	1
	CKP	0.9	1.1	1
	SD	0.8	1.08	**0.86**
OR32	Conventional	1	1	1
	CKP	0.84	1.08	0.9
	SD	0.68	1.08	**0.73**

3.5 Low Voltage Operation of Wide Fan-In Domino Circuits

Dynamic circuits are faster than their static counterparts due to the reduced overall capacitance (the PMOS capacitance). However, dynamic circuits are inherently more power hungry due to the large power required by the clock distribution network. Reducing supply voltage of dynamic circuits leads to a reduction in the power dissipated by both the clock tree and the dynamic circuit itself. Given a positive slack time due to the fast switching nature of dynamic circuits, voltage can be optimally reduced till the point where timing requirements are met. However, noise immunity at reduced supply voltages becomes an issue for dynamic circuits. As supply voltage is reduced, the amount of charge stored on the dynamic node is reduced linearly. An undesirable charge loss during the evaluation phase at low voltage may have a higher probability of occurrence compared to that at regular voltage operation. A closer look at noise immunity of dynamic circuits at low supply voltage is required.

As mentioned earlier, keeper sizing is performed at worst case noise. In this analysis, worst case noise is assumed to be 30% of the supply voltage, V_{DD} (nominally is 1.2 V) in a 90 nm CMOS process. This level of noise reflects the worst case resulting from all sources of noise including a 15% due to crosstalk, 5% due to ground bounce, and 10% from the preceding stage. This is based in the assumption that ground bounce and crosstalk have almost doubled by going from the 130nm to the 90nm feature size. With worst case noise level scaling with V_{DD}, noise immunity of dynamic circuits is not greatly affected when voltage is reduced.

To validate the above statement, the drains of two parallel connected NMOS devices, V_D, are precharged to V_{DD} and a DC noise V_N of 30% of V_{DD} is applied to the gates of the two transistors as shown in Figure 3.8. Worst case leakage condition of FS, 125^{o}C is considered. This scenario mimics the dynamic node of a footless domino gate with the keeper device is disconnected. Therefore, any external factors are excluded from affecting the response of the dynamic node to the applied noise. Two cases for V_{DD}, 0.8V and 1.4V, are considered. Figure 3.8 shows that the discharge time of the drain connection for the two cases is approximately the same although the DC noise level is 75% higher for 1.4V supply compared to the 1.2V case. The discharge rate when supply voltage is 0.8V is slower compared to that at 1.4V. When the drain-to-source voltage, initially equals to V_{DD} in this case, is reduced from 1.4V to 0.8V, the subthreshold leakage current is reduced exponentially. For the real domino circuit, a keeper is connected to the node V_D. This keeper transistor operates in the linear mode. Therefore, the keeper current is reduced approximately linearly with drain-to-source voltage. As a result, the weak keeper at low supply voltages can compensate for the lower leakage level and maintain the required level of stability originally obtained at higher supply voltages. Therefore, noise immunity remains almost the same as supply voltage is scaled.

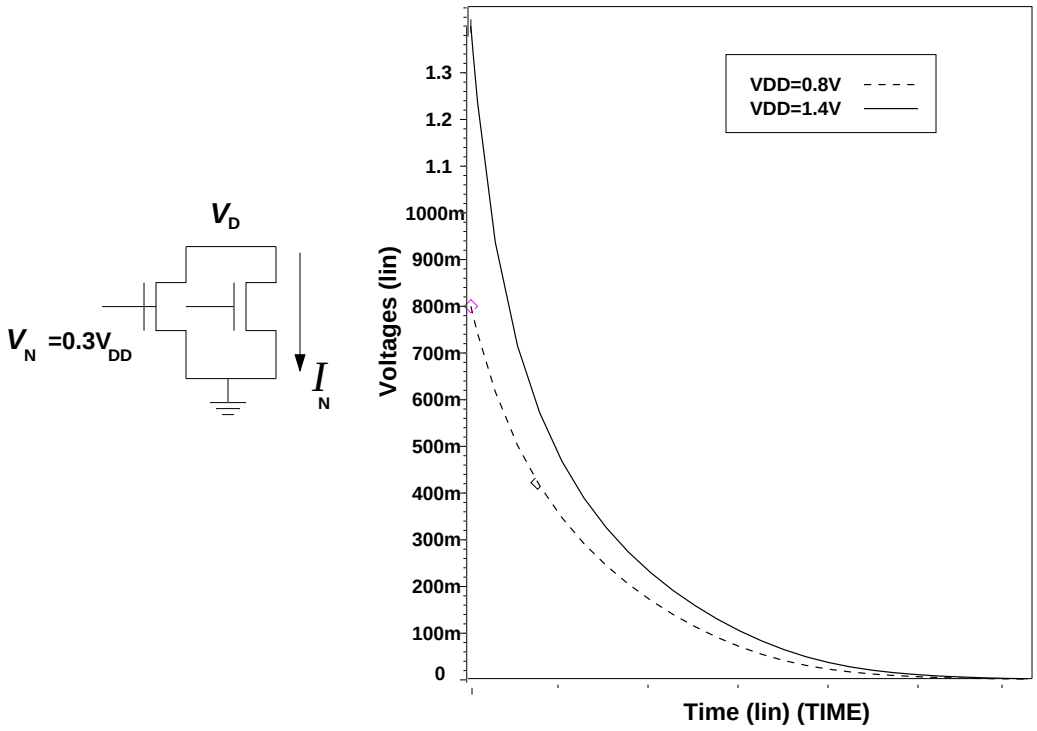


Figure 3.8: Discharge of the drain node of two parallel connected NMOS transistors at supply voltage of 0.8V and 1.4V when Drain is precharged to V_{DD} and Gate is subjected to 0.3V_{DD}.

The key to low voltage operation of dynamic circuits is to maintain the speed advantage and stability at low supply voltages. This can be accomplished through minimizing the diffusion capacitance and reducing contention current. Therefore, the SD structure is well suited for low voltage operation. The reduced diffusion capacitance of SD domino leads to a good delay scalability with voltage compared to conventional domino. Furthermore, the lesser contention current at the beginning of the evaluation phase leads to a faster operation. Therefore, SD domino has more power savings as supply voltage is reduced compared to conventional domino.

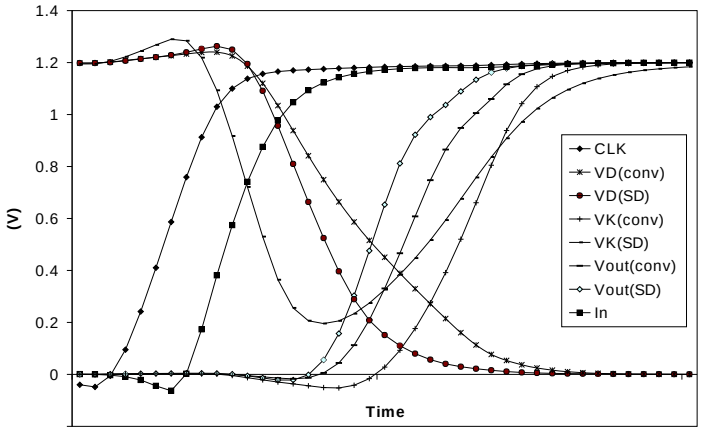
The above argument is verified through simulation of 8, 16, and 32-input conventional

and SD domino gates. Keeper transistor is sized in such a way that noise immunity of both the SD and conventional gates are approximately the same. Worst case delay condition, SF and 125^{o}C, with only one switching input is considered. All circuits are simulated at two different supply voltages, 0.8V and 1.2V. Figure 3.9 (a) and (b) show simulation results of 8-bit SD and conventional domino gates. Due to smaller diffusion capacitance and lesser contention current, the SD gate is 39% faster than conventional domino.

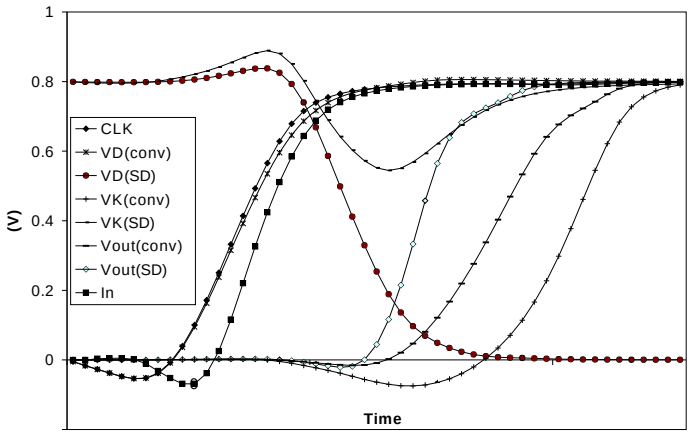
Transient waveforms of SD and conventional 16-bit OR gates at 0.8V and 1.2V supply are shown in Figure 3.10 (a) and (b) respectively. It can be noticed that the speed advantage of SD grows when the number of inputs increases from 8 bits in Figure 3.9 to 16 bits. This is due to the increased diffusion capacitance and leakage current per gate and the corresponding increase in keeper size. As a result, contention current in convectional domino increases with the increased number of pulldown paths. Contention current of SD remains at approximately half the conventional value due to the split nature of the pulldown network.

The delay enhancement, power savings, and power-delay-product (PDP) of the low voltage SD gate compared to conventional 8, 16, and 32-bit gates are shown in Figure 3.11, 3.12, and 3.13 respectively. All plots represent the SD figures normalized to the corresponding conventional domino results. The delay advantage of SD over conventional domino is improved as supply voltage is scaled down. Delay enhancement of SD increases from 10% to 28% when supply is scaled from 1.4V down to 0.8V for the 8-bit OR gate as shown in Figure 3.11. The speed advantage of SD is further improved as the number of inputs grows. Delay enhancement starts at 30% and 78% at 1.4V and reaches 61% and 87% at 0.8V for the 16 and 32-bit gates respectively.

Not only delay of SD compared to conventional domino is enhanced but also power dissipation is reduced as supply voltage is scaled down. Figure 3.12 shows that power dis-



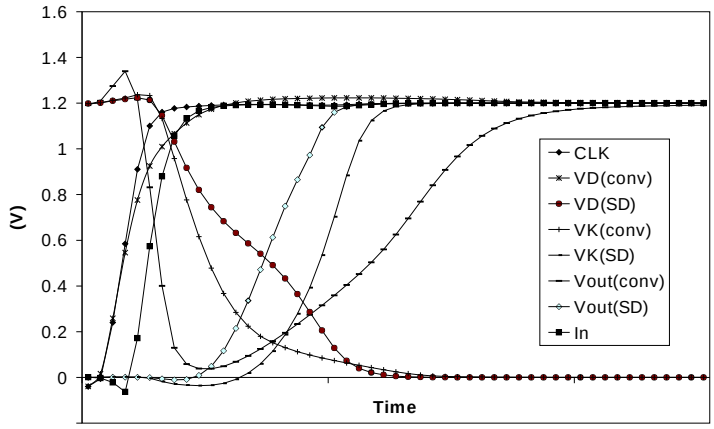
(a) 1.2V



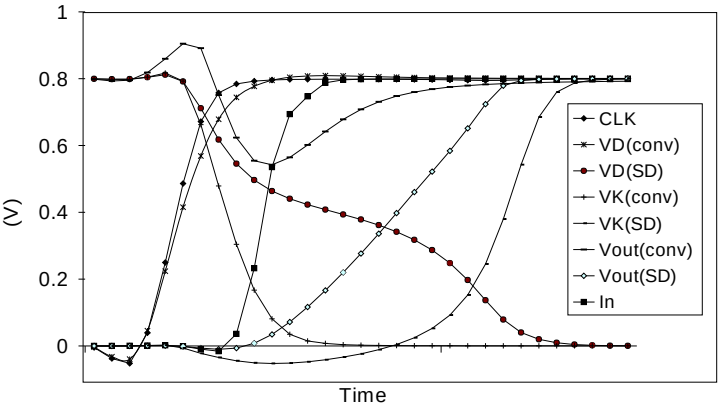
(b) 0.8V

Figure 3.9: Simulation waveforms for 8-bit Conventional and SD at (a) 1.2V and (b) 0.8V supply.

sipation of the 8-bit SD OR gate is slightly higher than the conventional at most operating voltages and it becomes lower at 0.8V. However, power dissipation of SD becomes less than that of conventional domino for the 16 and 32-bit cases. SD is 60% and 75% less in power dissipation at a supply voltage of 0.8V for the 16 and 32-bit gates respectively.



(a) 1.2V



(b) 0.8V

Figure 3.10: Simulation waveforms for 16-bit Conventional and SD at (a) 1.2V and (b) 0.8V supply.

As a result of the speed and power enhancements of SD over conventional domino, energy (PDP) is continuously improving as supply voltage is reduced. Energy reduction of SD over conventional is 1.4×, 3.3×, and 4.1× when supply voltage is scaled from 1.4V to 0.8V for the 8, 16, and 32-bit gates respectively as shown in Figure 3.13.

As mentioned earlier, keeper sizing, input noise, and pulldown transistor size, amongst

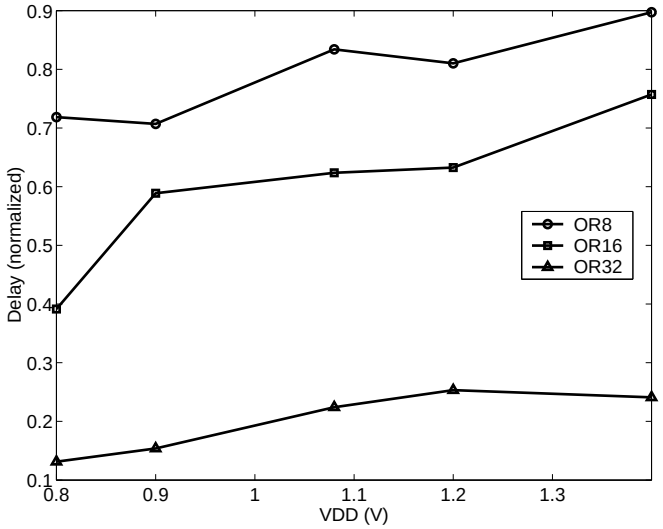


Figure 3.11: Delay plot of SD normalized to Conventional Domino as voltage is scaled down.

other parameters, require careful design and optimization. A methodology for the analysis and optimization of wide OR domino gates is presented in the following sections. First, an accurate leakage and active current model for the device is described in section 3.6. Then, a delay model for conventional and SD logic is presented and used for design optimization.

3.6 MOSFET Device Model for Circuit Analysis

In order to examine the behavior of DVT domino circuits, a simple, yet accurate, device model is utilized. Due to its simplicity and accuracy, the alpha-law power model [62] is used to model drain current in both linear and saturation modes of operation. However, subthreshold current, which is crucial for DC-noise analysis, is not represented in the alpha-law power model. We used the BSIM2 model presented in [63] to model drain current in the subthreshold region. There are several other complex models available to accurately

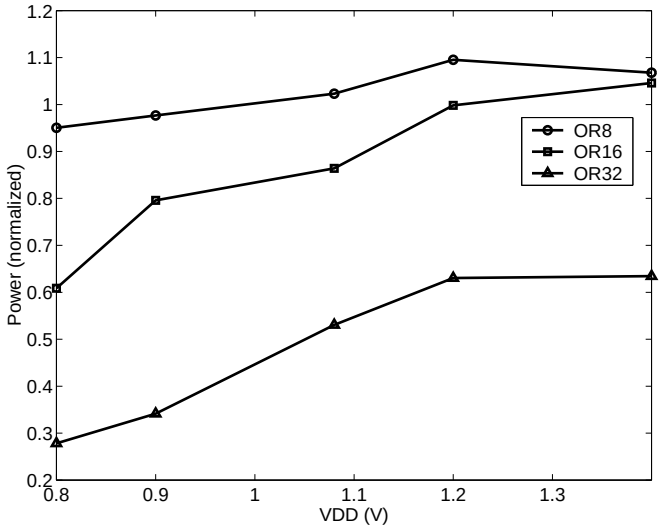


Figure 3.12: Power plot of SD normalized to Conventional Domino as voltage is scaled down.

model device behavior such as [64] and [65]. Such a high accuracy is beyond the scope of this work since the objective is to analyze and to optimize wide domino gates and not to accurately model their transient response. In this analysis, the drain current model of an MOSFET device is given by [66] [67]:

$$I_{\mathrm{D}} = \begin{cases} I_{DSAT} = I_{D0}\left(\frac{V_{GS}-V_{TH0}}{V_{DD}-V_{TH0}}\right)^{\alpha}(1+\lambda V_{DS}) & (\text{saturation}: \ V_{DS} > V_{DSAT}) \\ I_{DSAT}\left(2-\frac{V_{DS}}{V_{DSAT}}\right)\left(\frac{V_{DS}}{V_{DSAT}}\right) & \\ \qquad\qquad (\text{linear}: \ V_{DS} < V_{DSAT}) & \\ I_{\mathrm{sub}} \qquad\qquad (\text{subthreshold}) & \end{cases} \tag{3.1}$$

V_{GS} is the gate-to-source voltage, V_{DS} is the drain-to-source voltage, V_{DD} is the supply voltage, V_{TH} is the threshold voltage of the device, I_{D0} is the drain current at $V_{GS} = V_{DS} = V_{DD}$, α is the velocity saturation index, λ is the channel modulation index, and V_{DSAT} is

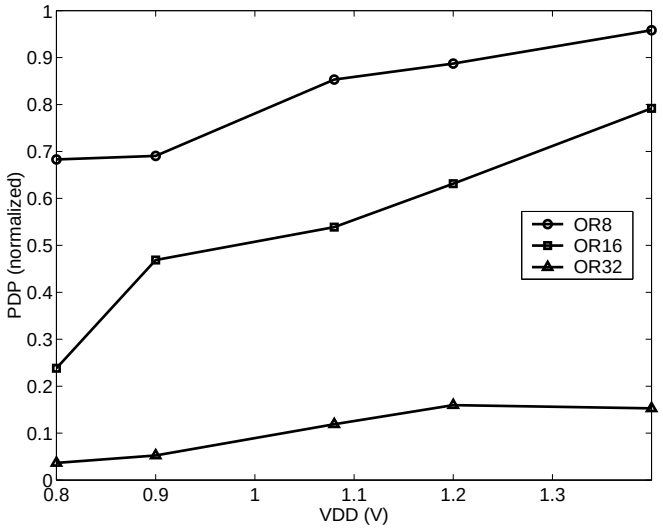


Figure 3.13: PDP (Energy) plot of SD normalized to Conventional Domino as voltage is scaled down.

the saturation voltage.

Threshold voltage, V_{TH}, which is a primary device parameter, is given by

$$V_{TH} = V_{TH0} - \eta V_{DS} + \gamma V_{\mathrm{S}} \tag{3.2}$$

where V_{TH0} is the threshold voltage at zero bias ($V_{DS} = V_{GS} = 0$), η is the Drain-Induced Barrier Lowering (DIBL) coefficient, and γ is the linearized body effect coefficient. All three parameters can be extracted from the I-V characteristics of the transistor.

Modeling subthreshold current, I_{sub} in (3.1), depends on V_{GS} of the device. When a DC-noise is applied to the gate of a certain device at a level below the threshold voltage, the device operates in the weak inversion region. The weak inversion current is given by

$$I_w = I_s \exp[(V_{GS} - V_{TH} - V_{\mathrm{off}})/(nV_T)] \tag{3.3}$$

with

$$I_s = \left(\frac{W_{\mathrm{eff}}}{L_{\mathrm{eff}}}\right) V_T^2 \mu_{\mathrm{eff}} \phi.$$

Where V_T is the thermal voltage (26 mV at 25^oC), V_{off} is the offset voltage which is a fitting parameter, n is the subthreshold swing coefficient, $\phi = \sqrt{q\epsilon_s N_{\text{CH}}/(2\phi_s)}$, ϵ_s is the silicon permittivity, N_{CH} is the channel doping, and $2\phi_s$ is the built-in potential barrier.

As can be seen from (3.3), mobility degradation results in a linear reduction in leakage. However, the weak inversion current increases exponentially with the reduction in threshold voltage. As a consequence, leakage is increased as V_{TH} is decreased. If $V_{\text{S}} = 0$, I_w can be written as [64]:

$$I_w = I_s \exp[(V_{GS} - V_{TH0} + \eta V_{DS} - V_{\text{off}})/(nV_T)]. \tag{3.4}$$

Consequently, I_w in (3.4) can be simplified by expanding the term $\exp(\eta V_{DS})$ using Taylor series expansion. Then, the subthreshold current, I_w, can be approximated by

$$I_w \approx I_{w_0}[1 + bV_{DS} + (bV_{DS})^2] \tag{3.5}$$

where $I_{w_0} = I_s \exp[(V_{GS} - V_{TH0} - V_{\text{off}})/(nV_T)]$ and $b = \eta/(nV_T)$. Comparison between HSPICE simulation and (3.5) yields 2% error and therefore the second order Taylor series expansion is sufficient. The quadratic rather than the exponential dependence of I_w on V_{DS} helps reducing the computational complexity.

On the other hand, when V_{GS} just exceeds the threshold voltage, transistors operate in the moderate inversion region [68]. The moderate inversion region can be considered as the transition region between weak and strong inversion modes of operation. However, the boundaries of such a region are fuzzy. In our model, the moderate inversion region is chosen to be $0.8V_{TH0} < V_{GS} < 1.5V_{TH0}$ based on several simulations that were performed using different transistor sizes, different threshold voltages, and process corners. When $V_{GS} < 0.8V_{TH0}$, the device operates in the weak inversion region while $V_{GS} > 1.5V_{TH0}$ defines the beginning of the strong inversion region. Figure 3.14 shows the weak, moderate and strong regions of operation.

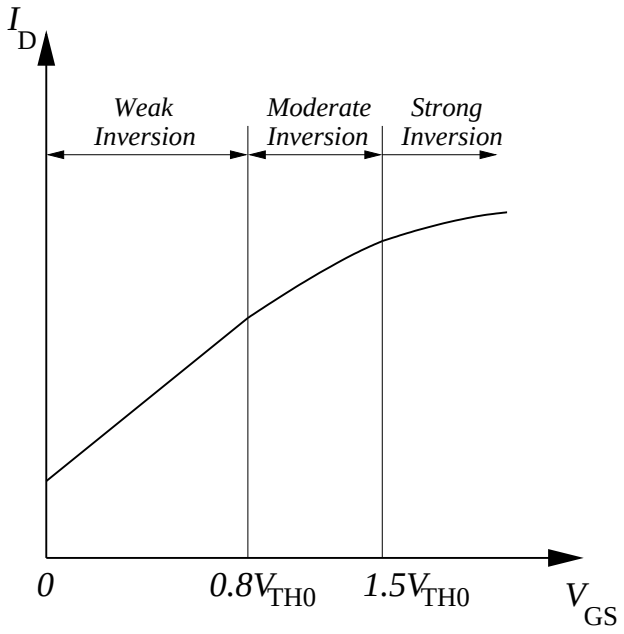


Figure 3.14: Different Inversion Modes for the MOSFET Device.

Enz *et. al.* [69] proposed a simple model for drain current in the moderate inversion region which is given by

$$I_m = I_{m_0}\left[\ln\left(1+\exp\left[\kappa\frac{V_{GS}-V_{TH}}{2V_T}\right]\right)\right]^2 \tag{3.6}$$

where

$$I_{m_0} = \frac{2\mu_{\text{eff}}C_{ox}{V_T}^2}{\kappa}.\frac{W_{\text{eff}}}{L_{\text{eff}}}$$

and

$$\kappa = \left(1+\frac{\beta}{2\sqrt{(1+\theta)\phi_F}}\right)^{-1}$$

where $\beta = \sqrt{2q\epsilon_s N_{\text{CH}}}/C_{ox}$, $\phi_F = V_T ln(N_{\text{CH}}/n_i)$, and n_i is the intrinsic carrier concentration which is temperature dependent. For silicon, the approximated value for n_i at a given temperature can be easily calculated [70]. θ is a parameter set between zero, for

extreme weak inversion (near depletion mode), and one, for the boundary between weak and moderate inversion. A value of $\theta = 0.5$ is used in our model for a closer fitting to HSPICE simualtions. The moderate inversion current, I_m, in (3.6) is independent of V_{DS} and therefore is simple to compute.

From (3.5) and (3.6), the current I_{sub} is given by

$$I_{\text{sub}} = \begin{cases} I_w & V_{GS} < 0.8V_{TH0} \qquad \text{using (3.5)} \\ I_m & 0.8V_{TH0} < V_{GS} < 1.5V_{TH0} \qquad \text{using (3.6).} \end{cases} \tag{3.7}$$

In order to account for the worst case leakage, temperature effect on carrier mobility has been considered. Carrier mobility is degraded with increasing temperature. The BSIM3 [64] models are used to take the temperature effect into account. The effective carrier mobility is given by

$$\mu_{\text{eff}} = \mu_0 \frac{(T/T_n)^{ute}}{1 + K_1 V_r + K_2 {V_r}^2} \tag{3.8}$$

where μ_0 is the carrier mobility at zero electric field and normal operating temperature. $V_r = (V_{GS} + V_{TH})/t_{\text{ox}}$, $K_1 = U_A + U_{A1}(T/T_n - 1)$, $K_2 = U_B + U_{B1}(T/T_n - 1)$, T and T_n are the operating and nominal (25°C) temperatures, respectively, t_{ox} is the oxide thickness, ute, U_A, U_{A1}, U_B, and U_{B1} are fitting parameters. Since DIBL coefficient, η, is around 100 mV/V and $K1$ and $K2$ are small for current technologies, the effect of V_{DS} on V_{TH} can be ignored and V_r can be approximated by $V_r \approx (V_{GS} + V_{TH0})/t_{ox}$. Therefore, effective carrier mobility can be considered to be independent of V_{DS}.

Using the MOSFET model described above, the analysis of DVT domino gates can be carried out in two phases. Firstly, for the worst case leakage condition, the optimum keeper size to constrain leakage within specification is obtained. Secondly, the optimum keeper size is used to determine performance during the evaluation phase. The two phases are described in more detail below.

3.7 Modeling of Conventional Wide Fan-In Domino Circuits

Domino logic is a member of the dynamic logic family. The output of a dynamic circuit (Figure 3.1) is precharged to LOW when CLK is LOW. Meanwhile, the dynamic node is charged to V_{DD}. The output of the circuit evaluates when CLK is HIGH. If no input is switching during evaluation, the dynamic node stays at V_{DD} and is vulnerable to charge loss due to leakage or any other noise source. For high performance circuits such as microprocessors, the worst case leakage condition occurs at high temperature, typically 110°C, and at the fast process corner. The keeper, shown in Figure 3.1, acts as a feedback transistor to maintain charge on the dynamic node during OFF state. The design criteria for the keeper is to accommodate for worst case leakage. Then, performance is evaluated at worst case, i.e. only one pulldown transistor is turning ON, at slow split, and hot temperature (110°C). Upsizing the keeper leads to a degradation in performance due to the larger contention between the upsized keeper and evaluation transistors.

3.7.1 Optimum Keeper Sizing

As mentioned previously, a keeper size should be optimized to realize the best possible performance under given stability constraints. A similar approach was proposed by Jung *et. al.* [71] to determine the optimum keeper size at normal operating conditions ($25C^{o}$ and typical process corner). In this work, we consider the effect of different environmental and process conditions on the design of the keeper. The keeper is analyzed to compensate for worst case leakage, i.e. for the fast split, elevated temperature (110°C), and a DC noise of 10% of V_{DD} applied to all the pulldown transistors. Keeper is optimized to conform with the UGDN metric as defined in [50] [72] such that noise level at the output never exceeds

that applied at the inputs.

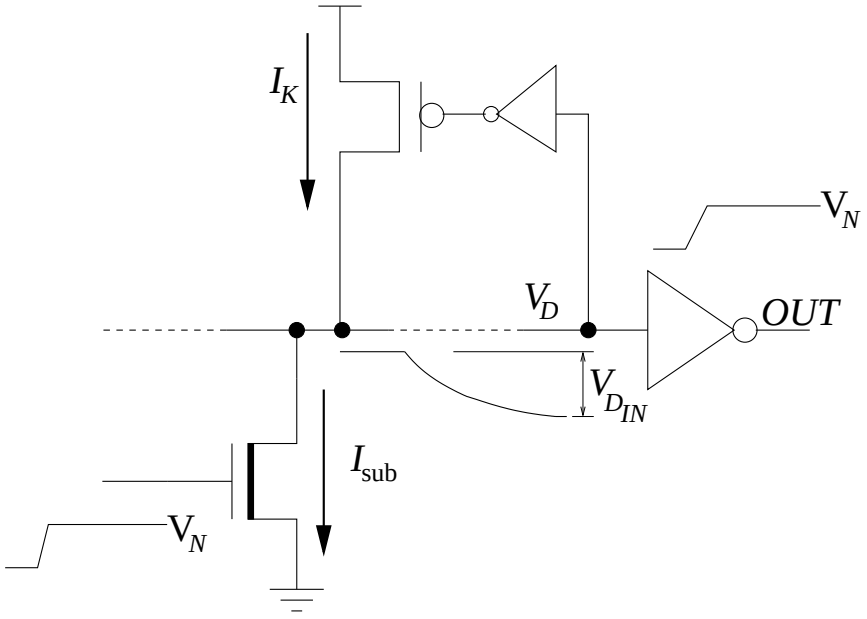


Figure 3.15: Different voltage and current components when all pull-down transistors are subject to DC noise.

The voltage droop of the dynamic node resulting from leakage in the pulldown network is shown in Figure 3.15. Using the transfer characteristics of the output inverter, the input voltage droop, $V_{D_{IN}}$, so that the output noise does not exceed the maximum allowable noise level, V_N, can be obtained.

Modeling keeper current during OFF state when all inputs are subject to DC noise is based on two simplifying assumptions:

- keeper is always operating in the linear mode and never enters in saturation.
- V_{GS} for the keeper can be considered fixed at V_{DD}.

The first assumption is valid since the keeper saturation voltage is larger than $V_{DD} - V_{D_{IN}}$. The second assumption is also valid since the feedback inverter is skewed to suppress most

of the noise on the dynamic node from affecting the keeper input. Assuming a 1μm wide keeper, and using the above assumptions, the keeper current can be expressed using (3.1) as

$$\begin{aligned} I_k = & I_{\mathrm{D0}_k} \left(\frac{V_{DD} - V_{\mathrm{TH0}_k}}{V_{DD} - V_{\mathrm{TH0}_k}} \right)^{\alpha_k} [1 + \lambda_k (V_{DD} - V_{D_{IN}})]. \\ & \left[A - \frac{(V_{DD} - V_{D_{IN}})}{V_{DSAT_k}} \right] . \left(\frac{V_{DD} - V_{D_{IN}}}{V_{DSAT_k}} \right) \end{aligned} \tag{3.9}$$

where the subscript k referes to the device paprameters of the keeper, and I_{k} is the keeper current per unit width. The parameter A is set to 2 in (3.1) [66]. However, we consider $A = 1.9$ in (3.9) for better fitting to HSPICE results.

The optimum feedback keeper size should be able to supply just enough current, I_k, to compensate for leakage in the pulldown network as shown in Figure 3.15. Using (3.7) and (3.9) the optimum keeper size, W_k, can be found by solving the following equation

$$N I_{\mathrm{sub}} = W_k I_k \tag{3.10}$$

where N is the number of transistors in the pulldown network.

The second step in our model is to use the keeper size obtained from (3.10) to estimate the delay of DVT domino circuits. First, the dynamic node capacitance is estimated. Then, the discharge rate of the dynamic node is determined to estimate the delay.

3.7.2 Dynamic Node Capacitance Estimation

The estimation of dynamic node capacitance in the domino circuit is critical for determining the discharge rate. The capacitance to this node is contributed by pulldown transistors, keeper, precharge transistor, feedback inverter and output inverter. Moreover, each of these elements may contribute different capacitive components such as diffusion, gate

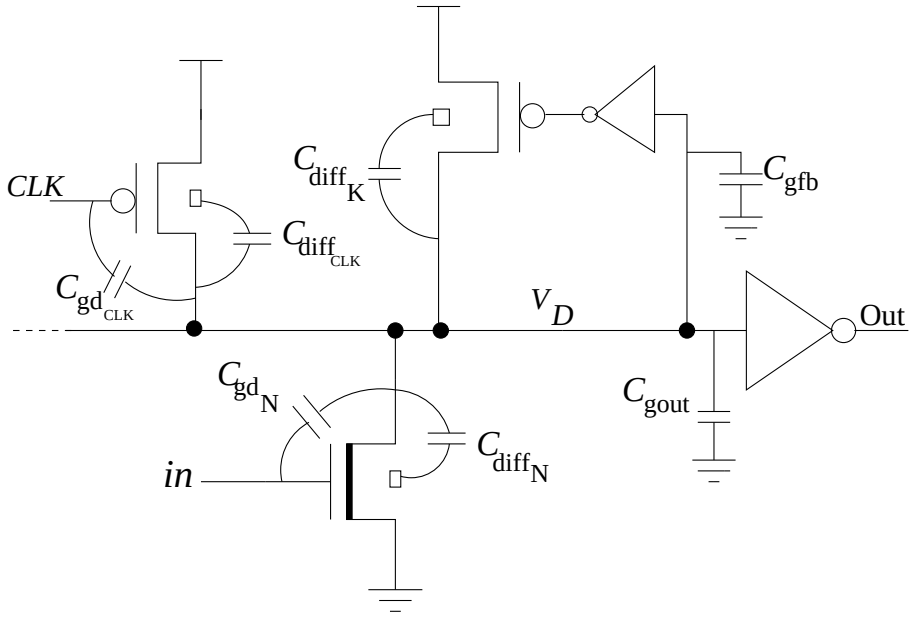


Figure 3.16: Dynamic node capacitive components

overlap, and gate oxide capacitances. These capacitive components can be extracted from technology and process parameters.

Figure 3.16 shows different components for the dynamic node capacitance. Some of these components have a non-linear behavior with respect to the applied voltage and some can be considered constant. The gate to drain overlapping capacitance, C_{gd}, gate capacitance, C_g, and the diffusion capacitance, C_{diff}, are the main capacitive components considered in this analysis.

Diffusion Capacitance: C_{diff}, is quite non-linear. Some approximation for computing C_{diff} was proposed in [7]. However, this approximation results in a relatively large error for DSM technologies. In this work, diffusion capacitance is extracted from simulation.

Gate-Drain Capacitance: C_{gd}, is the overlapping capacitance between the gate and drain areas. C_{gd} is computed for pulldown, keeper, and the precharge transistors and is

given by

$$C_{gd} = 2C_{\mathrm{GDo}}W_{\mathrm{eff}} \tag{3.11}$$

where C_{GDo} is the gate-drain overlapping capacitance per unit width and W_{eff} is the effective transistor width. The factor 2 is accounting for the Miller effect on C_{gd} when the transistor switches. If the gate voltage of the transistor is fixed and only the drain voltage is changing and vice versa, Miller effect is taken to be 1.

Gate capacitance: C_g, of the feedback and output inverters is composed of the gate-drain overlapping capacitance and the gate oxide capacitance for both the PMOS and the NMOS transistors. It can written as

$$C_g = 2C_{\mathrm{GDo}}W_{\mathrm{eff}} + W_{\mathrm{eff}}L_{\mathrm{eff}}C_{ox} \tag{3.12}$$

where C_{ox} is the oxide capacitance per unit area and L_{eff} is the effective transistor length.

Using (3.11), (3.12), and the value of C_{diff} extracted from simulation, the approximate total dynamic node capacitance, C_D is given by

$$C_D = NC_{\mathrm{N}} + C_{\mathrm{keeper}} + C_{\mathrm{CLK}} + C_{\mathrm{gout}} + C_{\mathrm{gfb}} \tag{3.13}$$

where C_{N}, C_{keeper}, and C_{CLK} represent the sum of the diffusion and gate-drain overlap capacitance for the pulldown, keeper, and precharge transistors, respectively. C_{gout} and C_{gfb} are the gate capacitances of the output inverter and the feedback inverter transistors, respectively. The dynamic node capacitance in (3.13) is used to estimate the gate delay.

3.7.3 Delay Estimation

The worst case delay of domino circuits occurs when only one pulldown transistor is turning ON for the slow process corner and high temperature (110°C). At the onset of evaluation, there is a contention between the switching pulldown transistor and the keeper. If the

keeper is large enough, the pulldown transistor might fail to discharge the dynamic node resulting in a logic error. If the keeper is sized properly, the dynamic node is discharged and the output switches high. In this case, the discharge rate of the dynamic node is governed by

$$C_D \frac{dV_D}{dt} = I_k - I_n \tag{3.14}$$

where I_k is the keeper current and I_n is the pulldown saturation current given by (3.1).

Discharging the dynamic node can be divided into three stages as shown in Figure 3.17. During the first stage, the gate-to-source voltage of the pulldown transistor, V_{GS_n}, is increased from $V_{DD}/2$ to V_{DD} and the keeper transistor operates in the linear mode. Before t_0, the dynamic node voltage is assumed to be fixed at V_{DD}. The first stage ends when V_{GS_n} reaches V_{DD}. Therefore, the elapsed time in stage 1 is equal to $\tau_r/2$, where τ_r is the input rise time. During the second stage, the keeper is still operating in the linear mode of operation, however, V_{GS_n} is fixed. The second stage ends when V_{DS} of the keeper exceeds the saturation voltage, V_{DSAT_k}, and the keeper becomes saturated. The period of interest in this analysis is the time between the moment when input is 50% of V_{DD} till when the output reaches 50% of V_{DD} (delay calculations). The keeper gate voltage is assumed to be fixed at zero during the entire period, t_0 till t_3. Simulation results in Figure 3.17 confirm this assumption. The keeper input and output voltages of the dynamic gate, hence V_{GS} of the keeper, are fixed during the time interval (t_0 to t_3) as indicated in Figure 3.17.

During the first stage, the gate voltage V_{GS_n} is not fixed. Rather, it is increasing to reach V_{DD}. Since V_{GS_n} is time dependent, solving (3.14) requires a simplified expression for I_n. I_n is simplified by expanding the term $(V_{GS_n} - V_{TH0})^\alpha$ into its Taylor series expansion. Details of solving (3.14) are given in Appendix A. The solution of (3.14) at input $V_{GS_n} = V_{DD}$, results in the value of V_{D1}, as shown in Figure 3.17.

The second stage of the dynamic node discharging process starts from V_{D1} and ends

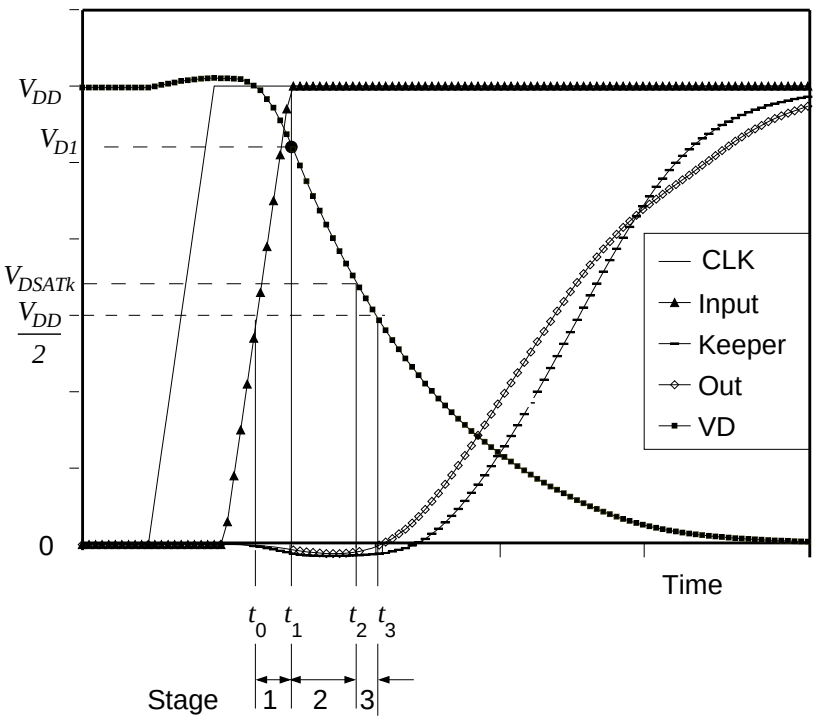


Figure 3.17: Transient Response of a Conventional 16-input Domino Gate

when V_D reaches V_{DSAT_k} for the keeper. V_{GS_n} is fixed at V_{DD}. Therefore, the current I_n in (3.14) depends only on V_D and is independent of V_{GS_n}. Hence, (3.14) can be solved analytically to compute t_2 using

$$t_2 = C_D \int_{V_{d1}}^{V_{DSAT_k}} \frac{dV_D}{I_k - I_n}. \tag{3.15}$$

Details of simplifying and solving (3.15) are shown in Appendix B.

Finally, the keeper becomes saturated in the last stage. Therefore, keeper current is the saturation current, I_{DSAT} in (3.1). V_D is the only time varying parameter in (3.14). Therefore, an analytical solution of (3.14) similar to (3.15) can be obtained. The resulting solution, t_3, is the time elapsed between $V_D = V_{DSAT_k}$ and $V_D = V_{DD}/2$ as shown in Figure

3.17, is given by

$$t_3 = C_D \int_{V_{\mathrm{DSAT}_k}}^{V_{\mathrm{DD}}/2} \frac{dV_D}{I_k - I_n}. \tag{3.16}$$

A simplified form of (3.16) is derived in Appendix B.

Using (3.15) and (3.16), the total delay for the dynamic node to discharge from V_{DD} to $V_{DD/2}$ is given by

$$T_d = \tau_r/2 + t_2 + t_3 \tag{3.17}$$

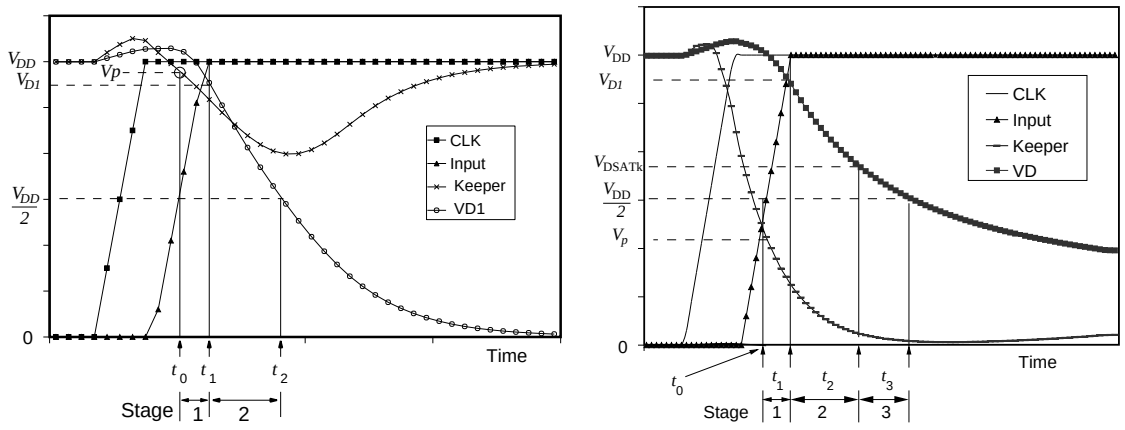
Using the above mentioned model, a domino circuit can be analyzed for its delay and noise stability for different design and technology parameters (e.g., fan-in, V_{TH}, etc.). In the next section, the model is extended to analyze more complex domino structures. In Section 3.9, the effect of V_{TH} reduction and fan-in on performance is analyzed.

3.8 Model Extension to Complex Designs

The delay model described above is extended to model more complex design styles. In the section, the SD gate, shown earlier in Figure 3.3, is chosen to demonstrate the generality and accuracy of the devised model described in 3.7. Due to the symmetrical nature of the circuit, the worst case delay can be analyzed using only one section of the circuit. For worst case leakage analysis, all transistors in both networks are subjected to noise. Therefore, one section is sufficient for the analysis. Worst case delay occurs when only one transistor in one of the two networks is switching ON. The other network does not influence the circuit performance. As before, analysis of SD circuit involves two steps: optimal keeper sizing and delay estimation.

Keeper sizing is performed according to the UGDN criteria. The two dynamic node voltages, V_{D1} and V_{D2}, are subjected to noise. The transfer characteristics of the NAND3 gate determines the maximum allowable noise, $V_{D_{IN}}$, at any of the two dynamic nodes.

Since both voltages are used as inputs to the NAND3 gate, the capability of the NAND3 gate to suppress noise is weaker than the feedback inverter in conventional domino. As a result, gate voltage of the keeper transistors is lower than that in the conventional case. For a specified output noise, the dynamic node droop is larger compared to that of the conventional circuit. Therefore, total keeper size of SD gate is usually larger than the conventional keeper to maintain same noise stability. The optimal keeper size can be found using (3.10) and the new $V_{D_{IN}}$. The number of pulldown transistors, N in (3.10) is divided by 2 since only one network is sufficient for the analysis. The resulting W_K represents the width of either *K1* or *K2*. The NAND3 gate is designed to drive the keeper inputs and is sized proportional to the keeper.



(a) $\tau_k \geq \tau_r$ (b) $\tau_k < \tau_r$

Figure 3.18: Simulation Waveforms for Split Domino

Delay estimation of SD circuit follows the same methodology described earlier. The worst case delay is when only one pulldown transistor in any of the two evaluation networks is turning ON for the slow split and hot temperature. As mentioned before, keepers in SD are OFF at the instance CLK makes a positive transition. If the circuit does not evaluate,

keepers turn ON after NAND3 gate delay. On the other hand, if the circuit evaluates, the keeper input voltage turners back OFF (as shown in Figure 3.18 (a)). The slope of the keeper input can be determined using the transfer characteristics of the NAND3 gate. In our analysis, the *keeper* input is assumed to be linear for the entire delay estimation period. Given the slope of the *keeper* input, τ_k, the starting keeper voltage, V_p, at t_0 can be obtained. Then, V_{GS_k} for the keeper can be approximated by

$$V_{GS_k} = V_p - V_p(t/\tau_k)$$

where V_p is the keeper gate voltage when $t = t_0$ as shown in Figure 3.18 (a) and (b). Data is assumed to arrive after *CLK*. Hence, the term $(V_{GS_k} - V_{TH0})^\alpha$ in (3.1) is time dependent and should be simplified using Taylor series expansion.

The slope of the *keeper* input signal determines the different keeper modes of operation throughout the discharge period of the dynamic node. When $\tau_k > \tau_r$, the keeper can fairly be assumed to operate in the linear mode and never in saturation. In this case, discharge of the dynamic node can be divided into two stages as shown in Figure 3.18 (a). On the other hand, when $\tau_k < \tau_r$, the keeper enters saturation after being in the linear mode. In this case, delay estimation is divided into three stages as shown in Figure 3.18 (b).

During the first stage in both of the above mentioned situations, the keeper operates in the linear mode while the pulldown transistor is saturated. Both the keeper and pulldown inputs are changing. Therefore, Taylor series expansion of $(V_{GS} - V_{TH0})^\alpha$ for both currents is used to simplify the expression in (3.14). Solving (3.14) for V_D at $t = t_1 = \tau_r/2$ results in node voltage V_{D1}. Subsequently, V_{D1} serves as the starting point for the second stage. During the second stage, the input of the pulldown transistor becomes fixed at V_{DD}. The *keeper* gate voltage is still time dependent and Taylor series expansion is used to simplify its drain current expression. The second stage is bounded by $V_D = V_{DSAT_k}$ when $\tau_k < \tau_r$ or $V_D = V_{DD}/2$ when $\tau_k > \tau_r$, as shown in Figure 3.18 (a) and (b) respectively. Substituting

of the current expressions in (3.14) and solving results in the time bound of the second stage, t_2.

When $\tau_k < \tau_r$, the keeper starts to operate in saturation where its V_{DS} exceeds V_{DSAT_k}. This marks the beginning of a third stage which ends at $V_D = V_{DD}/2$. The solution of the resulting expression is the time, t_3, elapsed from V_{DSAT_k} to $V_{DD}/2$. Therefore, total delay time for SD gate is given by

$$T = \begin{cases} t_1 + t_2 & (\tau_k \geq \tau_r) \\ t_1 + t_2 + t_3 & (\tau_k < \tau_r) \end{cases} \tag{3.18}$$

In the following section, delay models described in Sections 3.7 and 3.8 are compared with HSPICE simulation results. Moreover, delay models are also used to analyze the effect of threshold voltage reduction and fan-in on performance for a given noise constraint. A comparison between performance of the conventional and the SD structures is also presented using the previously mentioned model.

3.9 Optimization of Wide Fan-In Domino Gates

In order to verify our model, simulations of a 4-bit, 8-bit, 16-bit, and 32-bit DVT conventional and SD gates are performed. UGDN is chosen to be 10% of V_{DD}. All simulations are carried out in the 0.13μm CMOS technology. Threshold voltage of evaluation transistors is altered by changing V_{TH0} and channel doping, N_{CH}, in HSPICE BSIM 3.3 models. Table 3.2 indicates the different evaluation devices used in our model and the corresponding model parameters for worst case leakage condition, i.e. fast split and 110°C. As threshold voltage of devices in Table I is reduced, leakage current is increased exponentially. Leakage current increases by 34× for the low V_{TH} device 1 compared to high V_{TH} device 5. The effect of subthreshold slope and DIBL on the OFF current is magnified as V_{TH} is decreased.

Devices 1 and 2 in Table I operate in the moderate inversion region when DC noise is 10% of V_{DD} since V_{TH} is less than noise voltage. Therefore, subthreshold slope is not applicable for these transistors as indicated in Table 3.2.

Table 3.2: Leakage Current Model Parameters of different Pulldown Threshold Voltages at Worst case (Fast split, 110°C)

Device No.	V_{TH0} (V)	S_t (V/decade)	η (V/V)	$I_{\text{sub}}(@V_{DS} = V_{DD}, V_{GS} = 0.1V_{DD})$ (μA / μm)
1	0.123	N/A	0.086	22
2	0.148	N/A	0.084	15
3	0.198	0.123	0.081	6.2
4	0.256	0.115	0.079	1.96
5	0.310	0.110	0.077	0.65

Model parameters for worst case delay, i.e. slow split and 110°C for the same device depicted in Table 3.2, are shown in Table 3.3. Saturation current, I_{D0}, increases by 40% when V_{TH0} is decreased by 57% from 0.432V to 0.247V. λ remains almost constant while α decreases as V_{TH0} is decreased. The effect of threshold voltage reduction on the ratio between leakage and saturation currents can be seen from the data provided in Table 3.3. For example, leakage current is increased by 71× while saturation current is increased by only 1.4× for device 1 compared to device 5. Meanwhile, threshold voltage is affected by process variations. This can be seen from data provided in Tables 3.2 and 3.3. Threshold voltage of device 1 and the fast split is decreased by 60% compared to device 5. However,

V_{TH} of the same two devices is decreased by 43% for the slow split.

Table 3.3: Model Parameters for Worst Case delay (Slow split, 110°C) and different Threshold Voltages

Device No.	V_{TH0} (V)	α	I_{D0} (μA / μm)	λ	I_{OFF} (nA)
1	0.247	1.13	419	0.24	151.4
2	0.269	1.14	403	0.24	85.02
3	0.310	1.16	367	0.25	27.36
4	0.380	1.17	330	0.25	7.22
5	0.432	1.19	300	0.25	2.11

As threshold voltage is reduced, keeper has to be upsized to compensate for worst case leakage. Such trend can be predicted using the proposed model. Figure 3.19 illustrates the optimized keeper size using our model as a function of V_{TH} for different fan-in. It can be seen that keeper should be enlarged when V_{TH} of pulldown transistors is reduced. The extent of keeper upsizing becomes larger as the fan-in is increased. For example, keeper is only 1.5× and 2× larger when using low threshold (device 1) instead of high threshold (device 5) for a 4-bit contentional and SD gates, respectively. However, for larger fan-in such as 32-bit, the keeper is 15× and 18× larger for low threshold compared to the high threshold implementation for the conventional and SD gates, respectively. These modeled results are closely matched by HSPICE simulations.

Delay modeling as a function of V_{TH0} is illustrated in Figure 3.20. We utilized different

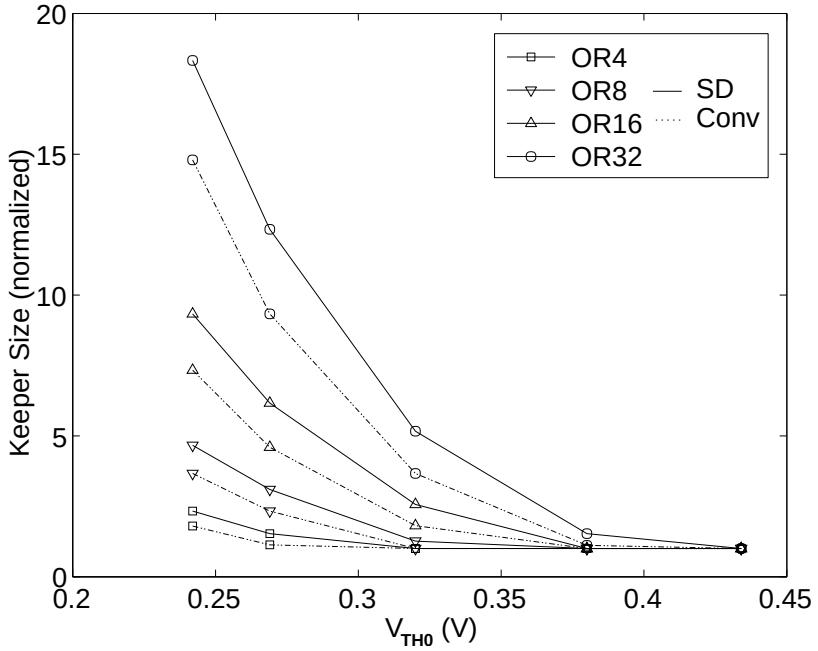


Figure 3.19: Keeper Device Size for 4, 8, 16, and 32 -input Conventional Domino.

device parameters as indicated in Table 3.3 for 4, 8, 16, and 32-bit conventional domino circuits. The delay estimation method detailed in Section 3.7 closely follows delay extracted from HSPICE simulations. The maximum error is 6%.

The effect of V_{TH} reduction and fan-in number on performance of conventional domino can be quickly analyzed using the proposed model. When V_{TH} is decreased, delay of 4 and 8-bit gates monotonically decreases. When using device 1 instead of device 5, the maximum performance gain is 27% and 23% for the 4-bit and 8-bit gates, respectively. On the other hand, delays of the 16 and 32-bit gates is decreased as V_{TH} is decreased to a certain extent. If V_{TH} is reduced any further, delay starts to increase. The reduced threshold results in an increased leakage current and the keeper has be upsized accordingly. For example, for $V_{TH0} = 0.246$V in the 32-bit gate, delay increases by 15% compared to the delay at the

original threshold of 0.432V. Similar results for 8-bit DVT domino circuit were presented in [57]. Therefore, based on the fan-in number, reducing V_{TH} of pulldown transistors can result in a performance degradation rather than performance enhancement.

The optimum V_{TH} which results in the best performance can be found using our model described earlier. The optimum V_{TH} conforms with that obtained using HSPICE simulation as shown in Figure 3.20. For both 16 and 32-bit gates, a V_{TH0} of 0.310V results in the best performance under the predefined noise constraint. Any further reduction in V_{TH} for those gates results in less performance enhancement.

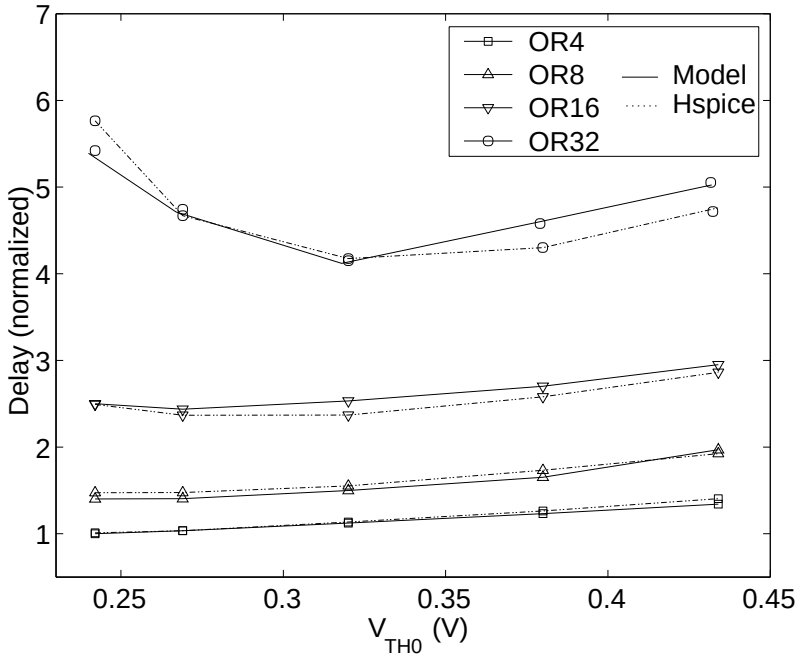


Figure 3.20: Simulated and Modeled Delay of 4, 8, 16 and 32-bit Conventional Domino gates vs. Threshold Voltage for the slow split and 110°C normalized to Delay of a High Threshold 4-bit Conventional Domino Gate.

The effect of V_{TH} reduction on performance of 4, 8, 16, and 32-bit SD gates compared to HSPICE simulation is shown in Figure 3.21. The proposed model can predict delay within

7% accuracy compared to HSPICE. Delay reduction when V_{TH0} is reduced from 0.432V to 0.246V is estimated to be 21% for the 4-bit gate. Delay of the 8-bit gate remains almost unchanged when using device 1 instead of device 2. The maximum delay improvement is 19% for the 8-bit gate. Delay reduction almost vanishes for the 16-bit gate while it increases by as much as 68% for the 32-bit gate when $V_{TH0} = 0.247$V compared to delay at $V_{TH0} = 0.432$V. From Figure 3.21, it is apparent that the optimal V_{TH0} for the 16 and 32-bit SD gates is 0.310V which is similar to conventional domino gates.

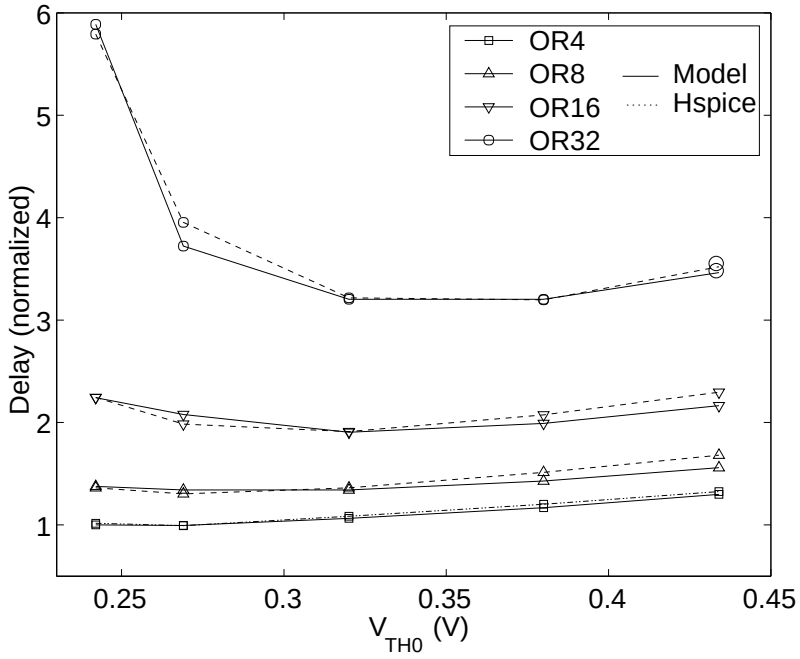


Figure 3.21: Simulated and Modeled Delay of 4, 8, 16 and 32-bit Split Domino gates vs. Threshold Voltage for the slow split and 110°C normalized to Delay of a High Threshold 4-bit Split Domino Gate.

A comparison between performance of conventional and SD gates is shown in Figure 3.22. Performance enhancement of SD increases as fan-in number increases. For high threshold device 5 ($V_{TH0} = 0.432$V), delay improvement starts at 24% for 4-bit SD gate

and reaches 85% for a 32-bit gate compared to conventional gates. This is due to the reduced dynamic node capacitance and reduced contention.

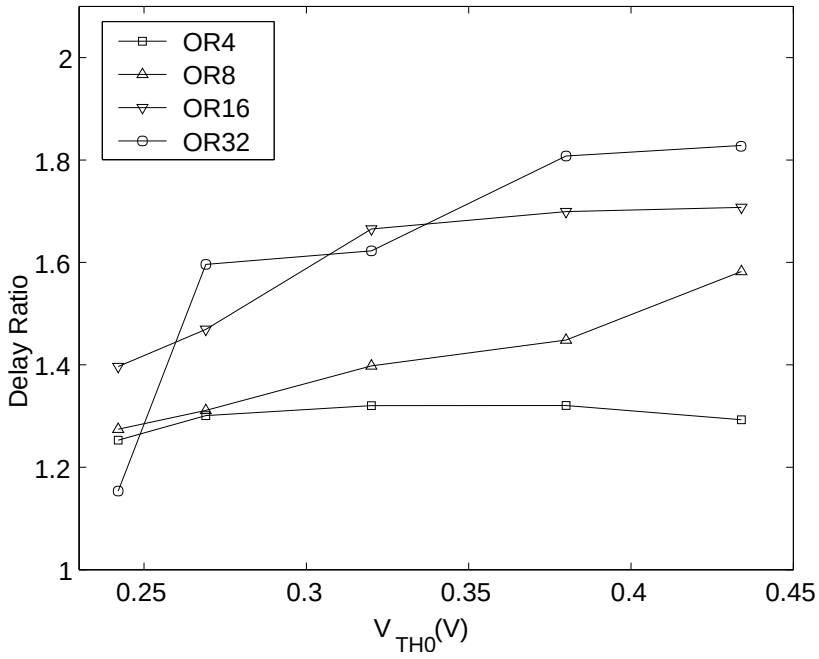


Figure 3.22: Ratio of Conventional Delay to SD delay.

The impact of V_{TH} reduction on performance enhancement of wide fan-in SD gates is more negative than in the case of conventional circuits. Performance enhancement is reduced from 85% to 28% in the 32-bit gate as a result of reducing V_{TH0} from 0.432V to 0.246V. The same trend applies to the 8, and 16-bit gates with a smaller impact on performance. For a 4-bit gate, SD performance advantage over conventional domino remains almost unchanged as V_{TH} is reduced. With V_{TH} reduction, keeper transistor should be upsized to limit output noise. At $V_{TH0} = 0.246$V, keeper of the SD gate is larger than that of the conventional regardless of fan-in as shown in Figure 3.19. The NAND3 gate in the SD structure should be upsized progressively with the keeper to maintain noise within bounds.

Therefore, performance enhancement of SD gates is degraded with reduced threshold.

3.10 Summary

A leakage tolerant energy efficient split domino (SD) circuit technique for wide fan-in gates has been presented. The SD 32-input OR gate is 32% and 12% faster compared to the conventional and the conditional keeper counterparts respectively. The proposed technique offers 12% and 27% energy reduction for 16-input and 32-input OR gates, respectively, compared to conventional domino. Relative to the conditional keeper technique, the corresponding savings are 12% and 18% respectively. Energy savings are a result of the faster evaluation of the proposed technique due to the reduced dynamic node capacitance and reduced contention power.

The speed, power, and energy advantage of SD over conventional domino is greatly enhanced as supply voltage is scaled down. Delay improvement of SD over conventional is shown to be 28%, 61%, and 87% compared to conventional at 0.8V for the 8, 16, and 32-bit gates respectively. Power dissipation of the 8-bit SD is slightly improved with supply scaling. However, 60% and 75% power dissipation reduction is shown for the 16 and 32-bit SD over conventional domino at 0.8V. As a result of delay and power dissipation improvement, energy reduction of SD over conventional is shown to be 1.4×, 3.3×, and 4.1× when supply voltage is scaled from 1.4V to 0.8V for the 8, 16, and 32-bit gates respectively.

In order to optimize domino circuits, performance of conventional domino circuits is analyzed as a starting point. Analysis is performed in two steps. First, the optimum keeper size is computed. Then, a delay estimation model has been proposed to predict performance. Delay estimated using the proposed model is within 6% accuracy compared

to HSPICE. Furthermore, the model is extended to estimate delay of split domino gates with 7% accuracy comapred to HSPICE. The proposed model is used to analyze the impact of threshold voltage reduction and the fan-in number on performance for a given noise constraint. It has been shown that V_{TH} reduction of the evaluation transistors beyond a certain extent has a negative impact on performance of wide domino gates. Using the proposed models, the optimal threshold voltage, for a particular fan-in, can be predicted to achieve the best performance under a given noise constraint. Also, the proposed models are used to compare performance of conventional and split domino gates. Split domino is faster than conventional domino due to less contention at the beginning of evaluation and lesser dynamic node capacitance. For high threshold implementations, a 32-bit SD gate has an 85% performance improvement over conventional implementation. As V_{TH} is reduced, the performance advantage of wide SD gates is degraded since keeper mechanism has to be upsized to counter the increasing leakage current. Delay advantage of 32-bit SD gate over conventional gate shrinks from 85% to 28% when threshold voltage is reduced by 43%.

Chapter 4

Robust and Efficient Voltage Scaling Architectures for Dynamic Power Reduction

4.1 Introduction

Portable devices such as personal digital assistants (PDAs), cellular phones, and portable computers are becoming part of the daily life. The high demand for more applications and functions integrated into these portable devices has pushed the design trend towards higher integration. Yet, the more sophisticated the portable device is, the more its energy consumption and the less its battery life. Long battery life is a very important design and marketing parameter. A great deal of design effort is devoted to extending battery life time while keeping the same level of performance.

Designing more versatile portable devices is becoming more feasible as the technology scales. With smaller feature size, more integration and more functions can be built within

the same area. Energy reduction techniques are essential in designing such systems. The most effective energy reduction method is supply voltage scaling due to the quadratic dependence of energy on voltage. The active dynamic energy dissipation for CMOS circuits is given by

$$E_{\text{act}} \propto C_{\text{avg}} V_{DD}^2 \tag{4.1}$$

where V_{DD} is the supply voltage. C_{avg} is the average switching capacitance and is given by $C_{\text{avg}} = C_{\text{gate}} + C_{\text{diff}} + C_{\text{wire}}$ where C_{gate}, C_{diff}, and C_{wire} are the average switching gate, diffusion, and wire capacitance for the chip respectively. C_{diff} is voltage dependent but its average value used here is assumed to be voltage independent.

The minimum allowable supply voltage V for a static CMOS inverter was derived in [73] and used in [6] and is given by

$$V_{\text{min}} = \beta V_T \tag{4.2}$$

where β is a constant between 3 and 4 and V_T is thermal voltage (26 mV at room temperature).

Lowering supply voltage has been proved to save energy. Recently, a fast fourier transform (FFT) unit was shown to work at 350 mV to provide optimal energy efficiency [45]. The FFT unit was also shown to function correctly at a supply voltage of 180 mV.

Peak supply voltage is selected based on peak performance requirements. Occasionally, peak performance is not required by the processing unit. Therefore, supply voltage can be scaled when maximum performance is not required. The software interface can provide information about performance requirements. This information can be used to reduce supply voltage based on the required speed. By exploiting the variation in computational requirements, supply voltage can be scaled and average energy of the system can be reduced while maintaining the required throughput. As a result, battery life time can be extended.

Figure 4.1 shows the power dissipation profile of a typical burst-mode application such as cellular phones. Such an application has only two modes of operation: active and idle. The device operates at full throughput for a small period of time (active mode) before entering the idle mode (standby). Given a deadline to be met for a given task, reducing supply voltage alone would lead to timing violations. Meeting the deadline requires satisfying the circuit delay constraint. This delay is generally expressed as

$$t_d \propto \frac{CVDD}{(V_{DD} - V_{TH})^{\alpha}} \tag{4.3}$$

where C is the switching capacitance. In order to meet the specified deadline at the scaled supply voltage, the threshold voltage of the device, V_{TH} is reduced. However, threshold voltage reduction leads to an exponential increase in subthreshold leakage. Therefore, leakage reduction methods are utilized to reduce leakage power during the idle mode.

Systems with more than two modes of operation where varying throughput is required during the period of operation require voltage to scale based on performance requirements to save power. Figure 4.2 shows throughput of various tasks and their scheduling. Task deadlines dictate the way voltage can be scaled. Given the flexibility of extending the various computational tasks over time without conflicting with a hard deadline, voltage can be reduced until task requirements are completed. The system enters the idle mode afterwards.

Figure 4.3 shows how flexible task scheduling can be achieved. The active period is extended while the idle period is shortened to help save power. When task scheduling is restricted to certain deadlines, there is no flexibility in reducing supply voltage of a running task and extending its execution time beyond the specified deadline. However, not all tasks require full throughput as shown in Figure 4.2. Therefore, supply voltage can be scaled dynamically based on throughput requirements as shown in Figure 4.4.

Dynamic voltage scaling is also effective in reducing standby power [74]. By using two

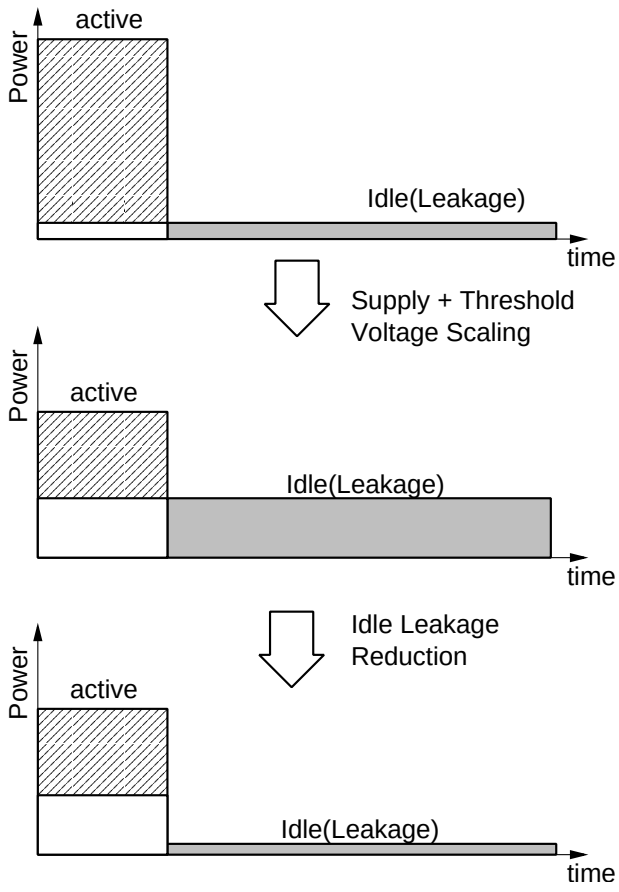


Figure 4.1: Power reduction through reducing supply voltage.

supply voltages, one for logic and one for flip-flops, standby power can be reduced. Figure 4.5 shows the two supply voltage configuration. Both the combinational and sequential supply voltages utilize dynamic voltage scaling to save power during the active mode. During standby, the combinational supply voltage is collapsed (shut down) using the sleep transistor technique [25]. Meanwhile, the sequential supply voltage is reduced to the level just before the stored state is destructed. Saving the state within the flip-flops reduces the power required to store and restore contents. Therefore, optimal power savings can be

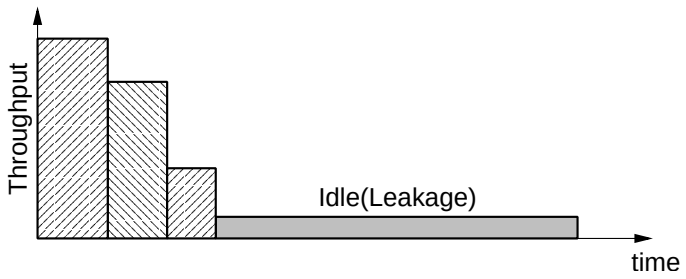


Figure 4.2: Throughput required for a certain application.

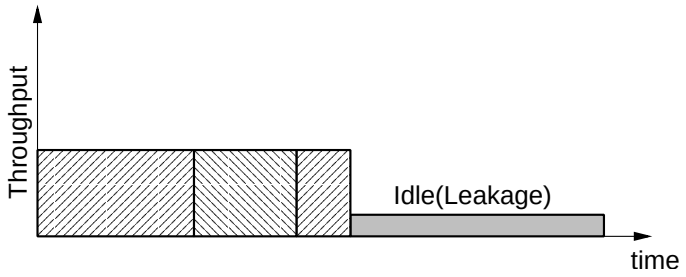


Figure 4.3: Task Scheduling with Dynamic Voltage Scaling.

achieved.

Similar to dynamic voltage scaling, adaptive body biasing has been proposed to control the increasing leakage current in deep sub-micron technologies. The body connection of the transistors is controlled by applying a reverse body voltage to increase the threshold voltage and decrease leakage current [75]. Combining dynamic voltage scaling and adaptive body bias results in more energy saving compared to each one alone [76]. A multiply-accumulate (MAC) unit was designed to work at low supply voltages by applying both dynamic voltage scaling and adaptive body bias [77]. The MAC unit was shown to work at 175 mV with frequency of 166 KHz. This powerful combination of both active and leakage current control capabilities has been used to mitigate the impact of process variations by adaptively changing supply voltage and body bias based on the actual silicon conditions

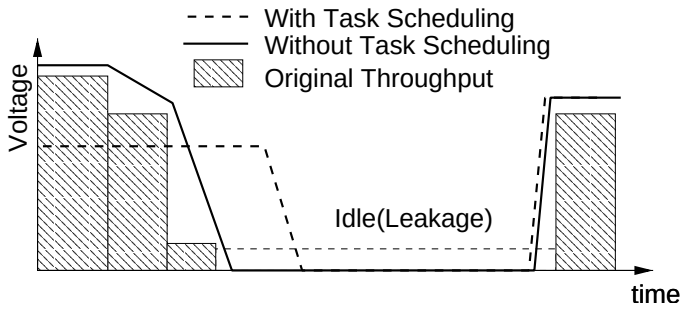


Figure 4.4: Dynamic Voltage Scaling with and without Task Scheduling.

[78] [79].

Adaptive body biasing can only be accomplished effectively when the body connection of the transistor is accessible. Such a feature is optional and usually adds extra cost to the fabrication process (e.g. triple well). In fact, modern CMOS technologies have started to shift towards the triple well option as a default to give designers better control over the device. However, the effectiveness of using body control is decreasing as technology scales [80] [81]. Other advanced technologies such SOI can be explored in the future to enable body biasing in the sub-100nm technologies.

4.2 Dynamic Voltage Scaling Systems for Deep-Submicron Technologies

Dynamic Voltage Scaling (DVS) systems adjust supply voltage according to throughput requirements. Figure 4.6 shows the overall architecture of a DVS system. The performance manager uses a software interface to predict performance requirements. Once performance requirement for the next task is determined, the performance manager sets the voltage and frequency just necessary to accomplish the task. The target frequency is sent to the

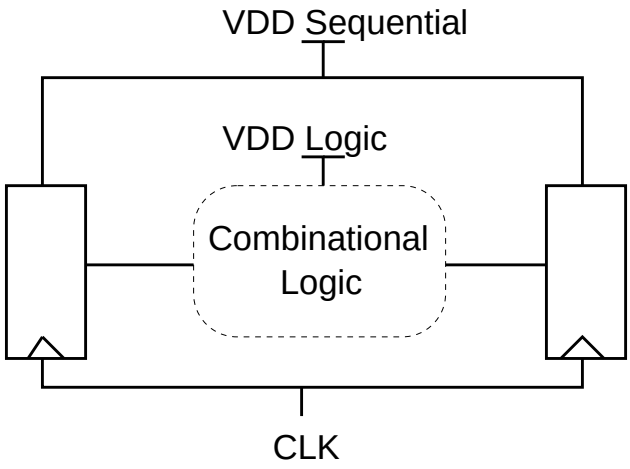


Figure 4.5: Two power supplies scheme for low standby power applications. V_{DD} **Combinational is gated during standby and** V_{DD} **Sequential is scaled to lowest voltage which preserves the state.**

phase-locked loop (PLL) to accomplish frequency scaling. Based on the target voltage, the voltage regulator scales supply voltage to meet performance target.

The actual performance of the core running under scaled voltage has to be characterized to guarantee a fail-safe operation while maintaining the required performance. A robust system should be able to meet the deadlines at any voltage, process, interconnect and temperature condition. system performance depends on the underlying voltage scaling methodology. The conventional approach to perform voltage scaling uses a target operating voltage for each required operating frequency. To guarantee a robust operation, the frequency-voltage relationship is determined via chip characterization at worst case conditions. This technique is utilized in open-loop dynamic voltage scaling (DVS) systems where the frequency-voltage relationship is stored in a look-up table (LUT). Since such LUT is pre-loaded with voltage-frequency points, DVS systems are not able to adapt to process variations or environmental conditions.

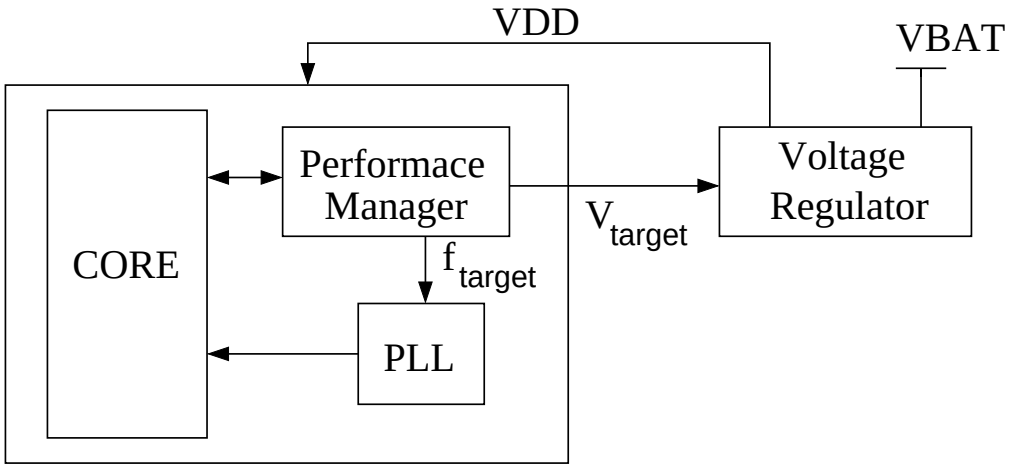


Figure 4.6: Architecture of a Dynamic Voltage Scaling System.

Alternatively, the critical path of the system can be duplicated to form a ring oscillator which adaptively responds to environmental and process variations. Also, the critical path replica can be replaced by fan-out of 4 (FO4) ring oscillator [82] or a delay line [83]. In both cases, a closed-loop mechanism based on adaptive voltage scaling (AVS) is formed by monitoring the actual silicon speed. Therefore, worst case characterization is no longer required. Since there is a direct relationship between the actual performance of the core and the speed of the ring oscillator (or the delay of the delay line), AVS systems adaptively adjust supply voltage to nearly the minimum level required to meet performance targets. A safety margin is added to account for any mismatch between the ring oscillator (or the delay line) and the actual critical path.

Different parameters are involved when selecting between the two different configurations. Stability against temperature change is a main design parameter. The conventional open-loop DVS stores the worst case performance numbers. Therefore, worst case process variation is covered and temperature stability is guaranteed. The large margin added to compensate for process and temperature variation can reduce energy savings significantly.

When monitoring the actual system performance, the AVS system compensates for process and temperature variation. Closed loop parameters and system response determine the time required by the feedback system for voltage fine-tuning. If the rate of change in temperature is faster than the closed loop response time, the system enters the panic mode and voltage has to be ramped up immediately to its worst case setting. That worst case setting corresponds to the worst case process and worst case temperature. The panic mode, where voltage has to guarantee the maximum performance under all circumstances, has not been addressed properly in closed loop AVS systems such as [82] and [22]. Energy efficiency of such systems should be carefully analyzed when the panic mode is frequently encountered.

4.2.1 Open-loop DVS

The open-loop system usually uses supply voltage as the control variable to adapt and reach the specified target performance. As mentioned earlier, open-loop system is not able to respond to environmental variation (e.g. temperature). Therefore, worst case scenario is taken into account. Figure 4.7 shows the entire control loop for the open-loop DVS system. The buck regulator and a comparator are the main components in the open-loop system. The target voltage is specified by the system's software or by a LUT and stored in the target voltage register.

The buck converter (sometime called switching regulator), shown inside the dotted box in Figure 4.7, is a DC-DC converter used to control supply voltage to reach the target. More details about the buck converter and its components are published in [84] and [82]. The main components of the buck converter are the filter, the Pulse Width Modulator (PWM), an analog-to-digital converter (A/D), and the power switches (the PMOS and NMOS). The A/D converter converts the analog supply voltage to a digital word which is

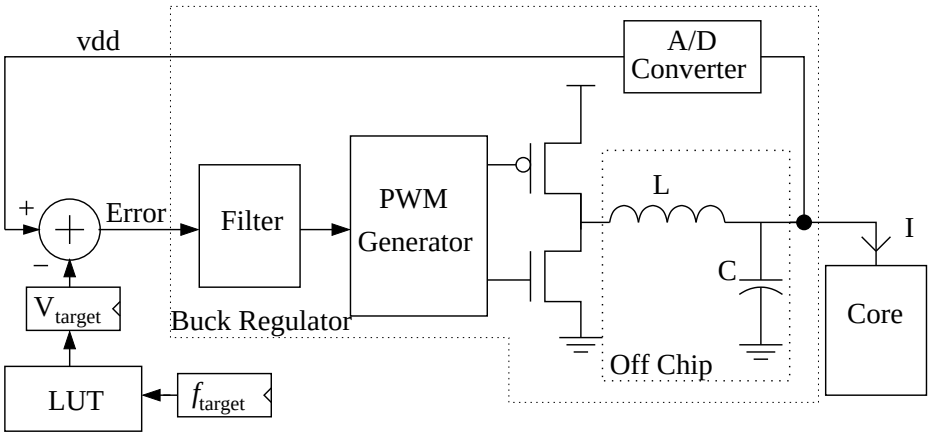


Figure 4.7: Open-loop DVS

compared to the target digital word. The resulting error is filtered down using the filter. The PWM module is used to generate a pulse-width modulated signal where the pulse width of the output signal is proportional to the target voltage. The power transistors in addition to the off-chip inductor and capacitor convert the PWM signal into a DC voltage which eventually (after the system reaches stability) will be close to the target voltage (within a certain error allowed by the system).

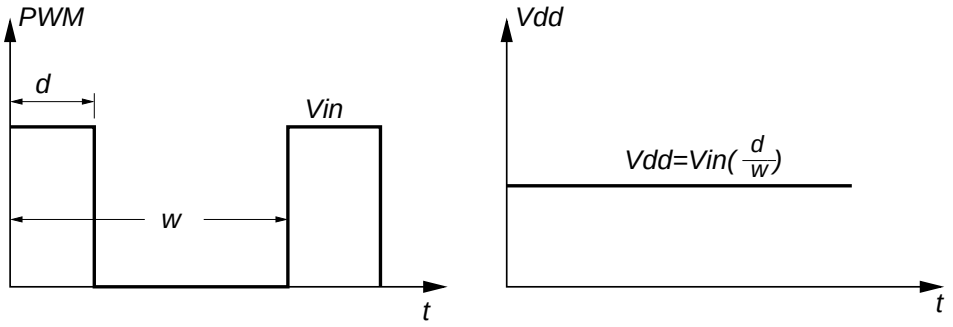


Figure 4.8: Converting the PWM signal to a DC voltage.

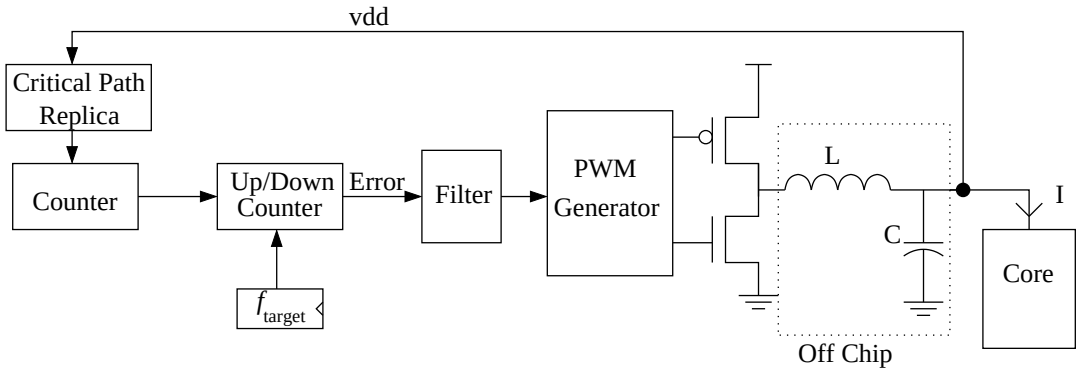
The typical switching frequency of the buck converter is around 1 MHz. Such a slow

switching frequency and large values for the inductor and capacitor are essential in achieving high efficiency. Since the load current is typically high, a small integrated inductor and capacitor is inefficient. A recent study showed that increasing the switching efficiency can yield highly efficient fully integrated switching regulators [61] [85]. The error between the target and the actual voltages is passed through the filter and used to fine tune the system response until the error goes below a certain threshold and the system enters a stable condition. When the target voltage is changed, the system adapts to the change and a stable supply voltage that satisfies performance constraint is achieved.

4.2.2 Closed-loop DVS

One of the main drawbacks of the open-loop system is that the voltage is set to accommodate for worst case scenario. There is no chance to get feedback about how close the system is running compared to a target performance. This kind of feedback forms the essence of the closed-loop DVS system (often called Adaptive Voltage Scaling, AVS). The actual performance is monitored using on-chip structures. A fan-out of 4 (FO4) inverter chain in the form of a ring oscillator is often used. This is is due to the fact that voltage scaling behavior of the FO4 inverter mimics that of most other static CMOS gates [22] [86]. The frequency of the ring oscillator is sampled using a counter as shown in Figure 4.9. The frequency count is then compared to the frequency required by the system and the difference (error) is filtered using the system's filter. The rest of the loop is similar to that described above for the open-loop DVS.

The FO4 inverter ring oscillator has a difference of approximately 14% compared to certain types of gates (e.g. 3-input NAND gate) [86]. Moreover, voltage scaling characteristics of dynamic gates (diffusion dominated) has up to 28% difference with respect to that of the FO4. This difference has to be built in the ring oscillator to accommodate for



(a) Frequency based loop

Figure 4.9: Closed-loop DVS.

all types of gates and all conditions. A better approach is to use a critical path replica as shown in Figure 4.9. The combination of gates forming the critical path of the system for which supply voltage is dynamically scaled is duplicated including the interconnection wires between the gates. The critical path replica provides the closest behavior to the actual critical path except for cross coupling capacitances which are difficult to duplicate. This difference was somewhat accounted for in [87]. Figure 4.10 shows that duplicating the critical path with 3-5% margin yields an efficient DVS system. The regulated voltage *RVDD* is used to supply two copied of the critical path. One of the two copies has a 3-5% for any mismatch with respect to the actual critical path. The two critical path replicas are inserted in between flip-flops representing a single stage of the pipeline running under DVS. A third path only includes the flip-flops so that only clock delay is considered. The system operates by adjusting the supply voltage *RVDD* to guarantee that the middle path runs without timing errors and the top path is failing. That means the supply voltage is just enough for correct functionality plus less than 5% margin. When the top path also passes timing, the timing controller is programmed to reduce the duty cycle of the PWM

controller and lower the supply voltage until the top path fails and the middle path passes timing. The digital type filter introduced in [87] saves power and improves the overall efficiency.

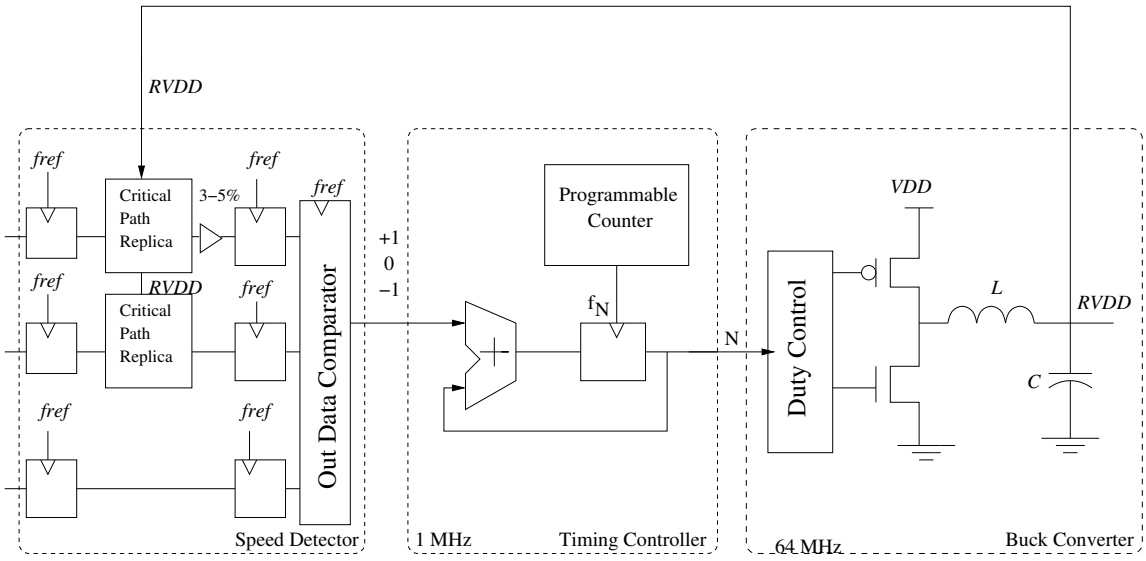


Figure 4.10: Closed-loop DVS system using a critical path replica.

Both the open and the closed-loop DVS systems work efficiently as long as the critical path is unique. However, this requirement is difficult to establish in modern VLSI circuits. In fact, the critical path can change when supply voltage is scaled. One path can be critical at one supply voltage while another path can be critical at another voltage. Furthermore, at a fixed supply voltage, the critical path can change from die-to-die based on process and temperature variations.

In order to eliminate such safety margin, Ernst *et.al* [88] proposed the Razor approach based on a speculative-timing pipeline. In this approach, an extra latch, *shadow* latch, is introduced at the critical path flip-flops. The latch is triggered by a slower version of the main clock as shows in Figure 4.11. As supply voltage is scaled, the value latched

in the *Master* flip-flop can be different from that latched by the shadow latch triggering an *Error* signal. The error signal serves two purposes. First, the error is propagated to the control system to increase supply voltage. Second, the pipeline is flushed and the correct value, now held by the shadow latch is fed back to replace the erroneous value held by the master flip-flop. The additional shadow latches are introduced where sub-critical paths become critical at worst case voltage operation. If the number of sub-critical paths is limited, the overhead of the razor approach can be ignored. However, in order to guarantee a robust operation, system characterization at all conditions is required. This may require an increased number of razor latches. Therefore, the overhead of the error detection circuitry may increase and the error probability may also increase resulting in a reduced efficiency.

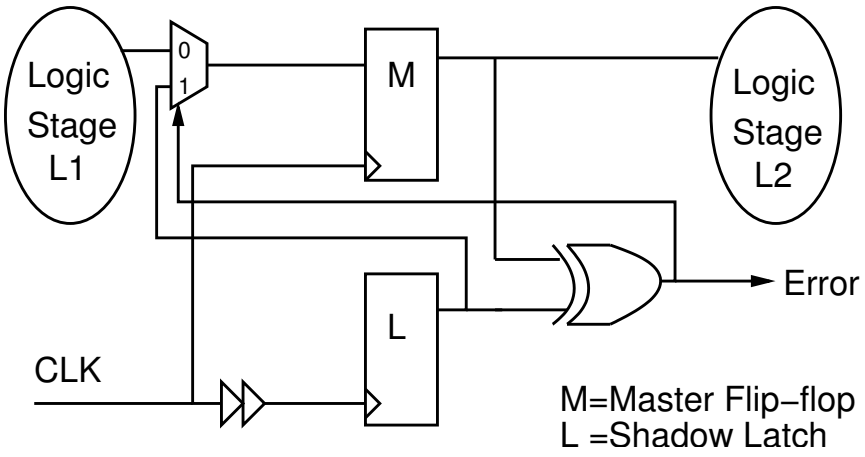


Figure 4.11: Razor approach to reduce the voltage margin dictated by worst case characterization.

Identifying the critical and sub-critical paths in a digital system is growing in complexity. Process variability and interconnect parasitics are negatively impacting the process of critical path identification. The ITRS technology roadmap predicts that delay due interconnect wires in the 65 nm technology node will be 8× that of the 180 nm technology.

Meanwhile, logic delay at 65 nm feature size is predicted to decrease by 2× compared to current technologies [9]. Moreover, the increasing usage of dual- and multi-threshold technologies to suppress leakage power adds further complications in the determination of a unique critical path for a system.

This chapter describes two different voltage scaling architectures. First, a hybrid approach between the open-loop and the closed-loop systems is presented. The proposed system saves energy by automatically identifying the process. In this case, the system selects frequency and voltage data points which correspond to the actual process split (silicon characteristics) not the worst case. Therefore, the proposed architecture minimizes the safety margin required by conventional open-loop systems to account for process variations. Once the voltage reaches the target dictated by the LUT, the system starts performance monitoring via a critical path replica to compensate for temperature variation. During the panic mode, the proposed DVS system switches to the LUT mode from the closed-loop mode and ramps up supply voltage to the maximum specified according to the actual process split. The proposed architecture is described in detail in Section 4.3. The analysis of the proposed DVS system and a comparison to the conventional system are given in Section 4.4.

The second proposed technique, described in section 4.5, is designed to reduce the growing complexity in identifying the actual critical path in the presence of the increasing interconnect delay. The proposed architecture follows the actual critical path delay at different process and interconnect parasitic conditions. The proposed technique uses an emulated critical path that has nearly the same voltage scaling behavior of the actual critical path at all conditions. Design details of the proposed system are described in section 4.5. Analysis and comparison of the critical path emulator to conventional systems is given in section 4.6.

4.3 Hybrid Dynamic Voltage Scaling Architecture

The proposed hybrid architecture works in two configurations, a LUT mode and a performance monitoring mode. During chip characterization, an automated mechanism to identify the process corner to which a particular chip belongs is performed. The LUT uses the process identification information extracted to determine the closest performance and voltage data points based on worst case temperature. When performance needs to be changed, the system starts in the LUT configuration. Once voltage is adjusted according to the LUT setting, the system switches to performance monitoring to fine tune supply voltage and to compensate for temperature variations.

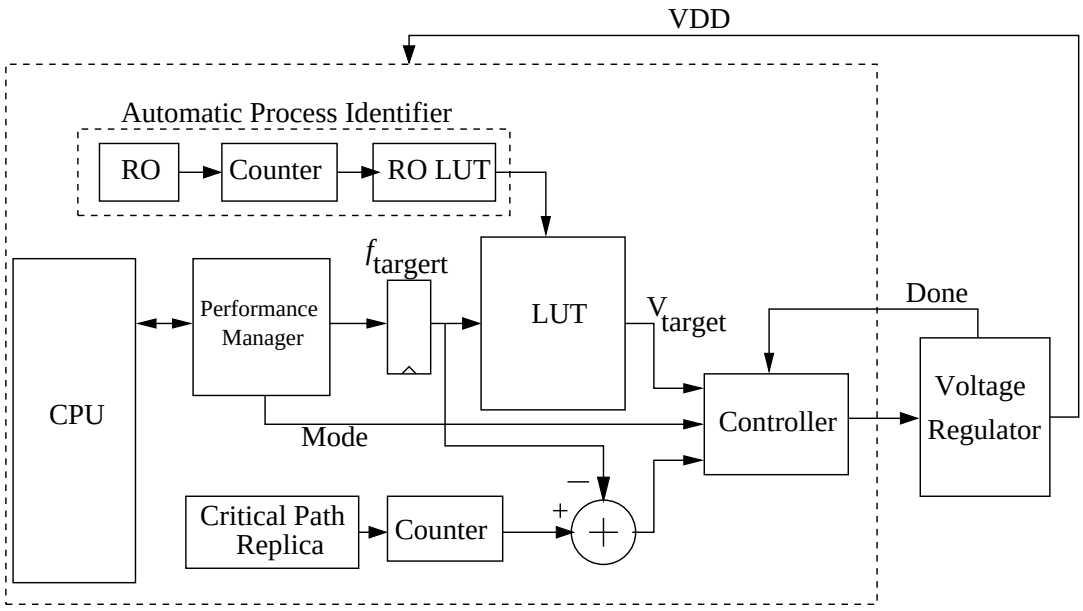


Figure 4.12: Architecture of the proposed hybrid DVS system

Silicon wafers are categorized based on their characteristics. There are three main corner cases, typical, slow, and fast. In between these cases, there are many wafers that lie within these corners and categorized as splits. The proposed system uses an automatic

process identifer to identify the process split during calibration as shown in Figure 4.12. Process variations and temperature are the main factors directly affecting performance. For example, a slow split at cold temperature can be faster than a typical split at hot temperature. Figure 4.13 shows the simulated frequency versus voltage characteristics for a critical path at different splits in 0.13μm CMOS process. Process identification is difficult to accomplish at high voltages. The distinction between the different frequency characteristics becomes fuzzy due to the larger impact of temperature on performance at high voltages. For example, at 1.5V, performance for the Fast process at hot temperature (125C) is almost the same for the Typical process corner at cold temperature (-40C). Therefore, it is necessary to fix temperature at a certain level in order to identify the process corner during calibration. Temperature adjustment adds extra time and cost to the calibration process.

The extra calibration time can be saved when the process corner is identified by measuring performance at a specific voltage for which performance is insensitive to temperature [89] [90]. When temperature changes, performance is affected by two main technology parameters, threshold voltage and channel mobility. Frequency at which a logic path can operate is given by

$$f = \frac{I_{\text{avg}}}{L_D C_{\text{avg}} V_{DD}} \tag{4.4}$$

where V_{DD} is the supply voltage and L_D is the logic depth. The average switching capacitance, C_{avg}, can be assumed independent of temperature. The average current, I_{avg}, is proportional to

$$I_{\text{avg}} \propto \mu(T)(V_{DD} - V_{TH}(T))^{\alpha} \tag{4.5}$$

where $\mu(T)$ is the channel mobility at temperature T, and $V_{TH}(T)$ is the threshold voltage at zero bias and at temperature T. Channel mobility and threshold voltage dependence

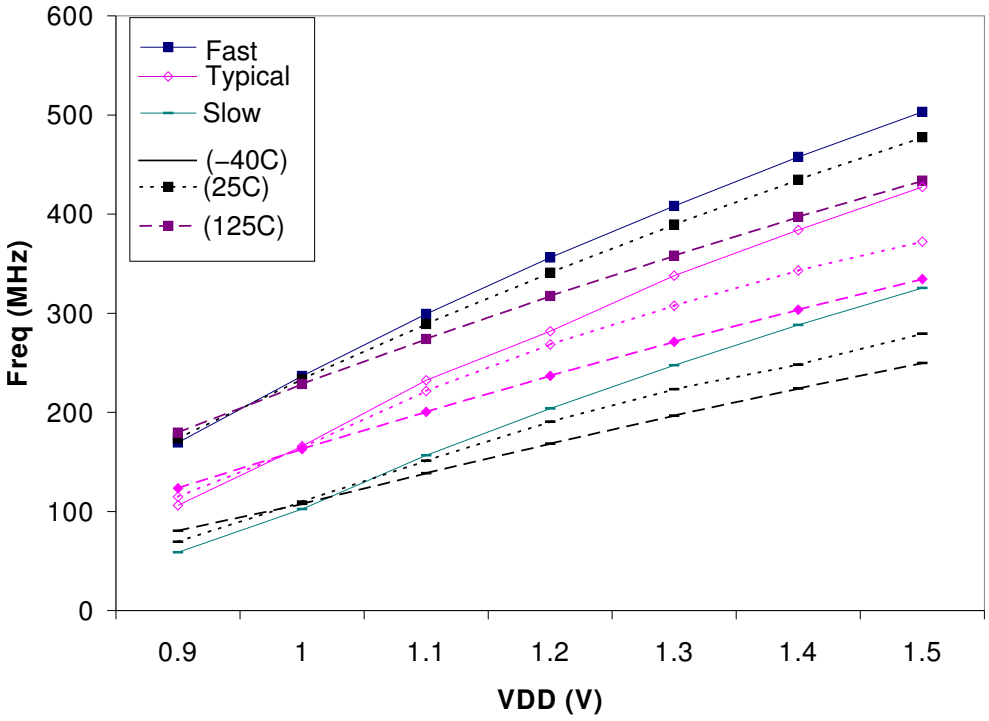


Figure 4.13: Critical Path frequency scaling with voltage for different process splits and different temperatures.

on temperature are given by [89]

$$\mu(T) = \mu(T_0)\left(\frac{T}{T_0}\right)^{-M} \tag{4.6}$$

and

$$V_{TH}(T) = V_{TH}(T_0) - \kappa(T - T_0) \tag{4.7}$$

where $T_0 = 300K$, M is the mobility temperature exponent, κ is the threshold voltage temperature coefficient. Typical values for M and κ are 1.5 and 1.8 mV/K respectively.

By lowering the supply voltage, the temperature effect on threshold voltage starts to cancel out the temperature effect on mobility. At a specific voltage, logic performance becomes insensitive to temperature. This voltage is independent of the type of logic im-

plementation. Therefore, process split can be identified since temperature effect has been canceled out and the only influence on performance is through process variations.

Figure 4.13 indicates that at approximately 1.0 V, the critical path frequency is insensitive to temperature across all process corners. A small ring oscillator (RO) is used to identify the process. A RO LUT is built using the characterized RO frequencies for different splits at the temperature insensitive voltage as shown in Table 4.1. The LUT entries are indexed by the RO frequency. During calibration, the voltage is set to the temperature insensitive value and the RO frequency is read. Process split can be identified using the RO frequency. If the frequency lies between two expected values then the split happens to be between two corners. The system must select the slower corner. For example, if the split lies between fast and typical corners, the typical corner is selected.

Table 4.1: RO LUT for Process Split Identification

Index	Process
f_{ro_1}	Process_1
f_{ro_2}	Process_2
...	

Compensation of process variations can be accomplished by performance characterization at worst case temperature. Three process splits (slow, typical, and fast) are considered. Characterization data is stored in a lookup table (LUT). The LUT format is indicated in Table 4.2. Total number of rows in the table is equal to the required number of target frequencies set by the software interface. The voltage settings corresponding to the identified process split are selected and all other voltage entries in the LUT are ignored. For

example, when a slow process is identified by the process identifier, V_{s1}, V_{s2}, ... etc., are selected. Based on the target frequency, the proper voltage is selected.

Table 4.2: LUT for Split Compensation

f	Slow	Typical	Fast
f_1	V_{s1}	V_{t1}	V_{f1}
f_2	V_{s2}	V_{t2}	V_{f2}
...	...	...	...

The proposed system works as follows: The automatic process identifier identifies the process during system calibration phase. The target voltage is set to the temperature insensitive value. The RO frequency indexes the different values stored in the RO LUT. Accordingly, the RO identifies the process to the main LUT. The target voltage is set according to the target frequency for the process identified . Target voltage is used by the voltage regulator to adjust supply voltage to reach the target.

Once the voltage settles at the specified target voltage, the system switches to the performance monitoring mode. The target frequency is compared to the frequency of a critical path replica for voltage fine tuning. A small voltage margin is added to compensate for any mismatch between the real critical path and its replica. The system switches back to the LUT mode when performance is to be increased or when a drastic temperature change occurs leading the system to enter the panic mode. The voltage is set to the peak voltage required by the split rather than the worst case split. The energy saving of the proposed hybrid system compared to the conventional system is analyzed in detail in the next section.

4.4 Analysis of the Hybrid DVS system

When the proposed system operates in the LUT mode, energy can be saved by setting supply voltage to the worst case temperature for the closest split. By contrast, conventional DVS systems set supply voltage based on worst case split (slow). Figure 4.14 shows the voltage distribution as a result of process variations (assuming a Gaussian distribution) at a fixed frequency. For the slow split, the voltage is maximum while for a fast split, voltage can be reduced to $V_{\min}$ while maintaining performance.

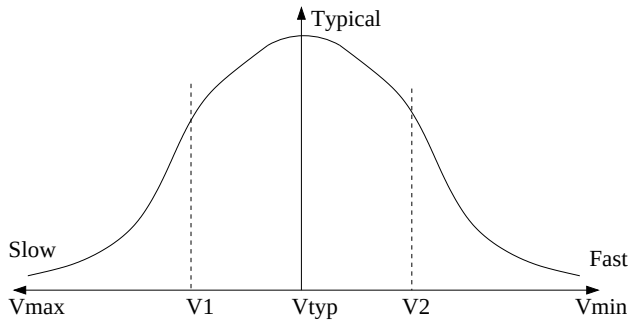


Figure 4.14: Voltage Distribution due to Process Variation at a fixed frequency.

From (4.1), energy has a quadratic dependence on voltage. As the number of process splits is increased, the expected savings are increased due to the finer granularity the system can offer. The effect of increasing the number of process splits on energy savings is analyzed by considering the 3-σ process distribution. The LUT contains information about three different splits, slow, typical, and fast. The cumulative probability density function (CDF) of having the voltage between fast and typical voltage is 50% and having the voltage at fast conditions is only 1%. Therefore, 50% of the parts can save energy by reducing supply voltage from $V_{\max}$ to V_{typ}, inidicated in Figure 4.14, while only 1% of the parts would benefit from reducing voltage from $V_{\max}$ to $V_{\min}$. Then, the energy saving

resulting from using voltage-frequency data of three different splits in the LUT is given by

$$\mathrm{E}_{\mathrm{savings}} = 0.5\left[1-\left(\frac{V_{\mathrm{typ}}}{V_{\mathrm{max}}}\right)^2\right] + 0.01\left[1-\left(\frac{V_{\mathrm{min}}}{V_{\mathrm{max}}}\right)^2\right] \tag{4.8}$$

Taking $V_{\mathrm{min}} = 1.0V$ and $V_{\mathrm{max}} = 1.5V$ for a frequency of 200 MHz as shown in Figure 4.13, energy saving is around 15%. If the number of splits stored in the LUT is increased, energy savings are increased. For example, assuming that 4 different splits, slow, -σ, +σ, and fast, are used in the LUT. CDF for -σ and +σ are 62.5% and 18.75% respectively. Energy saving becomes 20%.

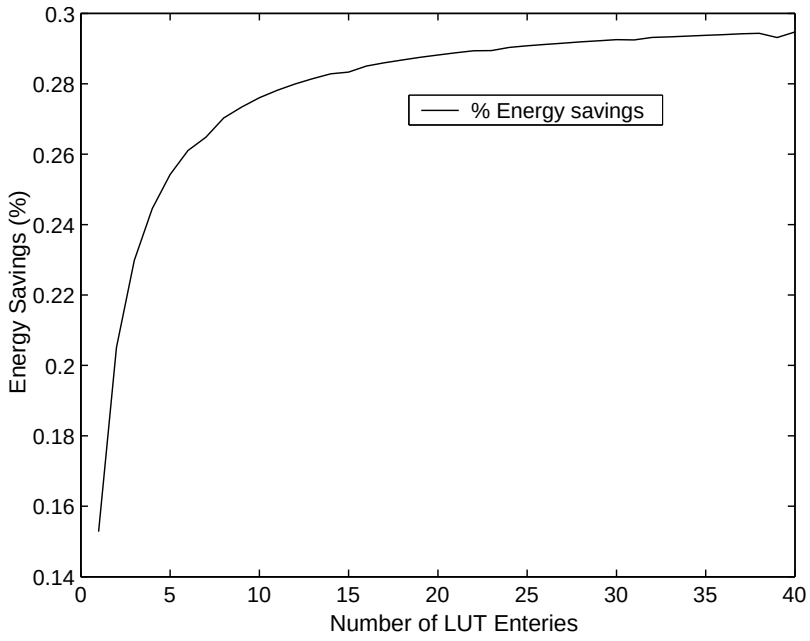


Figure 4.15: Energy Savings vs. number of entries in the LUT

Figure 4.15 shows the trend of energy savings when the number of entries (splits) of the LUT is increased. Energy saving is limited to approximately 29% using voltage and frequency data points for 40 different splits. However, using more than 10 different splits adds only 1% of savings. Furthermore, since the temperature insensitive voltage is not

exactly equal for all splits, increasing the number of entries in the LUT might result in a less accurate process identification.

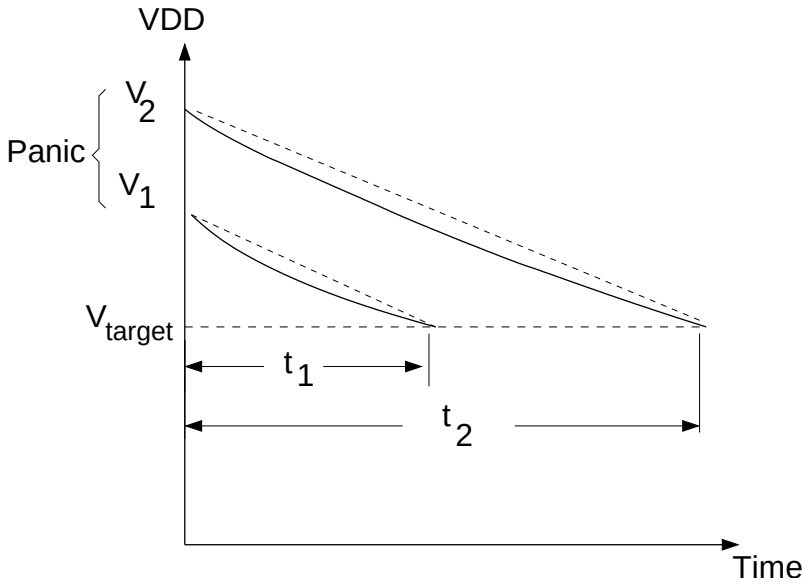


Figure 4.16: Voltage Waveform when going to panic mode

When the proposed system enters the panic mode and switches from the LUT mode to the performance monitoring mode, energy saving becomes highly application dependent. However, energy savings still can be achieved. Figure 4.16 shows the voltage waveform of both systems when entering the panic mode. Conventional systems ramp up supply voltage to the absolute worst case while the proposed system ramps to worst case for the actual process split. Therefore, supply voltage ramps up to a lower level than in the conventional case. Both systems eventually settle at the same voltage level but the conventional takes a longer time, $t_2 > t_1$, since it goes to V_2 while the proposed system goes to V_1, where $V_2 > V_1$. The discharge rate of both conventional and LUT case is assumed to be the same for the same load.

Energy saving during performance monitoring mode depend on the number of times

performance needs to be increased and how often the system enters the panic mode. The more the unnecessary performance glitches, the more energy savings the proposed system can achieve. This is highly application dependent and can vary from one system to the other. Test chip design and measurements are presented in Chapter 5.

4.5 Critical Path Emulator Architecture

As transistor dimensions are scaled every technology generation, the contribution of interconnect delay to the overall system delay increases. When several system paths have nearly the same delay while each one has different combination of logic and interconnect delay contribution, the process of selecting a unique critical path for the system becomes complicated. This phenomena can be illustrated using Figure 4.17. For a scaled supply voltage, delays of different paths implemented in the CMOS 0.13μm technology with different interconnect delay ratios are shown. The top set of delay plots represents delays for the slow process corner whereas the bottom set shows delays of the same paths at the fast process corner. For the slow process, the critical path, shown as a solid curve, is the reference path with an interconnect delay ratio of 50% at $V_{DD} = 1.3$ V. The dashed curves represent a number of potential critical paths with different interconnect delay ratios and delays close to the reference path delay. Since, logic delay scales faster with voltage than interconnect delay, delay scaling is different from one path to the other according to the contribution of logic and interconnects to the total delay of each path. When supply voltage is scaled based on performance needs, some potential critical paths become critical and their delays exceed that of the reference path. Once this happens, conventional systems which rely on characterizing or the monitoring the reference path alone tend to fail since supply voltage is not able to deliver the required performance. In order to accommodate

for the changing critical path, a delay margin has to be added to the reference path delay to guarantee that it remains the most critical at all supply voltages and for all interconnect parasitic variations.

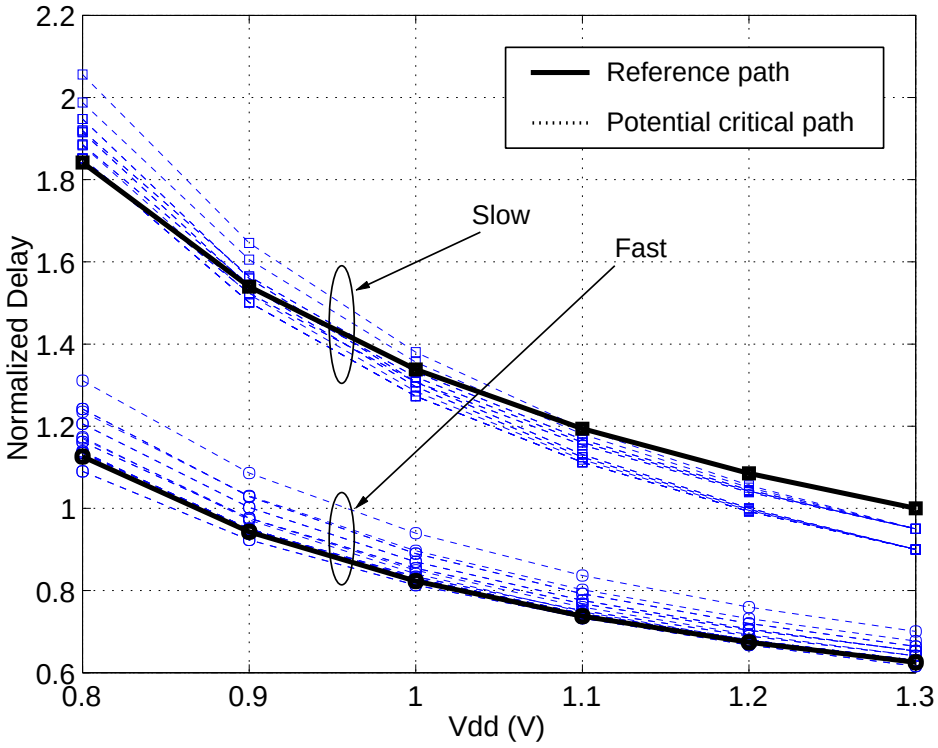


Figure 4.17: Reference path for the slow process changes due to the impact of interconnect delay and process variations.

Another factor that adds further complexity when designing a voltage scaling system is process variations and the impact of environmental conditions on performance. For example, at a certain voltage, a critical path at one process corner may not necessarily remain critical at another process corner or at a different temperature. Figure 4.17 shows this trend. The reference path at slow corner is no longer critical at the fast process (solid curve is moved down). As a result, conventional AVS systems require enough safety margin

to reliably scale supply voltage at any condition without causing a system failure. This margin is translated to a voltage overhead and the corresponding energy loss.

The reference path is Figure 4.17 has 50% interconnect delay at the slow process corner while a sub-critical path is due to majority logic. The delay margin is increased as voltage is scaled down as shown by the dotted line (sub-critical paths) are becoming critical. It is not sufficient to characterize and design the system based on worst case. One solution could be to use the logic path as the reference and add a small margin at the full scale voltage supply. This might not be sufficient if the logic process happens to be fast while interconnects remain at worst case as shown by the bottom set of delay plots in Figure 4.17. Therefore, enough delay margin is required to accommodate for variability in process and interconnect parasitics.

Both the conventional DVS and AVS systems tend to be less power efficient as interconnect delay contribution increases with technology scaling. Using either system requires a large voltage margin. Such margin reduces the power saved via supply scaling. Alternative to the conventional approach, a closer examination of the actual system behavior under different supply voltages and different operating conditions is necessary.

The objective of the proposed architecture is to emulate the critical path of a system at all conditions and at all supply voltages. Emulating the real critical path can be performed if the actual logic and interconnect speeds are measured on-chip. Consequently, the effect of process and interconnect variations on changing the critical path can be extracted. Based on the measured speeds, a critical path emulator is built using two delay lines. One delay line is composed of multiple stages of logic cells. This logic delay line is configured to have approximately the same delay as the logic delay portion of the actual critical path. Similarly, the other delay line is constructed using buffered interconnect wire segments with an overall delay approximately equal to the delay of interconnects in the real critical

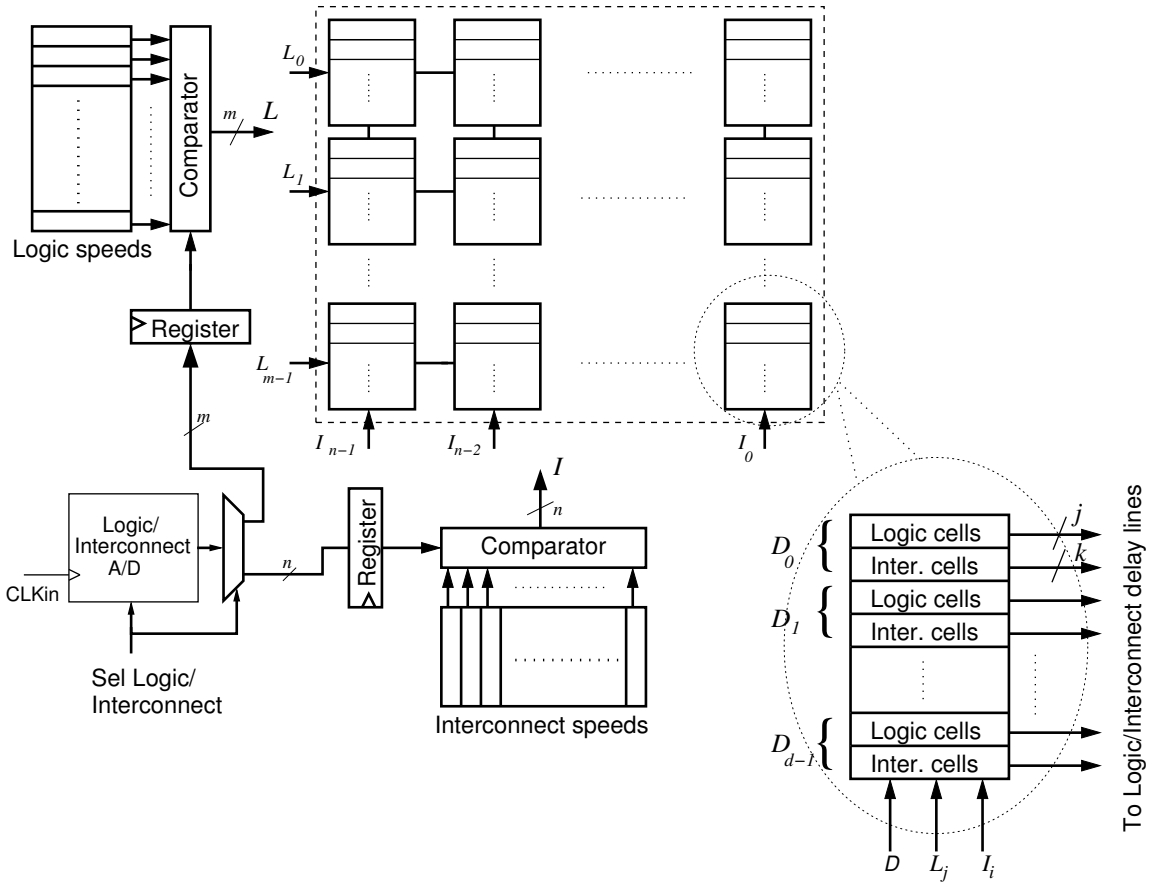


Figure 4.18: Critical Path Emulator Architecture.

path. The critical path emulator is monitored to form a closed-loop feedback system. By measuring the speed of the critical path emulator, which represents the actual speed of the system, supply voltage can be adapted to the actual environmental conditions.

In order to facilitate the subsequent discussion, a few terms used throughout this chapter are defined below.

- *Reference path*: the path that has the largest delay at worst case delay scenario (i.e worst case process, parasitics, and temperature.

- *Potential critical path*: a path which becomes critical at a certain voltage or at a certain process/interconnect corner.

- *Logic speed*: the actual on-chip logic speed. Logic speed is used to indicate how fast the actual process is compared to worst case.

- *Interconnect speed*: the actual on-chip interconnects speed. Interconnect speed is used to indicate the condition of the actual interconnect parasitics compared to worst case.

- *Interconnect delay ratio*: ratio of the delay caused by interconnect wires in a certain path to the total delay of that path.

- *Target delay*: the delay requirement specified by the system.

4.5.1 Proposed Architecture

The proposed architecture is shown in Figure 4.18. A logic and interconnect variations estimator is used to measure the effect of on-chip process and interconnect variations on logic and interconnect speeds relative to the worst case. This is represented by the logic/interconnect A/D described below. Logic and interconnect speed are represented by m and n-bits respectively. Based on the values of both vectors, a single LUT out of the LUT matrix is selected. For each target delay, the data stored in the selected LUT is used to construct two delay lines, one for logic and one for interconnects. The target delay, D, is determined by the system's software and is set by the d-bit vector. For each of the d-bit values, the number of logic delay cells represented by the vector j is used to construct the logic delay line, whereas the number of interconnect delay cells, k, is used to construct the interconnect delay line. The overall delay of the two delay lines (critical path emulator

delay) is approximately equal to that of the actual critical path. Furthermore, voltage scaling characteristics of the actual critical path and its emulator are nearly the same since their logic and interconnect delay compositions are approximately equivalent.

At system startup, on-chip process and interconnect variations are estimated by measuring logic and interconnect delays relative to the worst case. A low-power high-resolution A/D converter is used to determine logic speed [91] [86] as shown in Figure 4.19. FO4 inverters are used since their voltage scaling characteristics are nearly similar to most CMOS logic gates [92]. To eliminate the effect of temperature on the estimation process, supply voltage is adjusted such that performance is temperature independent [93]. At this voltage, temperature effect on delay is minimized leaving process and interconnect variations as the major factors affecting performance. As shown in Figure 4.19, the output of the counter represents the high-order bits of the logic speed vector whereas lower bits are represented by the output from the decoder. Similarly, interconnect speed is also measured using buffered interconnect segments. In order to avoid device mismatching between logic and interconnect buffers, the arrangement shown in Figure 4.19 is used. The two extra selectors are logic cells and should scale with voltage nearly the same way as the FO4 inverter [92].

The estimation process is performed in two steps. First, the selector is set to measure logic speed which is stored in a register. Then, the interconnect A/D converter is constructed by connecting the inverters through the long interconnect wire segments. To exclude inverter delays in the interconnect delay line, logic delay measured earlier is used to separate interconnect delay from buffer delay. Hence, interconnect parasitic variation is determined.

The output of the logic speed A/D is compared to the pre-stored logic speeds as shown in Figure 4.18. Based on this comparison, the appropriate selection line in the logic speed

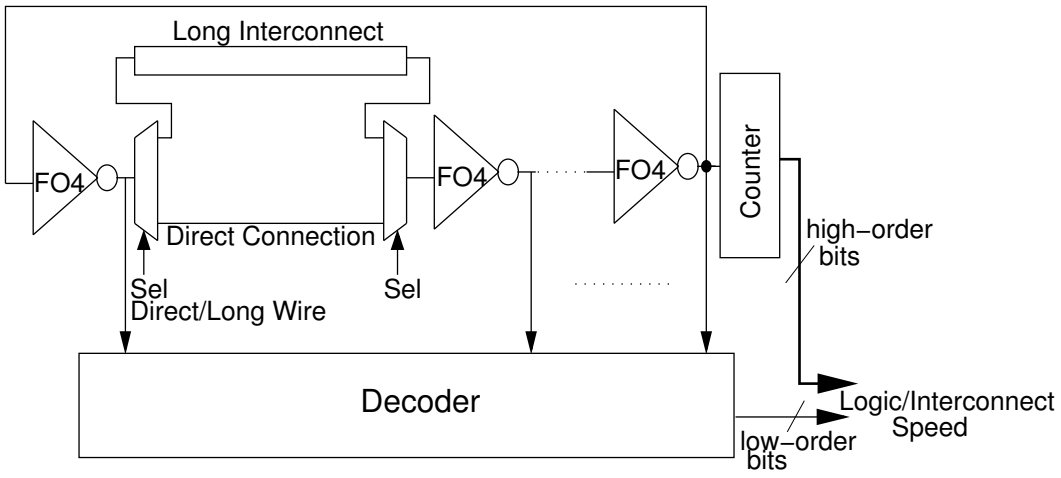


Figure 4.19: Logic and Interconnect low-power high-resolution A/D.

vector ($L = L_0L_1 \ ... \ L_{n-1}$) is activated to enable a row in the LUT matrix. Similarly, measured interconnect speed is used to activate the appropriate bit in the interconnect speed vector ($I = I_0I_1 \ ... \ I_{m-1}$) and the corresponding column is enabled. The architecture shown in Figure 4.18 depicts an m logic $\times$ n interconnect speed intervals and the corresponding LUTs. Using the estimated process and interconnect variations, the proper LUT is selected. The details of the LUT are shown in Figure 4.18. For each target delay, D, the corresponding number of logic cells, j, used to construct the logic delay line is selected. Similarly, the k-bit vector representing the number of interconnect delay cells is determined.

The delay line of the critical path emulator is constructed using the configuration shown in Figure 4.20. A similar approach was reported in [94]. The programmable delay line was used to emulate a single critical path and assumed that the critical path might not change. Adapting to process and parasitic effect on changing the critical path was not considered. The basic logic delay line used NAND gates with nominal and long channel devices [94].

In this work, the basic logic delay cell used to construct the logic delay line is the FO4 inverter. The interconnect delay cell is a long interconnect (e.g. minimum width and 1 mm long) with repeaters (FO4 inverters) at the driver and receiver ends of the wire. The logic delay line is programmed using the j-bit vector while the interconnect delay line uses the k-bit vector. The appropriate number of delay cells is selected using a multiplexer as shown in Figure 4.20. The critical path emulator is configured by connecting the output of the logic delay line to the input of the interconnect delay line.

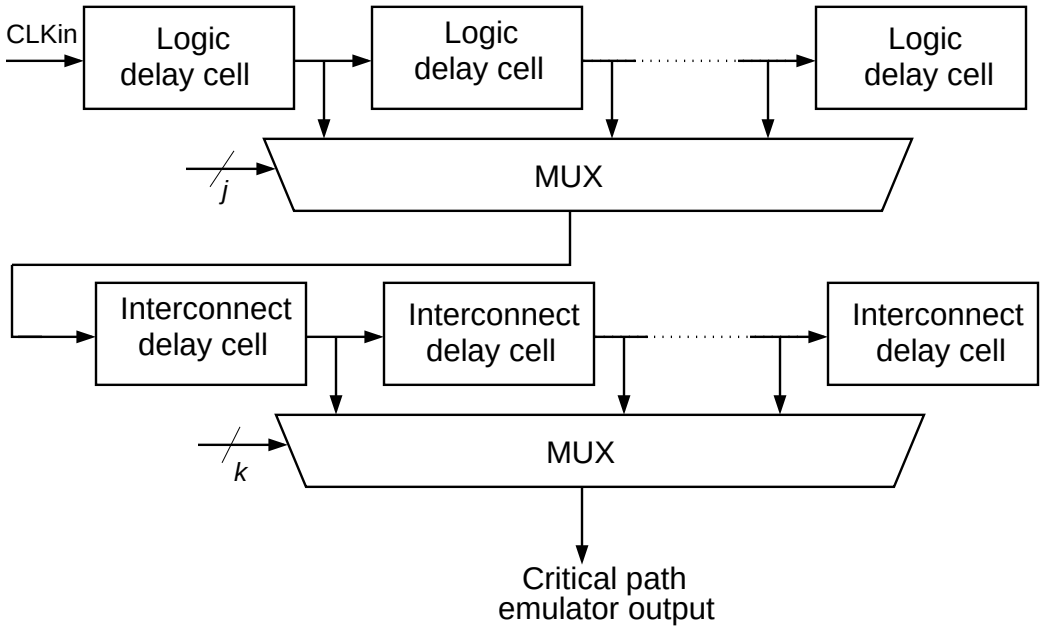


Figure 4.20: Implementation of logic and interconnect delay lines.

The number of logic and interconnect delay cells stored in the LUT matrix shown in Figure 4.18, can be determined through technology characterization. This process has to performed $m \times n$ times for the different process and interconnect splits. Instead of this lengthy and costly process, accurate modeling of both logic and interconnect delays is utilized. Using these models, the critical path delay at different conditions and different

target speeds can be predicted and stored in the LUTs.

4.5.2 Delay Modeling of Logic and Interconnects

As previously mentioned, a simple, yet accurate, model for delay of logic and interconnect delay lines can replace characterization in the development of the critical path emulator. In this work, the delay model for both logic and interconnects is based on previously published models [66]. Additionally, accurate modeling of the rising/falling input signals is used since the input ramp to one stage of the delay line is the output from the previous stage.

Traditionally, rise/fall time is often categorized into a fast and a slow input ramp. For our delay lines, since the input ramp to one stage of the delay line reaches full scale supply voltage (V_{DD}) before the output reaches the $V_{DD}/2$ point, the input ramp is considered fast. The output transition time, which is equal to the input rise/fall time to the next stage of the delay line, is defined in [66] and is given by

$$\begin{aligned} t_{TLH} = t_r &= \left(\frac{C_L V_{DD}}{0.7\ I_{Dp_{\max}}}\right) \frac{8v_{D0_p}^2(1+\lambda_p V_{DD})}{(4v_{D0_p}-1)(2+\lambda_p V_{DD})} \\ t_{THL} = t_f &= \left(\frac{C_L V_{DD}}{0.7\ I_{Dn_{\max}}}\right) \frac{8v_{D0_n}^2(1+\lambda_n V_{DD})}{(4v_{D0_n}-1)(2+\lambda_n V_{DD})} \end{aligned} \tag{4.9}$$

where C_L is the load capacitance, $I_{D_{\max}}$ is the maximum drain current at $V_{GS} = V_{DS} = V_{DD}$, v_{D0} is the drain saturation voltage at $V_{GS} = V_{DD}$ normalized by V_{DD}, and λ is the channel length modulation. The subscripts, p and n refer to the PMOS and NMOS parameters respectively.

Daga *et al.* [95] proposed an inverter delay model for fast input ramps based on the alpha-power model and the concept of inverter step response. The high-to-low and low-to-

high step response, t_{HLs} and t_{LHs} respectively, are given by

$$\begin{aligned} t_{HLs} &= \frac{C_L V_{DD}}{IDn_{\max}} \\ t_{LHs} &= \frac{C_L V_{DD}}{IDp_{\max}} \end{aligned} \tag{4.10}$$

where $IDn_{\max}$ and $IDp_{\max}$ are the inverter's NFET and PFET maximum drain current, respectively [95].

Using the rise/fall time given in (4.9), the delay time of a FO4 inverter delay for the fast input ramp case is given by

$$\begin{aligned} t_{HL} &= v_{TN}\frac{t_r}{2} + \left(1 + 2\frac{C_{GDP}}{C_L}\right) t_{HLs} \\ t_{LH} &= v_{TP}\frac{t_f}{2} + \left(1 + 2\frac{C_{GDN}}{C_L}\right) t_{LHs} \end{aligned} \tag{4.11}$$

where v_{TP}, v_{TN} are the zero-bias threshold voltage normalized to V_{DD} for the PMOS and NMOS respectively. C_{GDP} and G_{GDN} represent the input-to-output coupling capacitances for the PMOS and NMOS transistors respectively.

The velocity saturation index in (4.11) is considered to be unity. PMOS transistors usually have a velocity saturation index which is greater than NMOS transistors and greater than unity for current CMOS technologies. Generalizing (4.11) to include the non-unity velocity saturation index (α) results in

$$\begin{aligned} t_{HL} &= t_r\left[\frac{1}{2} - \frac{(1 - v_{TN})}{\alpha_n + 1}\right] + \left(1 + 2\frac{C_{GDP}}{C_L}\right) t_{HLs} \\ t_{LH} &= t_f\left[\frac{1}{2} - \frac{(1 - v_{TN})}{\alpha_n + 1}\right] + \left(1 + 2\frac{C_{GDN}}{C_L}\right) t_{LHs} \end{aligned} \tag{4.12}$$

HSPICE simulations are compared to (4.12) for a FO4 delay line implemented in 0.13μm CMOS technology. Figure 4.21 shows that maximum error between predicted delay model and simulations is 4-5%. This small margin is considered when designing the emulator.

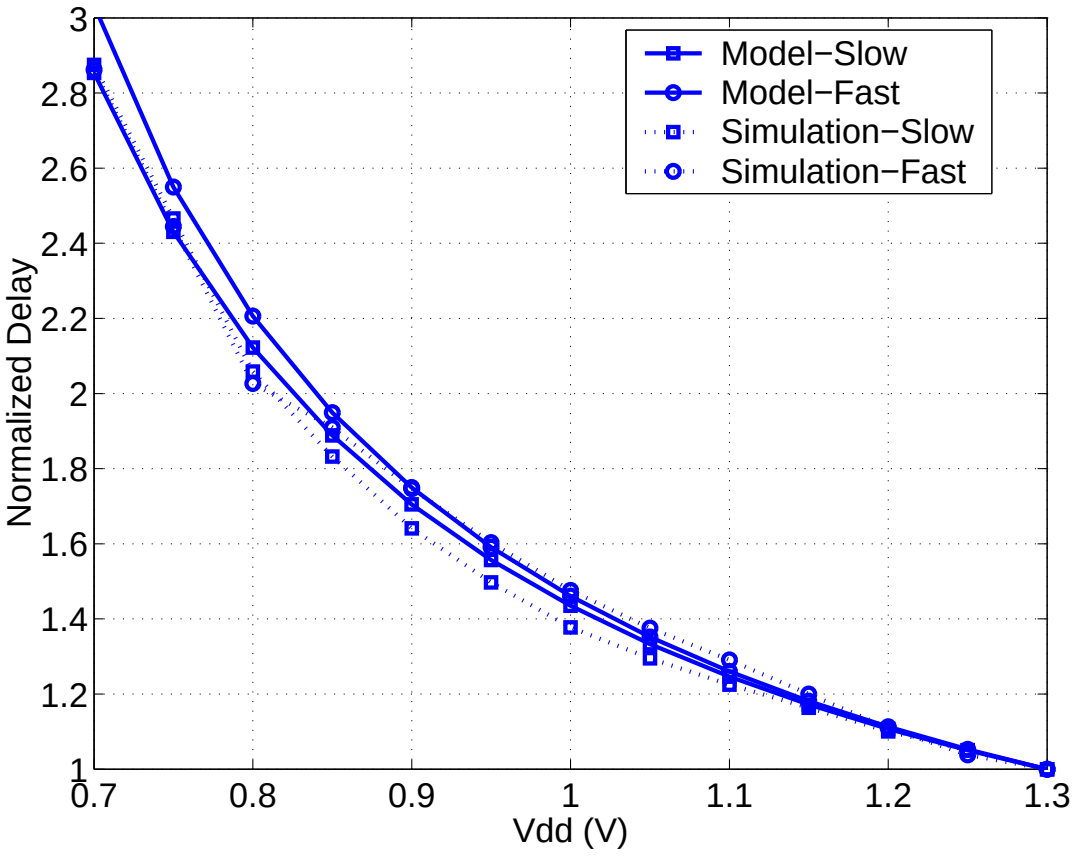


Figure 4.21: Logic delay vs. HSPICE simulations.

The FO4 inverter delay model described by (4.12) is used to model buffered interconnects. When buffers are inserted at optimal distances to minimize interconnect delay, overall delay of the buffered wire is found to be proportional to the square root of the buffer delay [96]. Therefore, the interconnect delay is related to the buffer delay, $t_{d_{\text{buf}}}$, by the following relation

$$t_{d_{\text{int}}} \propto \sqrt{RCt_{d_{\text{buf}}}} \propto \sqrt{RC}\sqrt{t_{HL}} \tag{4.13}$$

where R and C are the resistance and capacitance per unit length of the wire. Using (4.12)

and (4.13) to model voltage scaling behavior of both logic and interconnect delays takes into account process and interconnect variations. Therefore, the critical path at a certain process and certain parasitics corner can be predicted. Considering that worst case delay is the reference case, an algorithm is devised to determine such critical paths. This is described in detail below.

4.5.3 Algorithm

The algorithm used to generate the information stored in the LUTs for different process and interconnect corners is shown in Algorithm 1. Logic speed, L, and interconnect speed, I, are used as indicators of process and interconnect variations, respectively. In order to take process variations into consideration, the entire logic speed range is divided into increments with each increment is equal to L_{inc}. Similarly, the interconnect speed increment is I_{inc}.

The initial state of the algorithm is determined at worst case logic and interconnect corners. All logic and interconnect speeds are normalized to this reference case. In addition to the reference path, a set of potential critical paths is determined. Delay models given by (4.12) and (4.13) are used to predict the voltage scaling behavior of each path in the set. The ratio of interconnect delay to logic delay, I_{ratio}, for each path is also recorded. Based on the logic and interconnect unit delays at worst case in addition to I_{ratio} of each potential critical path, the number of logic, l, and interconnect, i, unit delays required to emulate each path are computed.

The next step is to determine which l and i to use in emulating the actual critical path for each target delay, D, specified by the system's software and for each specific logic speed, L, and interconnect speed, I. The delay of each path in the set of potential critical paths is computed using (4.12) and (4.13). Then, the path which has a delay equal to the target delay is selected. In this case, delay of all other paths should be less than the target

Algorithm 1 Critical path emulator

START:

$L = I = D = 1.0$

Find a set of potential critical paths

For each path in the set:

Compute (l, i)

for ($L = 1.0$: $L =$ Fast : $L = L$ - L_{inc}) **do**

for ($I = 1.0$: $I =$ Best : $I = I$ - I_{inc}) **do**

Find the reference path

Find the subsequent potential critical paths

while ($D <>$ Minimum) **do**

Find the critical path when ($t_d = D$):

Record its (l, i)

$(j, k) \leftarrow (l, i)$

$D =$ Next D

end while

end for

end for

delay. Once the critical path is selected, its (i, j) pair is stored as (j, k) and used for emulation. The same procedure is repeated for the next delay target. Once the generation of the critical path emulator at all target delays is finished, the data required for one LUT in the matrix shown in Figure 4.18 is determined. Each LUT is used to store the critical path emulator data for a specific logic and interconnect speed range. The information required for the entire LUT matrix can be determined by repeating the above for all logic and interconnect speed ranges. The resulting delay of the critical path emulator closely tracks that of the real critical path. More importantly, voltage scaling behavior is nearly the same for both the real critical path and its emulator.

4.6 Analysis of the Critical Path Emulator Architecture

The proposed architecture is designed in the CMOS 0.13μm technology. A reference path at worst case with a certain I_{ratio} is selected. The effect of interconnect delay on the selection of a unique critical path is illustrated through the examination of a set of paths which have delays close to the reference and lower I_{ratio} (more logic delay). Since potential critical path delays scale faster with voltage, a margin is required which is proportional to I_{ratio} of the reference path. The algorithm described earlier is applied to these paths using the CMOS 0.13μm technology parameters. Logic and interconnect speeds are divided into 10 ranges each. The critical path emulator information for the 10 logic splits and the 10 interconnect parasitic corners is extracted. Therefore, $m = n = 10$ in Figure 4.18, yielding a 100 different process and parasitic corners stored in 100 LUTs. In this design example, the number of bits used by the logic and interconnect delay multiplexers is equal to 5 (e.g. $j = k = 5$). Considering 4 target delays, approximately 4-Kbits of ROM are required to

form all the LUTs.

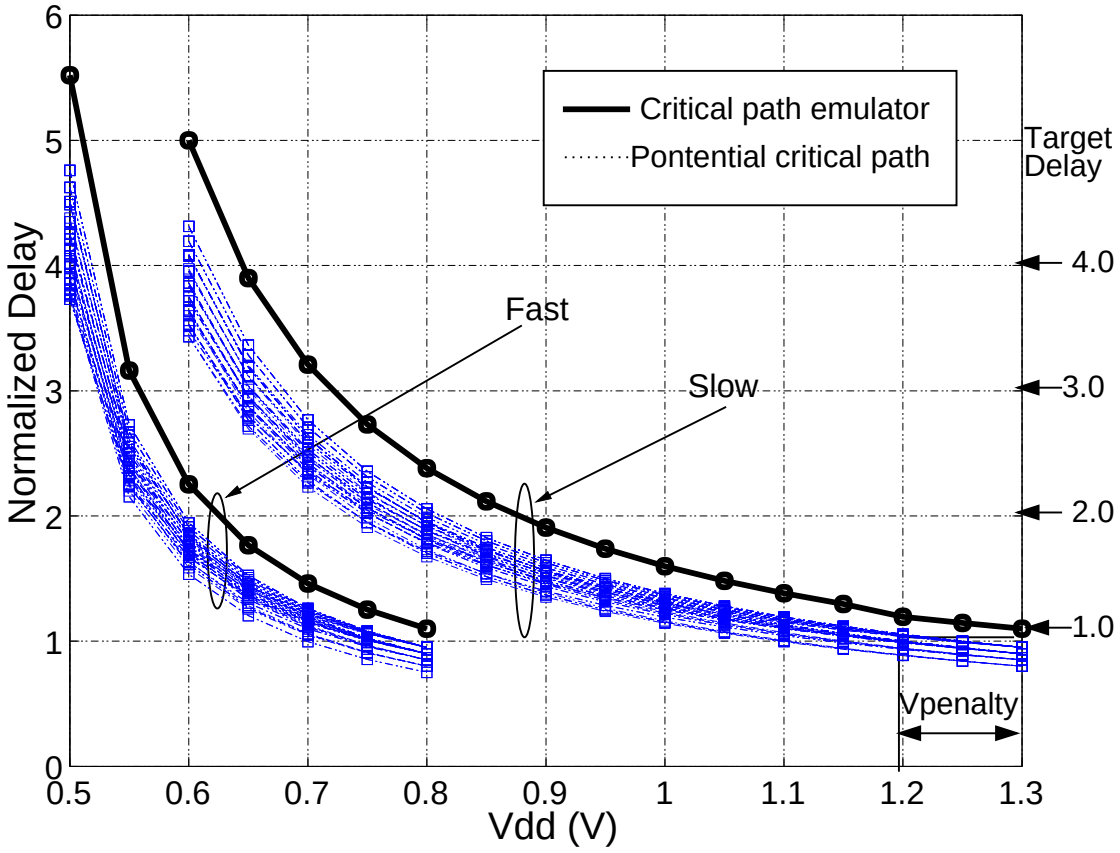


Figure 4.22: Delay of the critical path emulator exceeds delays of all other paths for the entire voltage range at both slow and fast process corners.

Figure 4.22 shows delays of the potential critical paths at the slow and fast corners and the critical path emulator for each case resulting from applying Algorithm 1. The reference path delay has an interconnect delay ratio of 50%. For both process corners, the critical path emulator, shown as a solid curve, has a safety margin above all the other paths at all target delays. Target delays, shown on the right of Figure 4.22, are set externally by the system's software.

The proposed critical path emulator architecture closely tracks the actual critical path at any given target delay. Therefore, the large delay margin required to account for worst case conditions can be saved. This delay margin is translated to a voltage overhead resulting in an extra dynamic energy dissipation which is given by

$$\text{Energy Loss} = 1 - (V_{\text{actual}}/V_{\text{worst}})^2 \tag{4.14}$$

where V_{worst} and V_{actual} are the supply voltages required to achieve the target delay with and without using a delay margin respectively.

When logic and interconnect speed intervals are taken to be equal to L_{inc} and I_{inc} respectively, the error range in determining the actual silicon condition becomes $\pm L_{\text{inc}}/2$ and $\pm I_{\text{inc}}/2$ for logic and interconnect, respectively. Assuming that $L_{\text{inc}} = I_{\text{inc}} = 10\%$, the maximum absolute error becomes 10% which is directly translated into a delay margin. In addition, a delay margin of 5% is added to compensate for model mismatch. Hence, the maximum delay margin required by the proposed system is 15%. From Figure 4.22, this delay margin corresponds to a voltage overhead of approximately 115 mV. Using (4.14), the maximum energy loss of the proposed system is approximately 17%. This energy loss can be reduced by increasing the granularity of process and interconnect speed sampling. However, increasing the granularity entitles more LUTs and additional selection overhead that reduces the energy efficiency.

Conventionally, the reference path is selected at the slow process corner and worst interconnect parasitics. Therefore, conventional open-loop systems require a delay margin to compensate for two factors, process variations in addition to the difference between voltage scaling characteristics of logic and interconnects. Energy savings obtained by adapting to process variations reach 27% when considering a sigma-distribution and the 10 process split information used by the proposed architecture [93].

On the other hand, utilizing a closed-loop feedback mechanism enables the system to

compensate for process variations. Therefore, a replica of the critical path can be sufficient to emulate the actual delay if both the critical and potential critical path delays are mainly due to logic delay. However, as the interconnect delay ratio, I_{ratio}, increases, the delay margin required to accommodate for any sub-critical path formed of pure logic delay also increases. This is due to the fact that logic delay scales faster than interconnect delay. Figure 4.23 shows the delay margin required by conventional closed-loop systems due to the difference in voltage scaling characteristics of logic and interconnects. At a supply voltage of 1.3 V, the critical path is due to mainly interconnect delay. As supply voltage is scaled down, a majority logic path becomes critical at supply voltage of approximately 0.8 V. Therefore, the critical path selection process should be preformed at both ends of the scaled supply voltage range. Furthermore, a delay margin is required to maintain a unique critial path at all conditions. In Figure 4.23, a 50% and 43% margins are required by the majority interconnect and majority logic paths respectively to guarantee a fail-safe operation. A general formula for the delay margin required by conventional systems is derived below.

When the reference path delay is assumed to have a certain I_{ratio} and the potential critical path delay is totally due to logic, this delay margin can be obtained when noting that at the delay of the reference path plus the required margin should be equal to that of the pure logic path. This can be expressed using the following relation

$$\frac{[\text{Margin} + (1 - I_{\text{ratio}})] * [t_{dl}]_{\text{V=V}_{min}} + I_{\text{ratio}} * [t_{di}]_{\text{V=V}_{min}}}{\left(1.0 * [t_{dl}]_{\text{V=V}_{min}} + 0\right)} = 1 \tag{4.15}$$

where $[t_{dl}]_{\text{V=V}_{min}}$ and $[t_{di}]_{\text{V=V}_{min}}$ are logic and interconnect delays at the minimum supply voltage respectively. Consequently, the delay margin can be expressed in terms of I_{ratio} as

$$\text{Margin} = I_{\text{ratio}} \left(1 - \left[\frac{t_{di}}{t_{dl}}\right]_{\text{V=V}_{min}}\right). \tag{4.16}$$

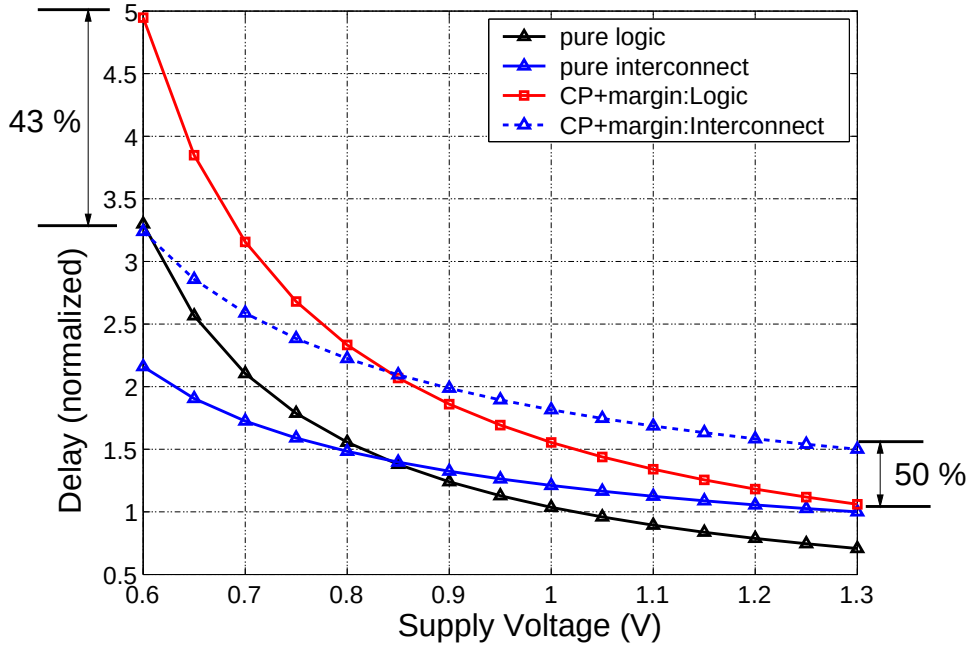


Figure 4.23: Delay margin required by conventional systems to compensate for the difference in voltage scaling behavior of logic and interconnects.

Using (4.16), the delay margin required by the conventional closed-loop system at the slow process corner due to the effect of interconnect delay on performance is shown in Figure 4.24. No delay margin is required when the reference critical path is due to pure logic delay. Meanwhile, the delay margin increases as the ratio of interconnect delay increases. The worst case parasitics require more delay margin than best case parasitics. For example, when the reference path delay is mainly due to interconnect (100%), delay margin required is 90% and 70% for the worst and best case parasitics respectively.

Based on (4.16), (4.14) is used to compute the energy efficiency of the proposed architecture compared to both the conventional open-loop and closed-loop systems as shown in Figure 4.25. Since open-loop DVS systems are designed at worst case process and parasitic conditions, an energy loss of 27% is incurred by open-loop compared to closed-loop

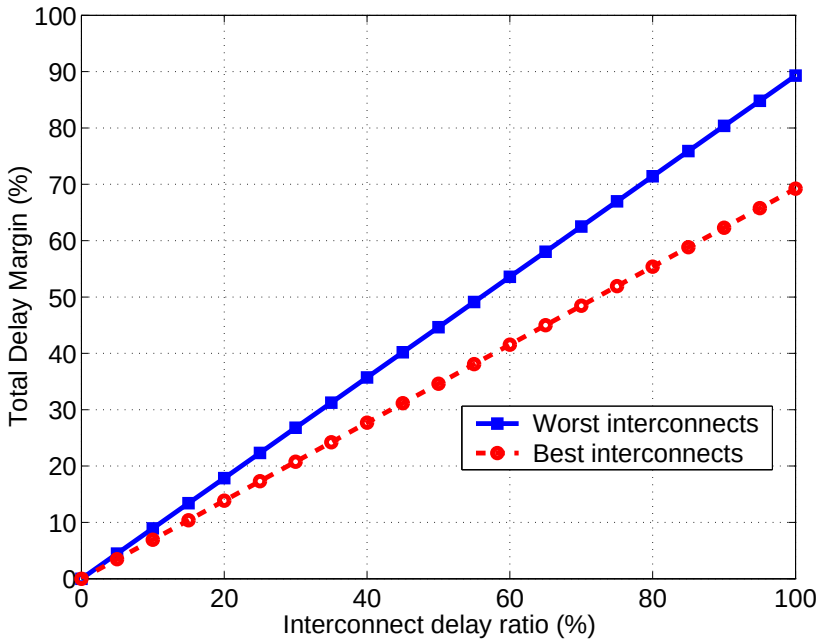


Figure 4.24: Delay margin required by conventional AVS systems as a function of interconnect delay ratio of the reference path.

systems. Therefore, the proposed system is up to 43% more energy efficient compared to conventional open-loop systems. Meanwhile, only the delay margin given by (4.16) is required by conventional closed-loop systems since process variations can be factored out. Therefore, energy efficiency of the proposed system compared to conventional closed-loop systems approaches 23%.

4.7 Summary

In order to meet the challenges of increased energy dissipation, a dynamic voltage scaling architecture was presented. The architecture regains the energy loss due to worst case characterization used in conventional systems. A lookup table based approach was utilized.

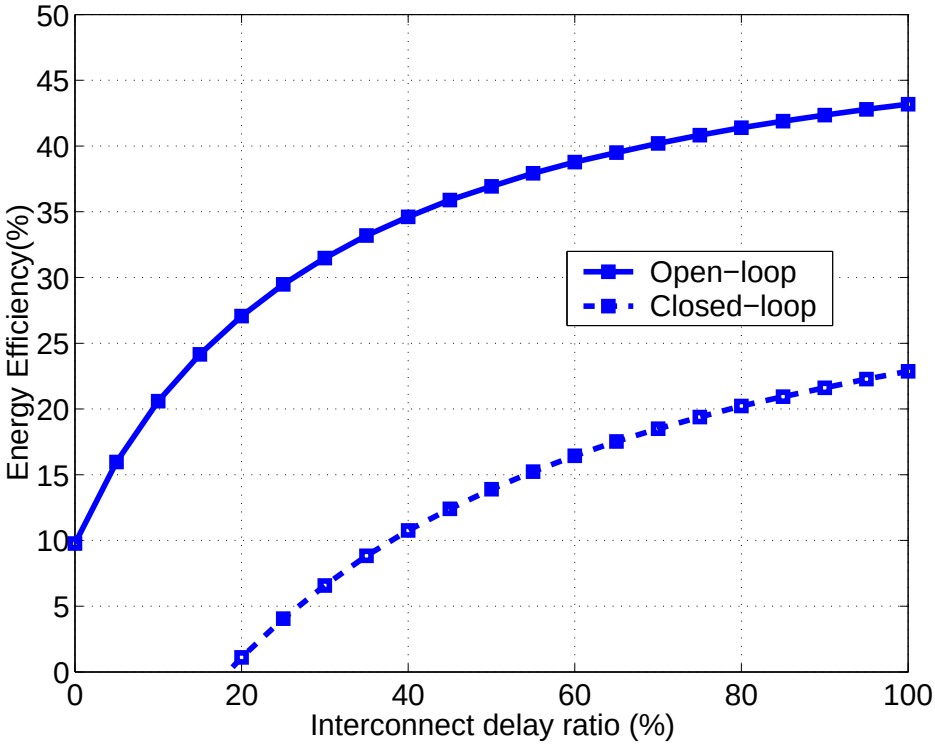


Figure 4.25: Energy efficiency of the proposed architecture compared to the conventional DVS and AVS systems as a function of interconnect delay ratio of the reference path.

The LUT holds characterization data for performance vs. voltage scaling of the critical path for three different process corners. Number of entries can be increased to gain more savings. Process is identified using an automated process identifier. The frequency-voltage entries corresponding to the identified split are used to set the voltage when performance is to be tuned. During panic mode, the proposed system ramps up supply voltage to worst case required by the actual process corner not the absolute worst case resulting in more energy efficient operation. Energy savings can be up to 29% compared to conventional DVS systems.

Conventional systems rely on identifying a unique critical path for all process and interconnect variations. However, with the increasing impact of process variations and interconnect delay on the determination of a unique critical path in modern VLSI systems, characterizing the critical path is becoming time and resources consuming. Therefore, conventional voltage scaling systems add a large delay and voltage margins to guarantee a robust operation even when the critical path changes under any circumstances. In order to recover this large margin required by conventional systems, an adaptive voltage scaling architecture with an on-chip critical path emulator was presented. The proposed system has the ability to adaptively track process and parasitic variations and environmental changes through a closed-loop feedback mechanism. Efficiency of the proposed architecture compared to conventional systems depends on the interconnect delay ratio of the reference path. The proposed architecture is up to 43% and 23% more energy efficient compared to open-loop and closed-loop systems, respectively.

Chapter 5

DVS System Experimental Results

This chapter demonstrates the implementation of the two dynamic voltage scaling architectures described earlier in Chapter 4. The first test chip demonstrates the open-loop DVS approach and the ability to adapt to process variations through an open-loop configuration. The critical path emulator (CPE) concept is demonstrated by the second test chip. The CPE system is shown to closely track the changing critical path at different conditions in a closed-loop configuration. Both test chips are implemented using the 0.18μm CMOS technology. Section 5.1 describes the design and implementation of the first test chip followed by post-layout and experimental results. The CPE test chip is described in section 5.2. Post-layout simulations are also shown.

5.1 Open-loop DVS Test Chip

The test chip architecture is composed of the open-loop DVS scheme described earlier in Chapter 4 connected to an off-chip programmable DC-DC converter. The open-loop DVS scheme is used to identify the actual process corner through an on-chip process identifier.

Once the process corner is identified, the corresponding voltage code used to program the DC-DC converter is issued. The programmable DC-DC converter is used to control the supply voltage according to performance requirements based on the actual process corner not the worst case.

The test chip architecture and test setup is shown in Figure 5.1. The two main components are the main look-up table (LUT) and the automatic process identifier. Considering 3 process corners, the LUT is divided into 3 different tables with each table corresponding to a specific process corner. Each LUT contains the frequency-voltage relationship for the specified process corner. By identifying the process corner and selecting the target frequency, the voltage codeword is selected.

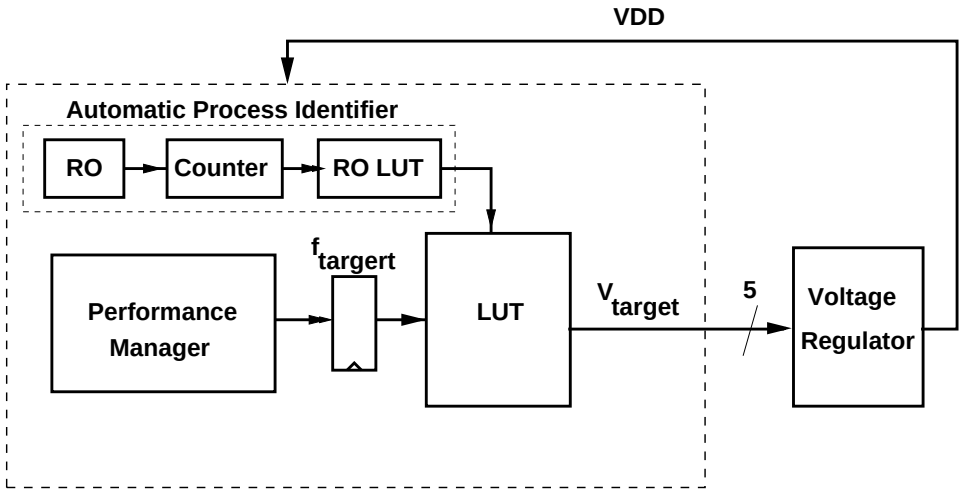


Figure 5.1: Architecture of the Open-loop Dynamic Voltage Scaling System Test Chip

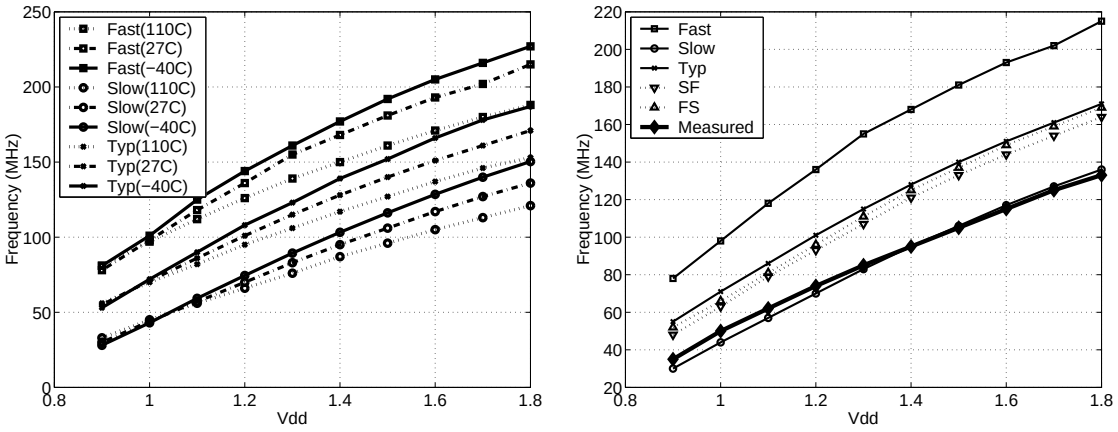
The process identifier is formed of a ring oscillator and a counter in addition to a small LUT. The ring oscillator is constructed using 123 stages. FO4 inverters are used in 122 of these stages. Stage number 123 uses a 2-input NOR gate to externally enable/disable the oscillation. The ring oscillator output is used as the clock input to the counter to

sample the actual on-chip speed. The reset input of the counter is activated every 1μsec. Therefore, the sampling rate of the counter is 1 MHz and the resultant count represents the ring oscillator frequency in MHz. A small LUT is used to store pre-characterization data of the ring oscillator frequencies at the different process corners. The measured frequency is comapred to the different frequencies stored in the LUT leading to the identification of the actual process corner. As a result, the corresponding system frequency versus voltage characteristics stored in the main LUT can be selected.

The frequency versus voltage behavior of the system underlaying dynamic voltage scaling should to be characterized based on silicon measurements. Characterization data for different process corners is then stored in the LUT. Ideally, the proposed open-loop DVS system can use a ROM-based LUT to hold this data. However, the characterization process requires performing measurements of the actual silicon behavior. This requires characterizing chips in several lots and several wafers in the same lot. To save time and resources, some test structures can be used to characterize process variations. In our test chip, the ring oscillator, used as part of the process identifer, was used for chip characterization. Accordingly, the LUT was formed using a serial shift register instead of using a ROM. Originally, the LUT is loaded with post-layout simulation data. Characterization data is then used to tune the information already loaded based on the actual silicon measurements. The actual silicon process parameters are extracted by measuring the ring oscillator frequency. At startup, characterization data is shifted serially into the LUT using a slow clock frequency. Once this data is completely stored in the LUT, the slow clock is disabled and normal operation begins.

The post-layout simulated frequency versus voltage relationship of the 123-stage ring oscillator is shown in Figure 5.2 (a). It is clear that at approximately 1.0 V, the ring oscillator frequency is temperature independent as discussed in detail in chapter 4. Therefore,

by adjusting the core supply voltage to 1.0 V, temperature effect can be excluded leaving process variations as the only factor affecting performance.



(a) Ring Oscillator post-layout simulation results for the Slow, Typical and Fast corners at -40C, 27C, and 110C.

(b) Measured and post-layout Ring Oscillator simulation results.

Figure 5.2: Post-layout and measured results for the Ring oscillator used in the Process Identifier.

The measured ring oscillator frequency is shown in Figure 5.2 (b). Since the frequency at a supply voltage of 1.0 V is approximately 45 MHz, it can be concluded that the process at hand belongs to the slow corner. Based on this decision, the corresponding frequency versus voltage LUT is selected.

The die photo is shown in Figure 5.3. The layout dimensions are 230×220 μm^2. The Agilent 81205 digital tester is used to test the functionality of the open-loop DVS test chip. Figure 5.4 shows the captured digital waveforms for the voltage code used to program the off-chip DC-DC converter. The output code starts at "00000" and switches to the appropriate code when the process is identified. When the target frequency is changed, the corresponding code is looked up in the LUT and used to reprogram the DC-DC converter

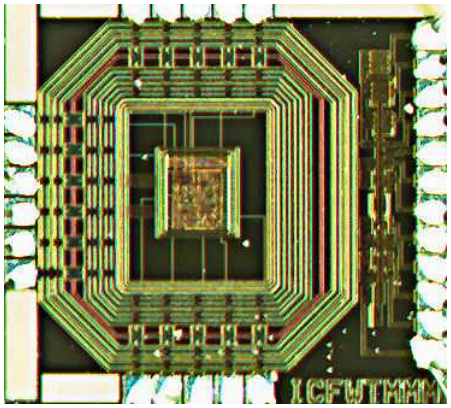


Figure 5.3: Die photo for the Open-loop DVS system implemented in the CMOS 0.18μm technology.

to achieve the required performance. Based on the selected frequency, the output code word starts at "10000" then changes to "11000" and finally to "11100".

The test chip is connected to the National Semiconductor's LM2633 DC-DC converter. The LM2633 is a 5-bit programmable DC-DC converter for mobile microprocessors. The 5-bit output voltage code word of the test chip is used to program the LM2633. Due to the way the LM2633 can be programmed, the generated voltage starts initially at zero and ramps up to the desired voltage. Therefore, the supply voltage generated is not fed back to the DVS system as shown in Figure 5.1. Figure 5.5 shows the captured waveform of the generated voltage when it ramps up from 0 V to 1.3 V according to performance requirements set by the test chip's output voltage code word.

5.2 Critical Path Emulator Test Chip

In order to validate the critical path emulator (CPE) architecture, a test chip is designed and implemented in the CMOS 0.18μm technology. The objective of this test chip is to

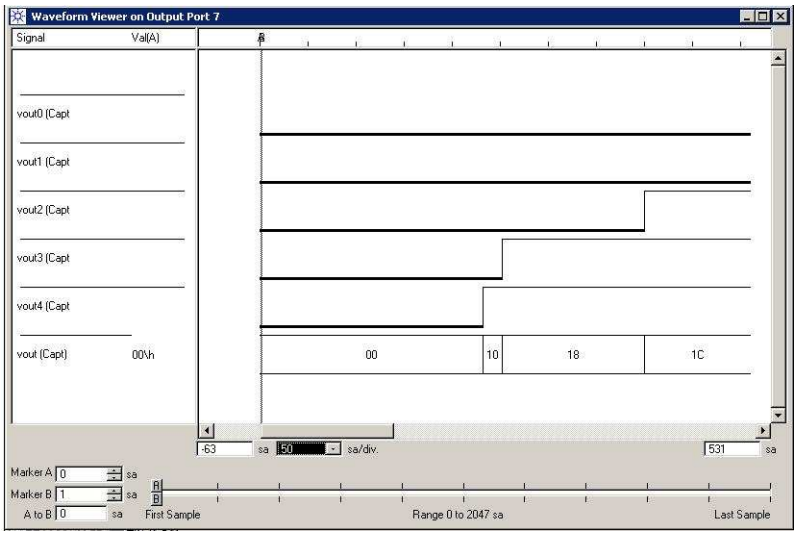


Figure 5.4: Captured output voltage codeword for different target frequencies using the Agilent 81205 Digital Tester. The voltage code word is initially set to "1000" then updated based on the performance command to "11000" then "11100".

validate the CPE architecture against two different delay paths, one is a majority interconnect and the other is majority logic path. The expected behavior is that the CPE output is able to closely track the most critical of the two paths independent of process or interconnect parasitic variations at all conditions and all target frequencies.

As described in detail in Chapter 4, the CPE architecture relies on estimating the on-chip logic and interconnect speeds and relate them to the actual process conditions. A logic ring oscillator is sufficient to measure the on-chip logic speed (with a limited accuracy). On-chip interconnect parasitics are probed by examining the delay of long interconnects. In order to maximize the capacitive effect of interconnects, the arrangement shown in Figure 5.6 is utilized. Since the top metal layer has the smallest sheet resistance, it is usually used for global signal routing such as clocks and long on-chip buses. In this test chip, the top metal, Metal 6 is used for interconnect delay estimation. The distance W is chosen

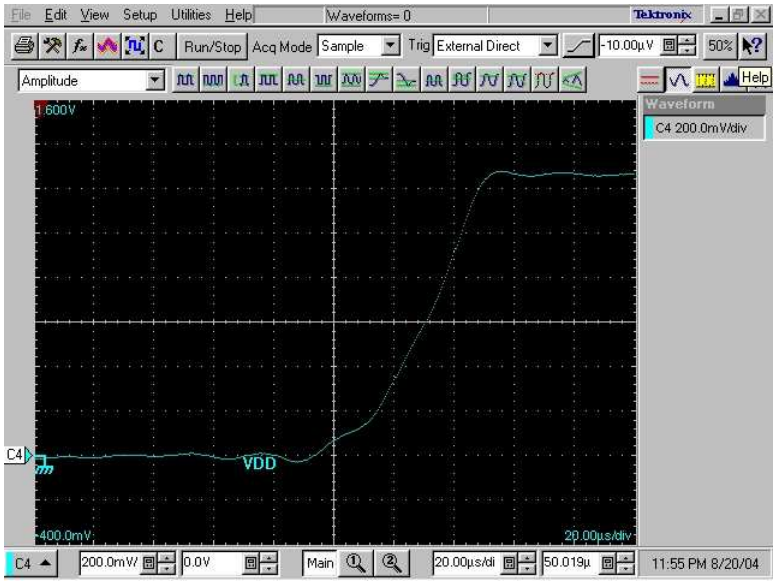


Figure 5.5: Measured output of the programmable DC-DC converter (LM2633) when it is ramping up from 0 V to 1.3 V.

to be the minimum distance allowed by the technology. In addition, the signal wire is sandwiched between two wires, one from each side, with minimum distance between them. The upper and the lower wires are acting as aggressors. The signal wire is set to switch in opposite direction with respect to the aggressors to maximize the coupling effect. From [7], the interconnect delay is calculated using the following equation

$$D_{int} = 0.38RC \tag{5.1}$$

where R and C are the resistance and capacitance of the interconnect wire respectively.

For the CMOS 0.18μm technology, the typical delay for a 1 mm wire of the top metal (M6) is estimated using (5.1) to be approximately 50 ps. Since the FO4 inverter delay for the typical process corner is approximately 100 ps, a 5 mm wire length is suitable to estimate the effect of interconnect delay with a reasonable accuracy. Meanwhile, the variability in sheet resistance and capacitance for metal M6 is $\pm 25\%$ and $\pm 20\%$ respectively.

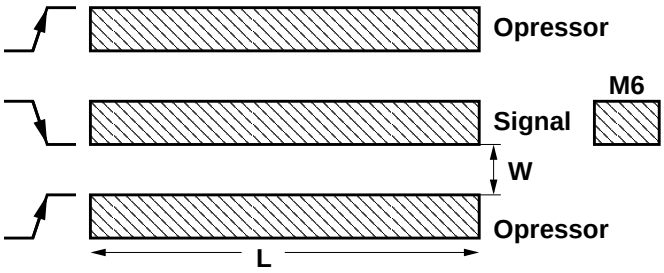


Figure 5.6: Worst Case coupling for interconnect capacitance.

Therefore, the overall variability in interconnect delay is ±45%. This range is divided into 5 different corners. Similarly, process variation of approximately ±35% is divided into 5 different regions.

The schematic of the CPE test chip is shown in Figure 5.7. The CPE system consists of a process/interconnect speed estimator, a LUT to store the information required to program the CPE delay line, and a programmable delay line. The logic/interconnect estimator uses a configurable ring oscillator to probe on-chip process and interconnect parasitic conditions. A small LUT is used to store pre-characterization data of the ring oscillator frequencies at the different process/interconnect corners. Similar to the open-loop DVS chip described in the previous section, the ring oscillator is configured to measure logic speed by connecting a loop of FO4 inverters and measuring the resulting frequency. The same ring oscillator is used to estimate interconnect parasitics by connecting a loop of buffered interconnect segments. Using a multiplexer at the output of each inverter, the FO4 ring oscillator is reconfigured by connecting each inverter to the next via a long interconnect instead of a direct short connection as previously shown in Figure 4.19. In order to maximize the capacitive coupling effect, these wire segments are implemented in the way shown in Figure 5.6.

Similar to the open-loop DVS chip described earlier, shift registers are used to construct

the LUT instead of using a ROM. The LUT is loaded with post-layout simulation data and is fine-tuned using the actual silicon data obtained after measurements. Considering five logic and five interconnect parasitic corners, the LUT included in the logic/interconnect estimator is formed using ten registers to store logic and interconnect speed information required. A serial shift register is constructed using these registers of 9-bit wide each. A 9-bit counter is used to measure the logic and interconnect ring oscillator speed by counting the number of cycles every 1 μsec. The ring oscillator is first configured to measure logic speed. The frequency count is then compared the logic speed information stored in the LUT to determine the process split. Similarly, interconnect speed is identified by configuring the ring oscillator in the interconnect speed measurement mode.

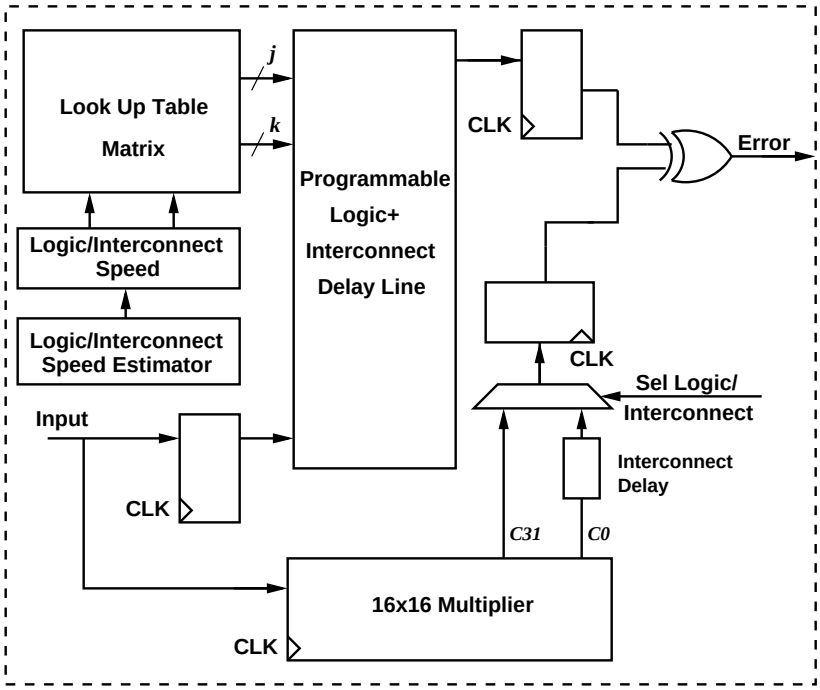


Figure 5.7: Test chip schematic for the CPE system.

The second component of the critical path emulator architecture is the LUT matrix.

Each LUT in the matrix corresponds to a specific logic and interconnect parasitic corner. The LUT is formed using serial shift registers to compensate for the lack of characterization silicon data. The number of cells required to construct the logic and interconnect portions of the CPE's programmable delay line is stored in each LUT. Based on the logic and interconnect corners identified, a specific LUT is enabled with the rest of the LUTs in the matrix are disabled. For a specific target delay, the required information required to program the delay line is extracted from the selected LUT. The delay line is configured in such a way that its total delay is approximately the same as the actual critical path delay. Moreover, the logic and interconnect delay portions of the delay line are approximately the same as that of the actual critical path. Therefore, voltage scaling characteristics of the actual critical path and its emulator are nearly equivalent.

A 16x16-bit unsigned multiplier is used as a test vehicle to verify the functionality of the CPE architecture. All the 32 inputs of the multiplier are tied together to form one input which is synchronized with the system clock (CLK). The same input is used as an input to the programmable delay line. This input toggles its value every clock cycle. Accordingly, the input to the multiplier switches from all zeros to all ones and back to all zeros and so on. Therefore, exercising the critical path in the multiplier is guaranteed through switching all inputs from zeros to ones. The frequency of the outputs is half that the of the system clock since the outputs switch from all zero to all ones once every clock cycle. Only two bits of the multiplier output are used in the verification process. The first product bit of the multiplier output, C_0, has the shortest logic delay. An interconnect delay line is added to this output bit to mimic a majority interconnect path as shown in Figure 5.7. The total delay of the original C_0 path and the added interconnect delay is approximately equal to the largest logic delay of the multiplier output, C_{31}. The second output to be emulated using the CPE architecture is the last product bit, C_{31}. This path

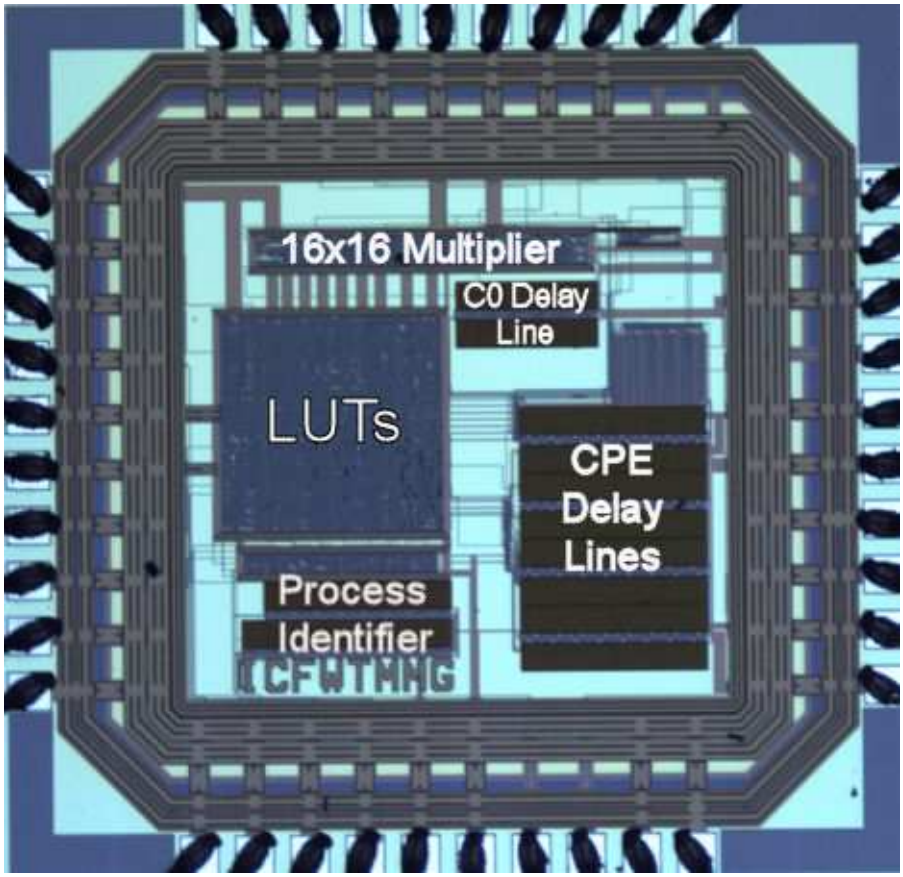


Figure 5.8: Die photo for the Critical Path Emulator system implemented in the CMOS 0.18μm technology.

is a majority logic delay path and has a different voltage scaling behavior compared to the C_0 path. Both multiplier outputs and the CPE output are latched using the system clock. When the CPE delay is longer than the system clock period, a wrong value is latched at the output CPE flip-flop. Using the 2-input XOR gate, the delay of the CPE system is compared to that of the multiplier and an Error signal is generated when the values stored in the corresponding flip-flops are different. This means that the CPE delay exceeded the required delay specification (including a 5% margin) and failed to emulate of the actual

delay of the multiplier. The die photo is shown in Figure 5.8. The layout dimensions are 1.6×1.6 mm^2. Excluding pads, the layout dimensions are 0.9×0.9 mm^2.

In order to verify the functionality of the CPE architecture, different supply voltage points are chosen and the ability of the CPE output to track the actual critical path is evaluated at each point. At each target supply voltage, the frequency of operation of the multiplier is determined by increasing the system clock gradually until the output flip-flops latch an incorrect value. Then the Error output signal is examined. If Error goes high then the CPE delay is longer than the actual critical path and the system fails to track. On the other hand, if Error remains low, the CPE delay is aligned with the actual critical path delay and the proposed system passes the test at this supply voltage point. Then, the same procedure is repeated again at a different target supply voltage.

Figure 5.9 shows the post-layout simulation results of the normalized delay of the CPE architecture. The logic and interconnect path delays of the two multiplier outputs are also shown. Results for both the Slow and the Fast corners are plotted. At the slow corner, the majority logic path, C_{31}, remains critical for most of the supply voltage range except for the range of 1.6 V to 1.8 V. During this short voltage range, the majority interconnect path, C_0, is critical. For the entire supply voltage range, the CPE output is shown to closely emulate the actual critical path with approximately 3% of additional margin. For the range from 1.6 V to 1.8 V, the CPE tracks the majority interconnect path with approximately the same margin. For the Fast corner, the CPE output emulates the majority interconnect path starting from a supply voltage of 1.8V down to approximately 1.0 V before switching to track the majority logic path. The maximum margin is 9% at 1.1 V.

On the test chip, the output of the CPE, the logic, and the interconnect paths before the flip-flops are observed off chip. The wire and mismatch effects between the three different paths are considered in the layout. The objective is to minimize the sources of error in the

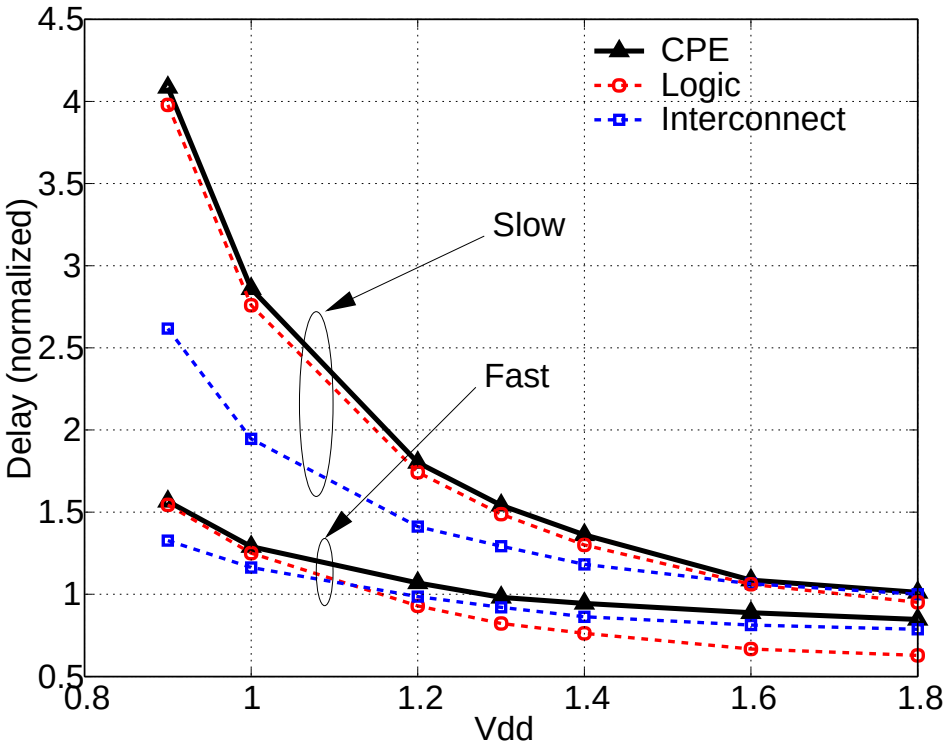


Figure 5.9: Post-layout simulation results for the CPE test chip showing the ability of the CPE delay line to track the actual delay of the multiplier for both the Slow and Fast process corners.

delay measurement. The measurement arrangement is shown in Figure 5.10. The input is toggled every clock cycle. Therefore, the outputs switch at half the clock frequency. For example, when the system clock is 50 MHz, the outputs switch at 25 MHz. The unlatched outputs are observed off-chip. The trace and pad delays are measured by directly routing the system clock to an output pad using an approximately the same wire length as the outputs. This delay is subtracted from the outputs delay measurements.

The phase error between the CPE and the multiplier outputs is measured. The magnitude of the phase error indicates how closely the CPE output tracks with the multiplier

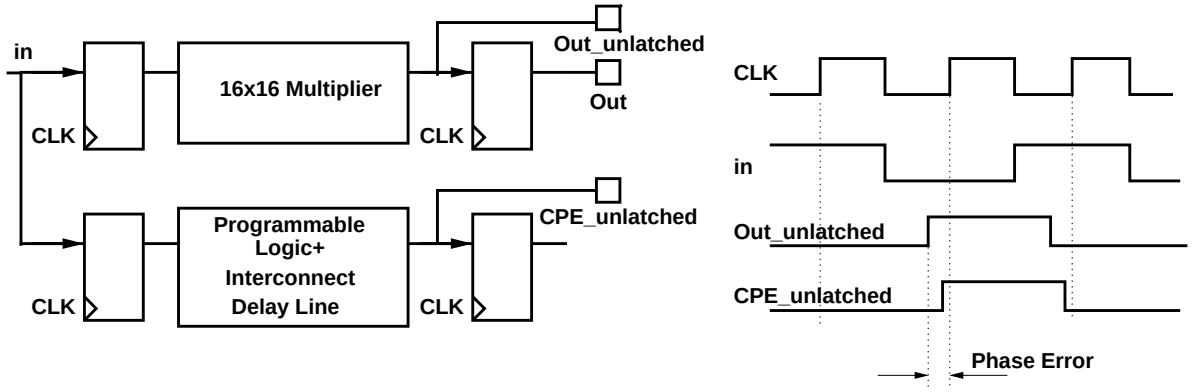


Figure 5.10: Delay measurement arrangement for the CPE chip.

output. The measured results for the CPE output and the multiplier output at 900 mV supply are shown in Figure 5.11. The phase error is approximately 57^o initially between the multiplier output and CPE output as indicated by the scope plots shown in Figure 5.11 (a). By programming the CPE delay lines to track with the actual critical path, the phase error is reduced down to approximately 15^o as shown in Figure 5.11 (b). Such a phase error can be almost eliminated. This is shown in Figure 5.11 (c) at a supply voltage of 1.1V and frequency of 40 MHz and is similarly done at the different target frequencies. The measured results for the CPE output and the multiplier output at 1.1V supply are shown in Figure 5.11.

The measured current consumption of the CPE architecture is shown in Figure 5.12. The current consumed by the CPE architecture is shown to scale well with the supply voltage. Therefore, the power dissipation overhead remains approximately constant across the entire supply voltage range which yields high efficiency at low supply voltages.

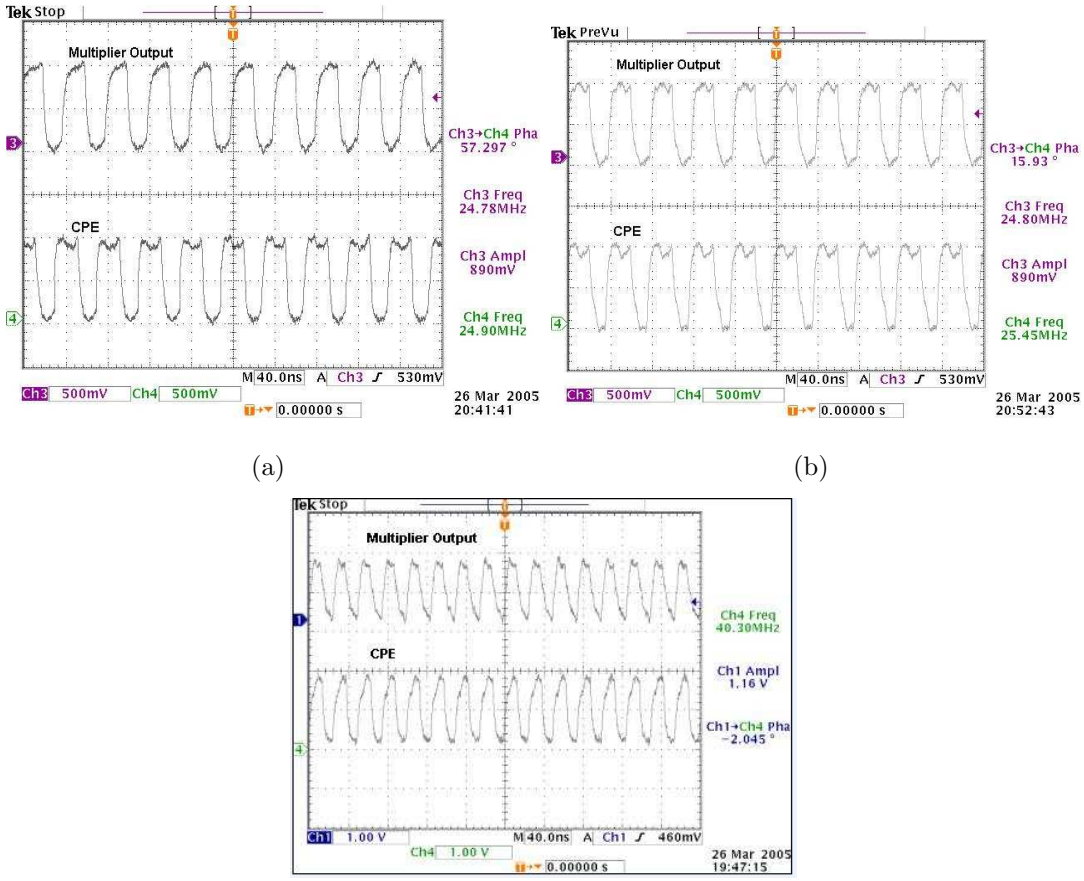


Figure 5.11: Measured results of the CPE test chip.

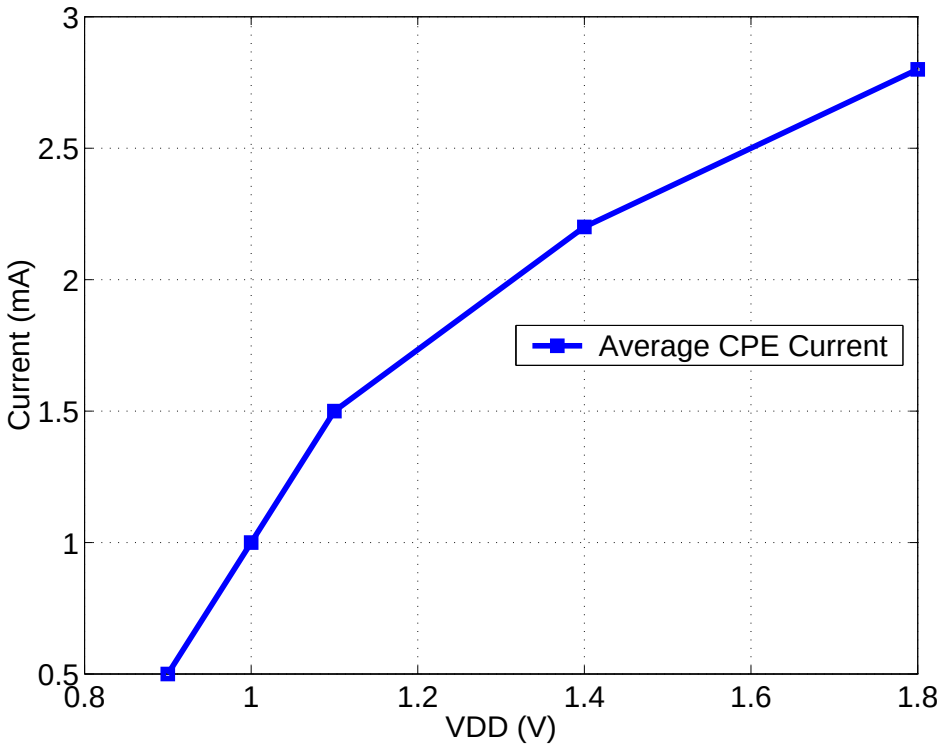


Figure 5.12: Measured current dissipation of the CPE architecture.

Chapter 6

Conclusions and Thesis Contributions

Aggressive technology scaling into the deep sub-micron regime has raised many challenges and obstacles in the road to follow Moore's law of doubling integration capacity every 2 to 3 years. The dramatic increase in transistor densities in modern VLSI systems has lead to a surge in power and energy dissipation to levels not seen before. This trend will continue to grow as more transistors are packed into the same area. The increased impact of process variations, interconnect parasitics, and reduced threshold voltage are some of the issues that are emerging as a result of device scaling. Therefore, power and energy-aware design flows are becoming popular in both ends of the design space, high-performance and portable applications. These design flows and techniques are developed on all design levels of abstraction. In this dissertation, power and energy reduction techniques on the architecture and circuit levels were presented.

An ultra low-voltage device structure was presented. By connecting the gate and the well of the PMOS transistor, the dynamic-threshold PMOS (DTPMOS) device structure

demonstrates a low threshold voltage during the ON state and a high threshold voltage during the OFF state. The new structure allows for sub-0.5 V operation in bulk CMOS technologies with a significant improvement in performance and a reasonable reduction in energy compared to conventional CMOS. However, above the 0.5 V, the efficiency of the DTPMOS scheme is reduced due to the increase in static power dissipation. Therefore, the DTPMOS scheme is mostly suitable for sub-0.5 V operation. The efficiency of the DTPMOS technique is demonstrated using a 16×16-bit multiplier designed in the 0.18 μm bulk CMOS technology. Simulation results show that the energy consumed by the multiplier is only 1.82 pJ at 0.48 V and a frequency of 2 MHz. Active leakage current during the ON state is controlled using a dynamic power management technique. Energy dissipation of the multiplier utilizing the dynamic power management technique is reduced to 1.68 pJ. The effect of the added well capacitance on performance of DTPMOS transistor should be carefully assessed early in the design stage. Such capacitance could reduce the competitive advantage of DTPMOS compared to conventional CMOS.

Noise immunity is another direct consequence of technology scaling. Supply voltage is scaled to maintain a sustainable electric field inside the device. Therefore, threshold voltage has to be scaled to meet performance requirements. As a result an exponential increase in leakage current is observed in deep sub-micron technologies. This mounting leakage current and threshold voltage reduction largely contribute to reduction in noise immunity especially for dynamic circuits. The dual-threshold technology has emerged as a viable solution for the decreased noise immunity of dynamic circuits. However, maintaining noise immunity becomes an issue and is usually associated with the cost of more energy dissipation. A circuit-level design technique is presented to address the trade-off between energy consumption and noise immunity. The evaluation transistors are split into two separate parts with the keeper transistor initially OFF during the precharge phase. Once evaluation

starts, the keeper turns ON if all inputs remain LOW otherwise the keeper remains OFF. This circuit technique offers a faster evaluation time and a lower energy dissipation due to two reasons. First, the dynamic node capacitance is reduced by half. Second, contention is virtually eliminated since the keeper transistor is OFF at the beginning of the evaluation period. The speed, power, and energy advantage of SD over conventional domino is further enhanced as supply voltage is scaled down. The less contention between the keeper and the evaluation network at the beginning of evaluation helps SD gates to switch faster compared to the conventional domino. Furthermore, reduced contention helps reducing power due to the less DC current flown. As a result, SD gates become more energy efficient as supply scales.

An optimal design methodology for dual-threshold domino circuit was also presented. The optimization methodology relies on analyzing the circuit behavior in both worst case leakage and worst case delay conditions. A delay model is devised with accuracy of 6% of HSPICE. The delay model is extended to the split-domino circuit technique with 7% accuracy comapred to HSPICE. The devised model can be used to examine different design trade-offs and offers designers a better handle on the different design decisions. First, the model was used to analyze the impact of threshold voltage reduction on performance for a given noise constraint. Although reducing the threshold voltage of evaluation transistors leads to higher ON current and faster discharge of the dynamic node capacitance, leakage current exponentially increases. The increased leakage current requires larger keeper transistor and results in larger contention power and slower evaluation. It was shown that reducing the threshold voltage beyond a certain point leads to slower, rather than faster, evaluation. The optimal threshold voltage was obtained using the proposed methodology.

Supply voltage scaling is the most effective way to reduce power dissipation. Dynamic voltage scaling is used to scale supply voltage based on performance requirements. Con-

ventionally, a worst case lookup table of frequency versus voltage information is used to control a programmable DC-DC converter based on performance demand. In this dissertation, a modified lookup table dynamic voltage scaling architecture was presented. In this architecture, an on-chip process identifier is utilized to factor out process variations. Based on the identified process corner, the corresponding frequency versus supply voltage lookup table is selected. The frequency-voltage entries are used to control the DC-DC converter when performance is to be tuned. The number of process splits and the corresponding lookup tables can be increased to achieve more savings. This architecture regains the energy loss due to worst case characterization utilized in conventional systems. During panic mode, the proposed system ramps up voltage to the worst case required by the actual process corner not the absolute worst case resulting in more energy efficient operation. Energy savings are up to 29% compared to conventional DVS systems. A test chip was designed in the CMOS 0.18μm technology to demonstrate the process identifier architecture. The architecture was connected to a 5-bit programmable DC-DC converter. It was shown that the process can be identified on chip and the voltage-frequency characteristics can be selected accordingly.

Process and temperature variations are conventionally compensated for by monitoring the on-chip speed of a ring oscillator or a critical path replica. However, the combined effect of process variations and interconnect parasitics on performance has lead to an increased complexity in identifying a unique critical path in deep sub-micron designs. The conventional approach is to add enough delay margin to the critical path replica to guarantee that it remains the most critical at all conditions even when the actual critical path changes. This delay margin is translated to a voltage margin which reduces the energy efficiency of conventional voltage scaling systems. In this dissertation, an adaptive voltage scaling architecture with an on-chip critical path emulator was presented. The

proposed system is capable of recovering the large margin required by conventional systems by adaptively tracking process and parasitic variations. The critical path emulator is constructed using a programmable delay line which has approximately the same delay as the actual critical path delay at any condition plus a small margin. This is accomplished by selecting a number of logic and wire delay cells corresponding to the actual logic and wire delay portions, respectively. Additionally, the proposed architecture forms a closed-loop feedback mechanism which adapts to temperature variations. Efficiency of the critical path emulator architecture compared to conventional systems is proportional to the ratio of interconnect delay to the total delay of the actual critical path. As the interconnect delay ratio increases, the probability that the critical path will change also increases and efficiency increases. The critical path emulator architecture is up to 43% and 23% more energy efficient compared to open-loop and closed-loop systems, respectively. The critical path emulator architecture was implemented in the CMOS 0.18μm technology. It was shown that the critical path emulator closely tracks the actual critical path at different conditions.

Appendix

Appendix A

During the first stage, V_{GS_n} is time-varying and can be expressed as

$$V_{GS_n} = \left(\frac{V_{DD}}{2}\right) + \left(\frac{V_{DD}}{2}\right).\left(\frac{t}{\tau}\right) \tag{6.1}$$

with $t_0 < t < t_1$ and $\tau = \tau_r/2$. Tylor series expansion is used to simplify the resulting α-power term $(V_{GS_n} - V_{TH0})^\alpha$. The result is a second order polynomial of the form

$$\begin{aligned}(V_{GS_n} - V_{TH0})^\alpha &= A_0 + A_1(t - 0.5\tau_r) + A_2(t - 0.5\tau_r)^2 \\ &= B_0 + B_1 t + B_2 t^2\end{aligned} \tag{6.2}$$

where

$$\begin{aligned}A_0 =&(0.75V_{DD} - V_{\mathrm{TH0}}),\\ A_1 =&A_0^{\alpha-1}(V_{DD}\alpha)/(2\tau_r),\\ A_2 =&V_{DD}^2.\alpha(\alpha-1)A_0^{\alpha-2}/(8\tau^2),\\ B_0 =&(A_0 + 0.5A_1\tau + 0.25A_2\tau^2),\\ B_1 =&(A_1 - A_2\tau), \text{ and}\\ B_2 =&A_2.\end{aligned}$$

Using (6.2), the pulldown saturation current can be expressed as

$$I_n = I_{0_n}(B_0 + B_1 t + B_2 t^2)(1 + \lambda V_D(t)) \tag{6.3}$$

Assuming that $V_G S$ for the keeper is fixed as explained in Section 3.7, the linear keeper current expression can be written as

$$\begin{aligned} I_k &= I_{\mathrm{D0}}[1 + \lambda_k(V_{DD} - V_D(t)] \left(A - \frac{V_{DD} - V_D(t)}{V_{DSAT_k}}\right) \\ &\cdot \left(\frac{V_{DD} - V_D(t)}{V_{DSAT_k}}\right) \\ &= I_{k_0}(K_0 + K_1 V_D(t) + K_2 V_D(t)^2 + K_3 V_D(t)^3) \end{aligned} \tag{6.4}$$

where

$$\begin{aligned} I_{k_0} &= \frac{I_{D0}}{V_{DSAT_k}^2} \left(\frac{V_{GS_k} - V_{\mathrm{TH0}_k}}{V_{DD} - V_{\mathrm{TH0}_k}}\right)^{\alpha_k}, \\ K_0 &= D_0 + D_0 \lambda_k V_{DD}, \\ K_1 &= D_1 - \lambda_k D_0 + \lambda_k V_{DD} D_1, \\ K_2 &= D_2 - \lambda_k D_1 + \lambda_k V_{DD} D_2, \\ K_3 &= -\lambda_k D_2, \\ D_0 &= A V_{DSAT_k} V_{DD} - V_{DD}^2, \\ D_1 &= V_{DD} - A V_{DSAT_k} + V_{DD}, \text{ and} \\ D_2 &= -1. \end{aligned}$$

.

From (6.3) and (6.4), the dynamic node discharge expression (3.14) is an *Ordinary Differential Equation* (ODE). Solving the resulting ODE in (3.14) can be performed numerically using Maple [97]. The resulting solution is the time V_{d1} at $t = t_1 = \tau$ where the input voltage reaches V_{DD}.

Appendix B

During the second stage in Fig. 3.17, the gate voltage of the pulldown transistor is fixed at V_{DD}. Therefore, the saturation current expression for the pulldown transistor in (3.14) can be written as

$$I_n = I_{n_1}(1 + \lambda V_D(t)) \tag{6.5}$$

where $I_{n_1} = I_{\mathrm{D0}} \left(\frac{V_G S - V_{\mathrm{TH0}}}{V_{DD} - V_{\mathrm{TH0}}} \right)^{\alpha}$

Using (6.4) and (6.5), the time t_2 in (3.15) can be expressed as

$$t_2 = C_D \int_{V_{d1}}^{V_{DSAT_k}} \frac{dV_D}{E_0 + E_1 V_D(t) + E_2 V_D(t)^2 + E_1 V_D(t)^3} \tag{6.6}$$

where

$$E_0 = I_{k_0} K_0 - I_{n_1},$$

$$E_1 = I_{k_0} K_1 - I_{n_1} \lambda,$$

$$E_2 = I_{k_0} K_2, \text{ and}$$

$$E_3 = I_{k_0} K_3.$$

. The time t_2 can be computed by solving (6.6) numerically.

In the thrid stage, both the keeper and the pulldown transistors are saturated. The saturated keeper current can be expressed similar to (6.5) as

$$I_k = I_{k_1}[1 + \lambda_p(V_{DD} - V_D(t)] \tag{6.7}$$

where $I_{k_1} = I_{\mathrm{D0}_k} \left(\frac{V_{GS_k} - V_{\mathrm{TH0}_k}}{V_{DD} - V_{\mathrm{TH0}_k}} \right)^{\alpha_k}$. The assumption that $V_{GS_k} = V_{DD}$ is considered to be still vaild. Hence, the time t_3 can computed using

$$t_3 = C_D \int_{V_{DSAT_k}}^{V_{DD}} \frac{dV_D}{G_0 + G_1 V_D(t)} \tag{6.8}$$

where

$$G_0 = I_{k_1} + \lambda_p I_{k_1} V_{DD} - I_{n_1}, \text{ and}$$
$$G_1 = -(\lambda_k I_{k_1} + \lambda I_{n_1}).$$

Bibliography

[1] J. Montanaro and *et.al.*, "A 160-MHz, 32-b, 0.5-W CMOS RISC Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1703–1714, November 1996.

[2] P. Gelsinger, "Gigascale Integration for Teraops Performance: Challenges, Opportunities, and New Frontiers," in *Proceedings of the 41st Design Automation Conference*, 2004, p. xxv.

[3] T. Sakurai, "Perspectives on Power-Aware Electronics," in *Proceedings of the International Solid-State Conference*, 2003, pp. 26–29.

[4] J. Rabaey and M. Pedram: Editors, *Low Power Design Methodologies*, Kluwer Academic Publishers, 1996.

[5] C. Belady, "Cooling and Power Considerations for Semiconductors Into the Next Century," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2001, pp. 100–106.

[6] J. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proceedings of the IEEE*, vol. 83, no. 4, pp. 619–635, April 1995.

[7] J. Rabaey, *Digital Integrated Circuits : A Design Perspective.*, Upper Saddle River, Prentice Hall, N.J. , 1996.

[8] A. Chatterjee, "An Investigation of the Impact of Technology Scaling on Power Wasted as Short-circuit Current in Low Voltage Static CMOS Circuits," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 1996, pp. 145–150.

[9] Semiconductor Industry Association, "ITRS, 2003 Ed. [http://www.public.itrs.net]," .

[10] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, April 1992.

[11] A. Keshavarzi, K. Roy, and C. Hawkins, "Intrinsic leakage in low power deep submicron CMOS ICs," in *Proceedings of Intl. Test Conference*, 1997, pp. 146–155.

[12] A. Chanadrakasan, W. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*, IEEE Press, 2000.

[13] A. Keshavarzi, S. Narenda, S. Borkar, C. Hawkins, K. Roy, and V. De, "Technology Scaling Behavior of Optimum Reverse Body Bias for Standby Leakage Power Reduction on CMOS ICs," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, March 1999, pp. 252–254.

[14] A. Chandrakasan and R. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," *Proceedings of IEEE*, vol. 83, no. 4, pp. 498–523, April 1995.

[15] N. Arora, *MOSFET Models for VLSI Circuit Simulation: Theory and Practice*, Springer-Verlag Wein New York, 1993.

[16] K. Yano *et. al.*, "A 3.8-ns CMOS 16×16 multiplier using complementary pass-transistor logic," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 4, pp. 388–393, April 1990.

[17] R. Zimmermann and W. Fichner, "Low-Power Logic Styles: CMOS versus Pass-Transistor Logic," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 7, pp. 1079–1090, July 1997.

[18] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. Ko, and C. Hu, "Dynamic Threshold-Voltage MOSFET (DTMOS) for Ultra-Low Voltage VLSI," *IEEE Transactions on Electron Devices*, vol. 44, no. 3, pp. 414–421, March 1997.

[19] L. Wei, R. Zhang, K. Roy, Z. Chen, and D. Janes, "Vertically Integrated SOI Circuits for Low-power and High-performance Applications," *IEEE Transactions on Very Large Scale Integration*, vol. 10, no. 3, pp. 351–362, June 2002.

[20] R. Zhang and K. Roy, "Low-power High-performance Double-gate Fully Depleted SOI Circuit Design," *IEEE Transactions on Electron Devices*, vol. 49, no. 5, pp. 852 – 862, May 2002.

[21] T. Kuroda *et.al.*, "A 0.9-V, 150-MHz, 10-mW, 4 mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1770–1779, November 1996.

[22] G. Wei and M. Horowitz, "A Fully Digital, Energy-Efficient Adaptive Power-Supply Regulator," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 4, pp. 520–528, April 1999.

[23] J. T. Kao and A. Chandrakasan, "Dual-Threshold Voltage Techniques for Low-Power Digital circuits," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 7, pp. 1009–1018, July 2000.

[24] J. T. Kao, S. Narenda, and A. Chandrakasan, "MTCMOS Hierarchal Sizing based on Mutual Exclusive Discharge Patterns," in *Proceedings of the Design Automation Conference*, 1998, pp. 495–500.

[25] S. Mutoh *et.al*, "1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 8, pp. 847–854, August 1995.

[26] H. Kawaguchi, I. Nose, and T. Sakurai, "A CMOS scheme for 0.5 V Supply Voltage with Pico-ampere Standby Current," in *Proceedings of the International Solid-State Conference*, 1998, pp. 192 – 193.

[27] H. Kawaguchi, K. Nose, and T. Sakurai, "A Super Cut-Off CMOS (SCCMOS) Scheme for 0.5-V Supply Voltage with Picoampere Stand-By Current," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 10, pp. 1498–1501, October 2000.

[28] Inukai *et.al.*, "Boosted Gate MOS (BGMOS): Device/Circuit Cooperation Scheme to achieve Leakage-free Giga-scale Integration," in *Proceedings of IEEE Custom Integrated Circuits Conference*, 2000, pp. 409 – 412.

[29] T. Inukai, T. Hiramot, and T. Sakurai, "Variable Threshold Voltage CMOS (VTCMOS) is Series Connected Circuits," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2001, pp. 201–206.

[30] A. Keshavarzi, S. Ma., S. Narendra, B. Bloechel, K. Mistry, S. Borkar, and V. De, "Effectiveness of Reverse Body Bias for Leakage Control in Scaled Dual Vt CMOS

ICs," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2001, pp. 207–212.

[31] Yang *et.al.*, "Experimental Exploration of Ultra-low Power CMOS Design Space using SOIAS Dynamic VT Control Technology," in *The IEEE International SOI Conference*, 1997, pp. 76–77.

[32] T. Sakurai, "Reducing power consumption of CMOS VLSI's through V_{dd} and V_{TH} Control," in *First International Symposium on Quality Electronic Design, IEEE*, 2000, pp. 417–423.

[33] S. Narendra, S. Borkar, V. De, D. Antoniads, and A. Chandrakasan, "Scaling of Stack Effect and its Application for Leakage Reduction," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2001, pp. 195–200.

[34] M. Johnson, D. Somasekhar, C. Yih, and K. Roy, "Leakage Control with Efficient use of Transistor Stacks in Single Threshold CMOS," *IEEE Transactions on Very Large Scale Integration*, vol. 10, no. 1, pp. 1–5, Feb. 2002.

[35] Gate Leakage Reduction for Scaled Devices using Transistor Stacking, "Leakage Control with Efficient use of Transistor Stacks in Single Threshold CMOS," *IEEE Transactions on Very Large Scale Integration*, vol. 11, no. 4, pp. 716–730, August 2003.

[36] K. Usami *et.al.*, "Automated Low-Power Techniques Exploiting Multiple Supply Voltages Applied to a Media Processor," in *Proceedings of IEEE Custom Integrated Circuits Conference*, 1997, pp. 131–134.

[37] D. Liu and C. Svensson, "Trading Speed for Low Power by Choice of Supply and Threshold Voltages," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 1, pp. 10–17, January 1993.

[38] S. Mahant-Shetti, P. Balsara, and Carl Lemonds, "High Performance Low Power Array Multiplier Using Temporal Tiling," *IEEE Transactions on Very Large Scale Integration*, vol. 7, no. 1, pp. 121–124, March 1999.

[39] C. Law, S. Rofail, and K. Yeo, "A Low-Power 16x16b Parallel Multiplier utilizing Pass-Transistor Logic," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 10, pp. 1395–1399, October 1999.

[40] T. Fuse, Y. Oowaki, M. Terauchi, S. Watanabe, M. Yoshimi, K. Ohuchi, and J. Matsunaga, "An Ultra Low Voltage SOI CMOS Pass-Gate Logic," *IEICE Transactions on Electronics*, vol. E80-C, no. 3, pp. 472–477, March 1997.

[41] M. Song and K. Asada, "Power Optimization for Data Compressors Based on a Window Detector in a 54×54 bit Multiplier," *IEICE Transaction on Electronics*, vol. E80-C, no. 7, pp. 1016–1024, July 1997.

[42] F. Mooler, N. Bisgaard, and J. Melanson, "Algorithm and Architecture of a 1V Low Power Hearing Instrument DSP," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED'99*, San Diago, CA, USA, 1999, pp. 7–11.

[43] H. Neuteboom, B. Kup, and M. Janssens, "A DSP-Based Hearing Instrument IC," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 11, pp. 1790–1806, November 1997.

[44] A. Sinha and A. Chandrakasan, "Dynamic Power Management in Wireless Sensor Networks," *IEEE Design and Test of Computers*, vol. 18, no. 2, pp. 62–74, March-April 2001.

[45] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, January 2005.

[46] W. Jin, P. Chan, and M. Chan, "On the Power Dissipation in Dynamic Threshold Silicon-On-Insulator CMOS Inverter," *IEEE Transactions on Electron Devices*, vol. 45, no. 8, pp. 1717–1724, August 1998.

[47] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance Microprocessors," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 1999, pp. 163–168.

[48] B. Davari, "CMOS Technology: Present and Future," in *Symposium on VLSI Circuits Digest of Technical Papers*, 1999, pp. 5–10.

[49] P. Larsson and C. Svensson, "Noise in Digital CMOS Circuits," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 6, pp. 655–662, June 1994.

[50] L. Wang, R. Krishnamurthy, K. Soumyanath, and N. Shanbhag, "An Energy-Efficient Leakage Tolerant Dynamic Circuit Technique," in *Proceedings of the IEEE ASIC/SOC Conference*, 2000, pp. 221–225.

[51] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 1998, pp. 239 – 244.

[52] M. Anders, R. Krishnamurthy, R. Spotten, and K. Soumayanath, "Robustness of sub-70nm Dynamic Circuits: Analytical Techniques and Scaling Trends," in *Symposium on VLSI Circuits Digest of Technical Papers*, 2001, pp. 23–24.

[53] A. Alvandpour, R. Krishanmurthy, K. Soumyanath, and S. Borkar, "A Sub-130-nm Conditional Keeper Technique," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 633–638, May 2002.

[54] R. Krishanmurthy, A. Alvandpour, G. Balamurugan, N. Shanbhag, K. Soumyanath, and S. Borkar, "A 130-nm 6-GHz 265×32 bit Leakage-Tolerant Register File," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 624–632, May 2002.

[55] L. Wang and N. Shanbhag, "An Energy-Efficient Noise-Tolerant Dynamic Circuit Technique," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 45, no. 11, pp. 1300– 1306, November 2000.

[56] G. Balamurugan and N. Shanbhag, "The Twin-Transistor Noise-Tolerant Dynamic Circuit Technique," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 2, pp. 273–280, February 2001.

[57] M. Allam, M. Anis, and M. Elmasry, "High-Speed Dynamic Logic Styles for Scaled-Down CMOS and MTCMOS Technologies," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2000, pp. 155–160.

[58] A. Alvandpour, R. Krishnamurthy, K. Soumayanath, and S. Borkar, "A Low-Leakage Dynamic Multi-Ported Register File in 0.13 μm CMOS," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 2001, pp. 68–71.

[59] R. Langevelde and F. Kaassen, "Timing Constraints for Domino Logic Gates with Timing-Dependent Keepers," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 1, pp. 96 – 103, January 2003.

[60] S. Jung and S. Kang, "High Performance Dynamic Logic Incorporating Gate Voltage Controlled Keeper Structure for Wide Fan-in Gates," *Electronics Letters*, vol. 38, no. 16, pp. 852–853, August 2002.

[61] V. Kursun, S. Narendra, V. De., and E. Friedman, "Analysis of Buck Converters for On-chip Integration with a Dual Supply Voltage Microprocessor," *IEEE Transactions on Very Large Scale Integration*, vol. 11, no. 3, pp. 514 – 522, June 2003.

[62] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, no. 2, pp. 584–593, April 1991.

[63] University of California Berkeley, *BSIM2 User's Manual*, 1994.

[64] University of California Berkeley, *BSIM3 User's Manual, Ver. 3.2.0*, 2001.

[65] R. Langevelde and F. Kaassen, "An Explicit Surface-Potential Based MOSFET Model for Circuit Simulation," *Solid State Electronics*, vol. 44, pp. 409–418, 2000.

[66] T. Sakurai and A. R. Newton, "Delay Analysis of Series-Connected MOSFET Circuits," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 2, pp. 122–131, February 1991.

[67] B. Sheu, D. Scharfetter, P. Ko, and M. Teng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 22, pp. 558–566, April 1987.

[68] Y. Tsividis, *Operation and Modeling of the MOS transistor: 2nd Edition*, McGraw-Hill Higher Education, 1998.

[69] C. Enz, F. Krummenacher, and E. Vittoz, "An Analytical MOS Transistor Model valid in all regions of operation and dedicated to Low-Voltage Low-Current Applications," *Journal of Analog Integrated Circuits and Signal Processing*, vol. 8, no. 1, pp. 83–114, May 1995.

[70] S. M. Sze, *Physics of Semiconductor Devices*, John Wiley and Sons, 1981.

[71] S. Jung, K. Kim, and S. Kang, "Noise Constrained Transistor Sizing and Power Optimization for Dual Vt Domino Logic," *IEEE Transactions on Very Large Scale Integration*, vol. 10, no. 5, pp. 532–541, October 2002.

[72] S. Thompson, I. Young, J. Greason, and M. Bohr, "Dual Threshold Voltages and Substrate Bias: Keys to High Performance, Low Power, 0.1μm Logic Designs," in *Proc. IEEE International Symposium on VLSI Technology*, 1997, pp. 163–168.

[73] R. Swanson and J. Meindl, "Ion-implanted Complementary MOS transistors in Low-voltage Circuits," *IEEE Journal of Solid-State Circuits*, vol. 7, no. 2, pp. 146–153, April 1972.

[74] B. Calhoun and A. Chandrakasan, "Standby Power Reduction Using Dynamic Voltage Scaling and Canary Flip-Flop Structures," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1504–1511, September 2004.

[75] S. Naredra *et.al.*, "Ultra-low Voltage Circuits and Processor in 180nm to 90nm technologies with a Swapped-body Biasing Technique," in *Proceedings of the International Solid-State Conference*, 2004, pp. 156–158.

[76] S. Martin, K. Flautner, T. Mudge, and D. Blaauw, "Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads," in *IEEE International Conference on Computer Aided Design*, 2002, pp. 721–725.

[77] J. Tschanz, S. Narendra, R. Nair, and V. De, "A 175mV Multiply-Accumulate Unit using an Adaptive Supply Voltage and Body Bias Architecture," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, pp. 1545–1554, November 2002.

[78] J. Tschanz, S. Narendra, R. Nair, and V. De, "Effectiveness of Adaptive Supply Voltage and Body Bias for reducing impact of Parameter Variations in Low Power and High Performance Microprocessors," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 5, pp. 826–829, May 2003.

[79] T. Chen and S. Naffziger, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for improving Delay and Leakage under the presence of Process Variation," *IEEE Transactions on Very Large Scale Integration*, vol. 11, no. 5, pp. 888–899, October 2003.

[80] S. Narendra, D. Antoniadis, and A. Chandrakasan, "Impact of using Adaptive Body Bias to compensate die-to-die V_t variation on within die variations," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 1999, pp. 229–232.

[81] I. Yand, V. Vieri, A. Chandrakasan, and D. Antoniadis, "Back Gated CMOS of SOIAS for Dynamic Threshold Control," in *Proceedings of IEDM*, 1995, pp. 877–880.

[82] T. Burd *et.al*, "A Dynamic Voltage Scaled Microprocessor System," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 11, pp. 1571–1580, November 2000.

[83] J. Kim and M. Horowitz, "An Efficient Digital Sliding Controller for Adaptive Power-Supply Regulation," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 5, pp. 639–647, May 2002.

[84] A. Stratakos, S. Sanders, and R. Broderson, "A low-voltage CMOS DC-DC converter for a portable battery-operated system," in *IEEE Power Electronics Specialists Conference, PESC '94*.

[85] V. Kursun, S. Narendra, V. De., and E. Friedman, "Low-voltage-swing Monolithic DC-DC Conversion," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 51, no. 5, pp. 241 – 248, May 2004.

[86] G. Wei *et.al*, "A Variable-Frequency Parallel I/O Interface with Adaptive Power-Supply Regulation," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 11, pp. 1600–1610, November 2000.

[87] T. Kuroda *et.al.*, "Variable Supply-Voltage Scheme for Low-Power High-Speed CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 454462, March 1998.

[88] D. Ernst *et.al*, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation," in *Micro Conf.*, December 2003.

[89] A. Bellaouar *et.al*, "Supply Voltage Scaling for Temperature Insensitive CMOS Circuit Operation," *TCAS-II*, vol. 45, no. 3, pp. 415–417, March 1998.

[90] K. Kanda *et.al*, "Design Impact of Positive Temperature Dependence on Drain Current in Sub-1-V CMOS VLSIs," *IEEE Journal of Solid-State Circuits*, vol. 36, no. 10, pp. 1559–1564, October 2001.

[91] A. Chandrakasan et. al, "Data-Driven signal processing: An approach for energy-efficient computing," in *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, 1996, pp. 347–352.

[92] R. Gonzalez and M. Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 9, pp. 1210–1216, August 1997.

[93] M. Elgebaly, A. Fahim, I. Kang, and M. Sachdev, "Robust and Efficient Dynamic Voltage Scaling Architecture," in *Proceedings of the IEEE ASIC/SOC Conference*, 2003, pp. 155–158.

[94] M. Nakai *et.al*, "Dynamic Voltage and frequency Management for a Low-power Embedded Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 28–35, January 2005.

[95] J. Daga and D. Auvergne, "A Comprehensive Delay Macro Modeling for Submicrometer CMOS Logics," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 1, pp. 42–55, January 1999.

[96] R. Ho *et.al*, "The Future of Wires," *IEEE Proc.*, vol. 89, no. 4, pp. 490–504, April 2001.

[97] Waterloo Maple Inc., *Maple 7.0 User's Guide*, 2001.