Application of Generalized Power-Delay Metrics to Supply and Threshold Selection in Deep Submicron CMOS

by

Dipanjan Sengupta

B.Tech, Institute of Radio Physics and Electronics, University of Calcutta, 2003

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Applied Science

in

The Faculty of Graduate Studies

Electrical and Computer Engineering

The University of British Columbia

June 2005

ABSTRACT

Application of Generalized Power - Delay Metrics to Supply and Threshold Selection in Deep Submicron CMOS

by

Dipanjan Sengupta

Power consumption has become as important as performance in today's deep submicron designs. As a result, high-level techniques and models must be developed to evaluate design changes in terms of power (energy) and performance tradeoff early in the design process. Recently, designers have been using the energy-delay product as a metric of goodness for CMOS designs due to certain perceived shortcomings of the more traditional power-delay product. As the industry moves to 90nm technology and encountered higher leakage currents, it is appropriate to revisit existing design metrics. In this thesis, a more general view of power and delay metrics for design optimization has been provided along with how these metrics can be used for design optimization.

Supply (V_{DD}) and threshold (V_T) voltage scaling are two popular methodologies of power reduction. As such, the effect on power and frequency are analyzed and the feasible region of operation is identified in the V_{DD} vs. V_T plane. A fundamental relationship is established between the optimal operating points and the generalized design metrics. In addition, new power and delay models are developed for logic blocks that incorporate the effect of V_{DD} and V_T. The effect of optimization on power and delay with respect to process, temperature and voltage variation has also been investigated.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor, Dr. Resve Saleh, for giving me opportunity to work with him. From the day I had entered UBC he has always given me tremendous enthusiasm in my work. I wish to thank him for the technical advice as well as the moral support he has given throughout my Masters.

I would also like to thank Dr. Naraig Manjikian of Queen's University for his suggestions and guidance. I am grateful to all the professors of the SOC group, as I have learned everything about VLSI while taking their courses. Thanks for their comments and feedback, which helped immensely in my work.

I am grateful for all the help provided by our CAD manager, Roozbeh, to cater all the needs relating to tools and simulations. Roberto, our Test Lab manager, and Sandy, our administrative assistant, have been particularly helpful in all the needs of the lab. It has been great to work with other students of the SOC Lab. I especially thank Mohammad, Neda, Partha, Cristian, Zahra, Samad, Peter, Zion, Nathalie, Baoshang, Scott and Victor.

Last but not the least, I would like to thank my entire family for all the love and support they have shown to me. Finally I would like to thank Madhuja for the encouragement and support she had been providing throughout my UBC years.

Many thanks to NSERC, PMC-Sierra, the Canadian Microelectronics Corporation and the University of British Columbia for the financial and tool support that made all this work possible.

Chapter 1

Introduction

1.1 Research Motivation

In the past, the main goal of digital design was to deliver the highest possible performance, often at the expense of area and power. However, if microprocessors continue to increase their speed by 2X in every generation, the power will quickly exceed all power density limits on the chip [1]. Even for designs with much lower performance requirements, power dissipation may be an issue. Many embedded systems, such as PDAs, require processing of the given applications with rigid low-power budgets [2], basically due to limitations in battery life. Recently, the IC industry has shifted its priorities and made power an equally important design issue in deep submicron technology [3]. Some have argued that power is even more important than delay in terms of design objectives for both high-end microprocessors and for low-power hand-held devices. As number of transistors and their associated leakage currents increase, power consumption is now widely recognized as a key design challenge [4].

Consider the trends associated with x86 microprocessor family of products from Intel[tm] over a 20-year time period. From one technology to the next the following characteristics are observed:

1. Frequency increasing by 2X (Fig. 1(a))
2. Transistor density is increasing by 2X (Fig 1(b))
3. Active power increasing by 2.7X (Fig 1(c)) [5]

The impact of scaling on power consumption is two-fold. Until recently, the major source of power consumption of a chip was the active power, which is consumed while performing the intended operation of the chip. Trends in the active power consumption of microprocessor chips (Fig. 1(c)) show that power consumption would exceed any realistic limit if no further measures were adopted to control it [5]. Static power, consumed when a circuit is idle, is the other source of concern due to scaling. In order to achieve increased frequency at each technology node, the threshold voltage also needs to be reduced. This causes an increase in static power due to an increase in subthreshold leakage current, which is presently its primary source. Projections show that the subthreshold power density (or the static power density) would soon be equal to, and may even surpass the active power density (Fig. 1(d)) [6].

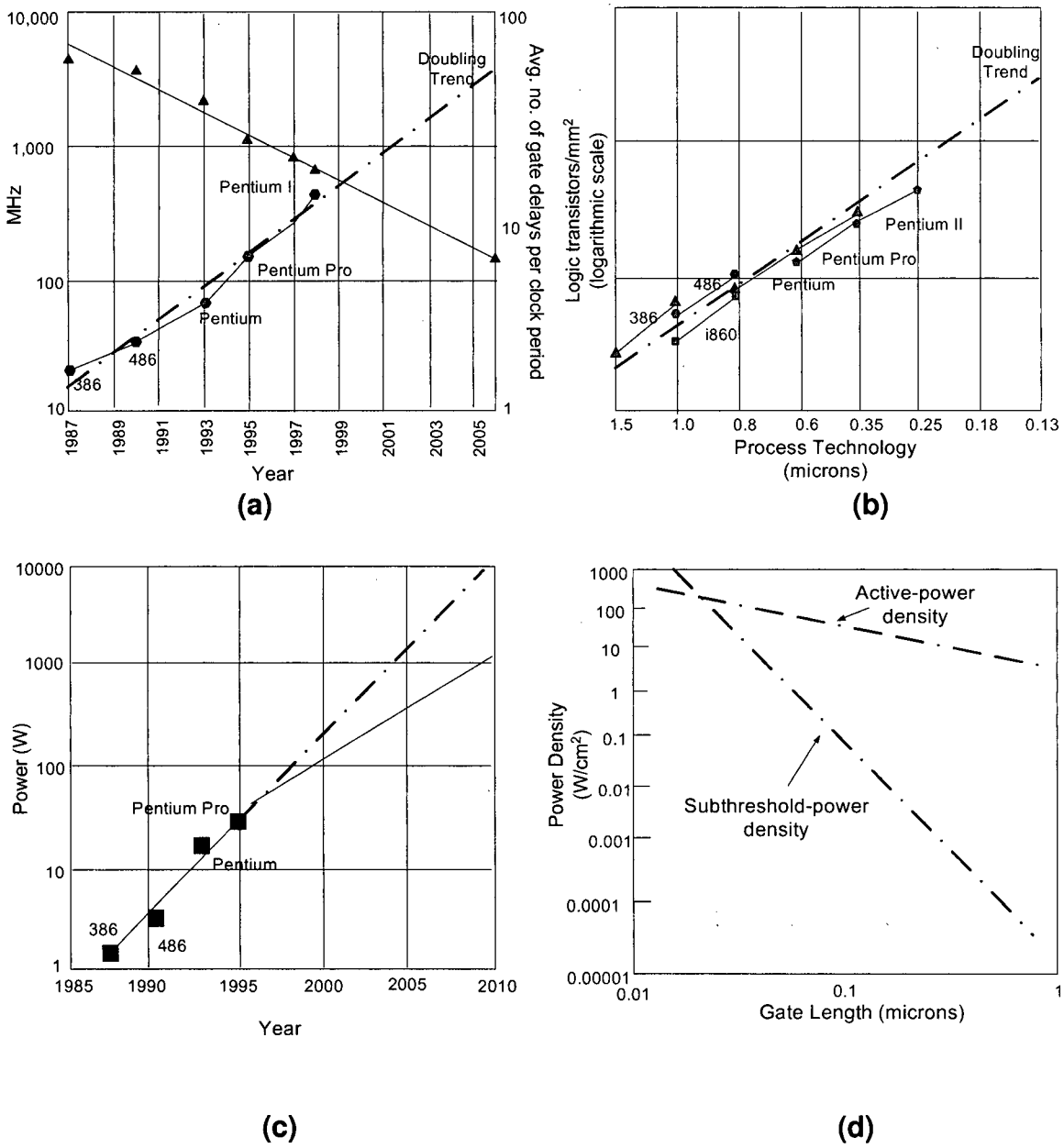


Figure 1. Technology Trends (a) Frequency and gate Delay (b) Transistor Density (c) Active Power (d) Active and Subthreshold Power Density.

Of the numerous techniques of reducing power, supply and threshold voltage scaling have become increasingly attractive solutions to designers [7, 8]. This is primarily due to the quadratic saving in power with lower supply voltage, and reduced leakage power with higher

threshold voltage. But this power saving is often at the expense of reduced performance, since a reduction in supply voltage or increase in the threshold voltage would in turn decrease the operating frequency of the design. Given that there is a tradeoff between power and frequency, their combined effects need to be investigated with technology scaling as a function of V_{DD} and V_T. That is, exploring power and delay metrics in a V_{DD} versus V_T plane would yield valuable information when designing low-power circuit blocks for a given application.

To investigate how the choice of supply and threshold voltage would provide power saving as well as meet the target frequency, power and delay analysis should be performed at an early stage in the design process. This leads to the requirement of power and delay models which can be used for the analysis. The models must be designed to be a function of V_{DD} and V_T. Once these models are developed, they need to be integrated together in a simulation tool to evaluate the proper balance between power and frequency to achieve optimal performance. The models can be developed at different levels of abstraction, with a commensurate tradeoff of accuracy and speed. In previous work [9], *Voltage Island* technique have been proposed, where individual IP cores are assigned certain supply voltages depending upon their power and frequency requirement. Any new model should allow both the supply and threshold voltages to be assigned at runtime.

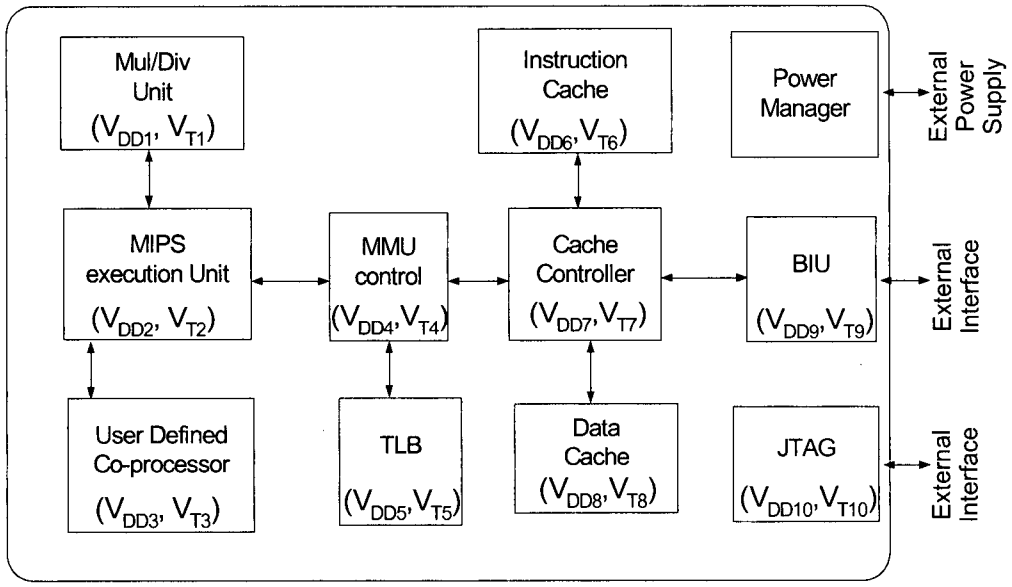


Figure 2. Block Diagram of Power Management Methodology.

As a representative application, consider the computer system in Figure 2, which has been adapted from previous work [10]. In this figure, each block is shown to have its own value of V_{DD} and V_T. The goal would be to select suitable values of V_{DD} and V_T to optimize the overall frequency and power of the design. The granularity of control of the V_{DD} and V_T would depend upon the overhead allowed for power management. The task of the power manager is to dynamically coordinate with the Operating System selection of the appropriate supply and threshold voltage values for each logic core. The information needed to design the power manager could be derived from simulation, assuming that detailed description of the hardware are provided to the tool along with the models of sufficient accuracy. For this purpose, new power models are needed that are data-dependent and vary according to the V_{DD} and V_T values. Moreover the models can be used to understand the relation between certain design metrics and their optimal solution for a given design block. The balance between the amount

of gain versus the granularity of control of V_{DD} and V_T can be understood clearly using these models and performing simulation with different applications.

Systematic and random variations in process, supply voltage and temperature (PVT) are posing a major challenge to the future high performance design [11, 12]. Process variation typically leads to a fluctuation of threshold voltage (typically 20% from nominal value). Both frequency and power are affected by this variation causing circuits to have non-ideal operating speed and over-budget power consumption, which may reduce efficiency. The impact of temperature variation is on the leakage current. Increase in temperature would raise the leakage current, requiring circuits to reduce operating speed for lowering the overall power consumption. Voltage variation across the chip would affect the frequency as well as the overall power consumption of the system. Thus whatever power and frequency optimization is done in terms of V_{DD} and V_T scaling; the PVT effect has to be considered. This would then provide a more generalized and practical solution to the optimization problem.

1.2 Research Objective

Previous work [13] suggests that the energy-delay product (EDP) is a useful metric for evaluating the quality of a design. But this metric may not be appropriate when low-power dissipation is a priority. As the recent 90nm technology has exhibited a high degree of leakage, it raises some interesting questions regarding the use of EDP as a design metric. Some new

insights into the metrics in use today and their relationship to the optimal operating point for a given design are provided here.

Process, Voltage and Temperature variation would always affect the decision on choosing the supply and threshold voltage. An ideal (V_{DD}, V_T) pair will quickly become a non-ideal one due to effect of one of the parameters or any combination of them. To provide a realistic view of this tradeoff, these varying parameters are also included.

To analyze a system, power and delay models would allow designers to make early estimates of power savings. Dynamic supply and threshold voltage scaling is one way of providing optimization in terms of power and delay. This work also focuses on how to develop these models. First, VHDL descriptions of circuits are developed and synthesized. Next, data-dependent power models are developed based on the VHDL descriptions. Further the models are enhanced to represent power and delay as a function of supply and threshold voltage. As a case study, an adder and a multiplexer are used for modeling purposes. This methodology can be used for developing other models and can be extended for larger logic blocks and then used for system level analysis.

1.3 Thesis Organization

In Chapter 2, a brief review of the power components and methodology of optimization in a chip are provided. Chapter 3 first provides an overview of the design metrics followed by

discussion on generalized power-delay metrics as well as comparison of three specific metrics. Chapter 4 illustrates the application of these metrics in the V_{DD} vs. V_T plane. In Chapter 5, the effect of Process, Voltage and Temperature variations is investigated. Chapter 6 deals with data-dependent power modeling and addresses some practical delay modeling issues. Conclusion and future work are presented in Chapter 7.

Chapter 2

Background

2.1 CMOS Power Components

Active and static power are the two components of power consumption of a chip. The active power can be further subdivided into dynamic power (including glitches) and short circuit power. The dynamic component is the power consumed due to charging and discharging of the load capacitance. Typically during a low-to-high output transition, supply current charges the output load through PMOS devices; and during the high-to-low output transition, the stored charge in the load capacitor is discharged through NMOS devices to the ground. The dynamic power depends linearly on the load capacitance, C, the average switching activity, α, quadratically on the supply voltage, V_{DD} and linearly on frequency, f_{dk} as shown in Equation (1):

$$P_{dynamic} = \alpha C V_{DD}^{2} f_{clk} \quad (1)$$

While the activity factor should be computed for each gate individually, the average activity factor and capacitance of the whole chip may be computed and used to estimate the average chip power due to switching [14].

The short circuit power, due to crowbar current, is dissipated during a switching event when there is a direct path from supply to ground. The finite slope of the input signal causes a direct current path between the supply rail and ground through the PMOS and NMOS devices for a short period of time during switching events [15]. The short circuit current is a strong function of the ratio of the input and output slope which leads to a tradeoff between the dynamic power of the previous gate and the short circuit power of the next gate [14].

The glitch power is expended when the inputs to a gate do not arrive at the same time causing a small glitch at the output. Glitches tend to propagate through the fan out gates and cause unintended transitions in the subsequent stages, increasing the power dissipation even further [14]. The effect of both the short circuit and glitch power can be incorporated into the α in Equation (1).

Static power is mainly due to subthreshold leakage current. This is the leakage current that flows between the source and the drain when the gate voltage is smaller than the threshold voltage. NMOS devices have a higher leakage than PMOS devices. Static power can be represented as:

$$P_{static} = I_{leak} V_{DD} \tag{2}$$

where I_{leak} is the subthreshold leakage current.

With technology scaling, the supply voltage has been progressively reduced. In order to sustain the traditional 30% improvement in gate delay for digital circuits in each generation, the threshold voltage has to be scaled aggressively [16] causing the subthreshold power dissipation to be of non-negligible value.

The total chip power can be simply viewed as the sum of dynamic power and static power[1]. A general formula for the power consumption on a chip can thus be estimated as:

$$\begin{aligned} P_{total} &= P_{dynamic} + P_{static} \\ P_{total} &= \alpha C V_{DD}{}^2 f_{clk} + N(1-\alpha) V_{DD} I_{leakage} \end{aligned} \tag{3}$$

where α is the activity factor (which also accounts for the short circuit and glitch power), C is the sum total of all load capacitances in the design, V_{DD} is the supply voltage, f_{dk} is the clock frequency, N is the number of gates, $I_{leakage}$ is the average gate leakage current. For a single gate, subthreshold current leakage can be modeled from [17] in a simplified expression as:

$$I_{leakage} = A e^{\frac{q}{nkT}(V_{GS}-V_T)} (1 - e^{\frac{-qV_{DS}}{kT}}) \tag{4}$$

where A is a constant which is technology dependent, V_{GS} is the gate-to-source voltage, V_{DS} is the drain-to-source voltage and V_T is the threshold voltage. The threshold voltage of a short-channel MOSFET transistor in BSIM model [18] is given by:

$$V_T = V_{T0} + \gamma(\sqrt{\phi_s - V_{BS}} - \sqrt{\phi_s}) - \theta_{DIBL} V_{DD} + \Delta V_{NW} \tag{5}$$

[1] In this thesis, dynamic power and active power would be used interchangeably while leakage power and static power are used interchangeably.

where V_{T0} is the zero-bias threshold voltage, ϕ_s, γ and θ_{DIBL} are constants for a given technology, V_{BS} is the voltage applied between body and source of the transistor, and ΔV_{NW} is a constant that models narrow width effects. Considering V_{GS}=0V for NMOS and V_{GS}=V_{DD} for PMOS we can represent the $I_{leakage}$ in a simplified fashion as:

$$I_{leakage} = I_0 e^{\frac{-V_T}{nV_{th}}} \tag{6}$$

where V_{th} is the thermal voltage, n is a subthreshold swing parameter and I_0 is the leakage current coefficient that effectively accounts for all the other terms of Equation (4). Combining Equation (3) and (6) the power consumption on a chip can represented as:

$$P_{total} = \alpha C V_{DD}{}^2 f_{clk} + N(1-\alpha) I_0 e^{\frac{-V_T}{nV_{th}}} V_{DD} \tag{7}$$

2.2 Background of Power Modeling

Previously, power estimates of a given design were performed at a very late stage of the design. Today, designers have to consider certain trade-offs that reduce the power consumption early in the design process. This growing demand of power analysis caused the CAD tools to include power analysis capability at higher levels of abstraction. The choice of abstraction level of power modeling depends upon the required accuracy and the amount of CPU time available for analysis. In this work, the higher level or microarchitectural-level power analysis is considered.

Initially, power estimation was based on the efficient way of estimating current drawn by any logic block. This approach has been used in [19, 20, 21, 22] where the current drawn by the CPU is noted for different instructions via some test program. In this way, an average current value is obtained for each instruction, which can then be used for power estimation. The main drawback of this method is that every time any architectural, process or technology change is made in the logic block, it would render the model inefficient.

Recently data-driven power estimation tools have been introduced [23, 24, 25]. The main motivation of such modeling is that the amount of switching within a circuit would have direct correlation to the input/output bit-flips of the circuit. By knowing how much flipping occurs in the input/output, rough estimates of the power consumption can be predicted. Wattch [23] is built on the SimpleScalar 3.0 simulator [26] that has been extended from 5-stage pipeline to 8-stage pipeline. Wattch developed basic components – array structures, content-addressable memories, combinational logic and wires and clocking [27]. The main difficulty of this tool is that models are not general-purpose; that is, the power consumption estimate for the datapath and the execution blocks are not scalable. Moreover the lack of granularity as well as the inability to report the percentage of error for each component leaves the user with little knowledge about the accuracy of the results.

Simplepower [24] is another execution driven, RTL energy estimation tool. It builds its models based on transition counts. It is integrated with the SimpleScalar tool set and has a LUT-based energy model for each functional unit based on the switching activity. The main purpose of this tool is to give a first-order comparison of the architectural and algorithmic trade-off

during the initial phase of design. The reference design, while accurate enough for the purpose of the trade-off analysis, is not easily modifiable to describe specific alternative designs that may have a different datapath width, smaller feature size, or different technology [28].

Another activity sensitive power estimation tool is the Cai-Lim power model [25]. It is built upon SimpleScalar 2.0 simulator. The SimpleScalar architecture is subdivided into smaller blocks and analyzed individually using HSPICE simulation for dynamic power, static power and area. These values are pre-computed and included as a part of the source code. The inconvenience of this tool is that lack of direct access to the models used to generate these values makes it difficult to scale units appropriately, even within the same process and architecture family [27]. Moreover, the simulator is based on 0.25μm TSMC technology files and is difficult to modify.

A parameterized and technology scalable microarchitecture-level power modeling technique has been proposed in [28]. Capacitance extraction of each logic block and thereby calculating the dynamic power is done in this model. This method can be used on a block-by-block basis and be incorporated in SimpleScalar toolset. One limitation of this method is the lack of static power modeling as well as the need of too much detailed information of the transistor sizing of the gates. In fact, none of the models to date incorporate the effects of V_{DD} and V_T scaling, so new models are needed to fill this gap.

2.3 Power Management Using V_{DD} and V_T scaling

There are numerous methodologies for power minimization starting from the software level, to the architectural level and eventually down to the circuit level. Equation (7) indicates which terms may be optimized at some level in order to reduce the overall power consumption. In this work, changes in power dissipation and clock frequency when modifying specific terms in the equation, namely V_{DD} and V_T, are of interest. While this thesis describes metrics and models the following sections illustrates the methods being developed for varying these voltages using on-chip circuitry.

2.3.1 Dynamic Supply Voltage Scaling

Supply voltage scaling [7, 29, 30] has been a widely accepted methodology for power optimization. Active power consumption varies linearly with frequency and quadratically with supply voltage so that the power reduction is roughly cubic when consequential reduction in frequency is included [31]. The major shortcoming of this solution has been that lower supply voltage causes the lowering of the circuit speed. Alternatively when a majority of the computation does not require maximum throughput, then the average energy consumption can be significantly reduced, for a given computation, which is important for battery-operated devices [7]. This is seen in portable devices like cell phones and PDAs, which encounter typically three types of operation, namely computation-intensive task, low-speed function and idle-mode operation. In the first case maximum throughput is demanded. In the second case, more relaxed deadline requires a fraction of the maximum throughput of the processor.

Finishing these tasks early has no benefit and thus voltage scaling at the cost of reduced speed can be applied here.

The method of dynamically adjusting the supply voltage at runtime is called *Dynamic Voltage Scaling (DVS)* [30]. It was found that completing a task before its deadline and then idling is less energy efficient than running the task more slowly to begin with, and meeting its deadline exactly [32]. The central issue on DVS is to set the right performance level for a processor so that energy is conserved while meeting the deadline. Also this methodology is enabled by the observation that the delays of most CMOS circuits are functions of supply voltage and track each other well over a range of supply voltages, which is a necessity for system operation under varying supply conditions [33]. Of course, there is an extra overhead on the decision of the correct V_{DD} and also the requirement of special circuitry for realizing such values. Moreover, circuits implementing this variable supply voltage, as proposed in [29], do have some fluctuations that must be taken into consideration.

Commercial chips like IEM926 that has ARM926EJ processor core have *Intelligent Energy Manager (IEM),* which implements DVS [30]. The IEM has a software as well as hardware unit for monitoring the system workload. The software component uses information from the Operating System to build up a historical view of the execution software running on the system and then some algorithms are used to classify the tasks and use them for global prediction about future workloads [34]. Additionally on the hardware side, there is an Adaptive Power Controller (APC) used for implementing DVS. In Figure 3, a block diagram of how dynamic voltage scaling is performed in [35] is depicted.

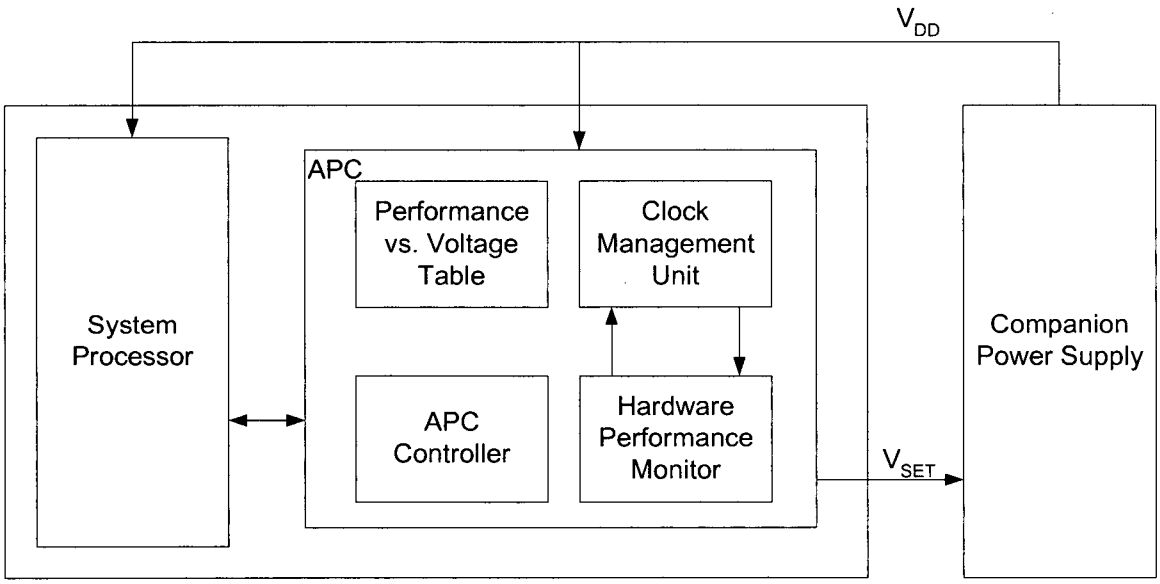


Figure 3. Circuit Used to Implement DVS [35].

The APC module performs the voltage scaling and clock management of the processor. The Performance vs. Voltage Table block gives the choice for setting the correct supply voltage for the required frequency of operation. The hardware performance monitor works in association with the IEM software for making predictions about systems performance.

2.3.2 Dynamic Threshold Voltage Scaling

Leakage power is now becoming an important concern for designers. Due to the inverse exponential relation of threshold voltage to the leakage power, threshold voltage scaling has become an attractive idea [23, 36, 37]. This can be achieved using more than one threshold voltage and assigning them to the appropriate circuits depending upon performance and low leakage requirement.

Dual threshold voltage [36] is a straightforward approach where a circuit is partitioned into high-V_T and low-V_T regions. Low-V_T is chosen for devices that are in the critical path and high-V_T is chosen for the non-critical path. This would give savings in leakage power for the high-V_T gates but performance would be maintained due to the low-V_T gates in the critical path. The main disadvantage of this method is the presence of many critical paths in a circuit that cause large portions of it to have low-V_T thereby reducing the effectiveness of the technique [16].

Multiple-Threshold CMOS (MTCMOS) reduces leakage power by using high-V_T sleep transistors to gate the power supplies for the low-V_T block [38]. Though significant saving in leakage power is achieved in the sleep mode, a relatively large size of the sleep transistor is required. Moreover, the internal nodes of the circuit might be floating and cause data loss. This has prompted designs implementing MTCMOS to have special circuits that have the capability

of retaining state during the standby mode. In Variable Threshold CMOS (VTCMOS), the threshold voltage is adjusted by biasing the body terminal [16]. Using this Adaptive Body Biasing (ABB), the transistors threshold voltage can be set to high value in standby mode and be reduced dynamically in the active mode depending upon the required performance level. Initially the circuit's threshold voltage is set at a low-V_T. Then, by applying reverse bias voltage (by suitably choosing V_{BS} in Equation (5)), the threshold voltage can be achieved for the targeted frequency of operation. This is illustrated in Figure 4(a) where $V_{BB,p}$ and $V_{BB,n}$ are used to control the threshold voltage of the PMOS and the NMOS transistors. In this way, leakage power in stand-by mode as well as active mode is saved. ABB can be applied to the entire die, which would cause the same threshold voltage for the entire chip or it could be specified according to block-by-block basis and controlled individually. This technique has been used in V_T- hopping where certain discrete threshold voltages are chosen dynamically [37] depending upon the workload of the processor.

Recently, *Dynamic Threshold Voltage Scaling (DVTS)* through substrate biasing has been proposed [8]. A block diagram of the DVTS scheme and the feedback loop is presented in Figure 4(b). A clock speed scheduler decides the intended frequency of operation. The DVTS controller then adjusts the PMOS and the NMOS body bias to have the same oscillation frequency as the intended one. The continuous control scheme would take into consideration supply voltage variation and temperature change, and would adapt the threshold voltage to regulate the delay of the circuit [39].

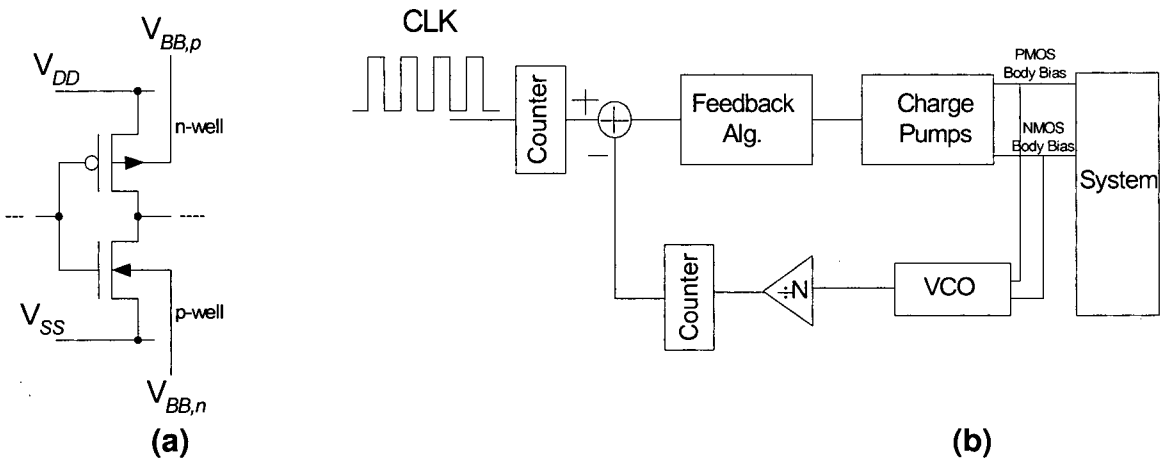


Figure 4. Circuit Used to Implement DVTS [8].

For the work described in this thesis, it is assumed that both DVS and DVTS are actively used to select the proper (V_{DD}, V_T) pair to achieve certain design requirements. It is also assumed that a given design block uses a single (V_{DD}, V_T) value although it can be varying over time. It should be mentioned that, while from theoretical point of view a range of (V_{DD}, V_T) may be possible, in practice only a subset of them might actually be feasible. From results of an embedded StrongARM it has been shown [40] that not all voltage levels are practical. This is because of the fact that beyond certain supply voltage the circuit will fail to operate correctly. However, in all the following analysis it is assumed that all values are theoretically possible.

The purpose of this thesis is to provide a framework in terms of metrics and models to carry out experiments and determine the optimal settings of V_{DD} and V_T. The selection of the best values of V_{DD} and V_T are based on the design metrics, as described in the next three chapters. The development of suitable high-level models for power and delay that vary with V_{DD} and V_T is the subject of Chapter 6.

Chapter 3

Design Metrics

3.1 Previous Metrics

Design metrics are widely used to provide a measure of goodness when comparing different designs that carry out the same function. They can also be used to optimize a design to achieve a minimum value of a given design metric, or a *weighted* set of metrics. For integrated circuits, power and delay (i.e., operating frequency) are the two most important design specifications. So the primary metric in use today is based on these two factors. The delay component here is simply the clock period ($T=1/f$), where f is the clock frequency. One can speed up or slow down a circuit by adjusting device sizes, the clock frequency, the supply voltage, or the threshold voltage, by modifying the body-bias voltage, V_{BS}.

The traditional design metric to minimize both power and delay of electronic designs is the power-delay product, which is essentially the energy per operation [14]. In 0.35μm CMOS

designs and above, dynamic power was the dominant component of the total power. For this specific case, the metric could be represented as follows:

$$Energy = P_{dynamic} \times Delay = \alpha C V_{DD}^2 f \frac{1}{f} = \alpha C V_{DD}^2 \quad (8)$$

Note that the frequency term gets cancelled and the designer is only left with only α, C and V_{DD} to adjust the energy. This is not a completely useful metric since one can reduce the energy by using smaller devices (to reduce C) or by reducing V_{DD}; unfortunately, this will also reduce the speed of the design and this is not directly reflected in the metric. Therefore, one cannot effectively trade-off power and delay. In fact, the optimal V_{DD} for minimum energy is achieved with V_{DD}=0, which is not a meaningful result.

An improved metric was developed to circumvent this problem as follows: The energy is multiplied by another delay term to obtain the energy-delay product [13]. This produces the following result:

$$Energy \times Delay = \alpha C V_{DD}^2 \frac{1}{f} = \frac{\alpha C V_{DD}^2}{f} \quad (9)$$

Now the delay term is reinstated in the design metric and adjustments in C or V_{DD} are immediately reflected in f and the metric is once again able to provide a tradeoff between power and delay. Also, the optimal value of V_{DD} is no longer zero.

3.2 Generalized Metrics

It is well known that, with the scaling of standard CMOS technology, there is a significant increase in subthreshold leakage current in each transistor when it is normally off. As a result of an increase in the number of gates on a chip, the leakage power has increased noticeably [1]. While the leakage of a single gate is very small, the leakage of all inactive gates on an entire chip can be large. Hence, leakage power management has become a major design issue [17]. Typically, a higher effective V_T is used to reduce the subthreshold current for inactive gates. Of course, this slows down the gates and contributes to an overall decrease in the clock frequency.

Since there is a significant increase in leakage power in 90nm CMOS, it is appropriate to revisit the effect of this change on the power-delay metrics [41]. First, consider the original power-delay product:

$$\begin{aligned} Energy &= \left(P_{dynamic} + P_{leak}\right) \times Delay \\ &= \alpha C V_{DD}^2 + \frac{N(1-\alpha) I_0 e^{\frac{-V_T}{nV_{th}}} V_{DD}}{f} \end{aligned} \tag{10}$$

In Equation (10), the energy metric now contains the frequency term. This implies that energy contains information about the changes in timing due to the changes in the design. Hence, energy is a useful design metric if significant leakage is present. Note that the frequency term is now only associated with the leakage power.

Similarly, the energy-delay product can be rewritten as:

$$Energy \times Delay = \left(Energy_{dynamic} + Energy_{leak}\right) \times Delay$$
$$= \frac{\alpha C V_{DD}^2}{f} + \frac{N(1-\alpha) I_0 e^{\frac{-V_T}{nV_{th}}} V_{DD}}{f^2} \quad (11)$$

Here, the frequency appears in both terms, except that it has a different exponent in each case.

These two metrics can also be viewed another way. If P represents power and D represents delay (in the critical path) of the circuit then the metrics can also be represented as follows:

$$Energy = P \times D$$
$$Energy \times Delay = P \times D \times D$$

So the energy-delay metric (EDP) gives a higher geometric weighting to delay than power, whereas the energy metric gives balanced weighting to both. In other words, the EDP prioritizes delay above power, while the energy metric assigns equal priority to both. The energy-delay metric is more suitable when performance is the primary concern. On the other hand, if power is of higher priority, then neither metric captures this aspect. Thus to prioritize power, another metric could be introduced that gives more emphasis on power rather than delay, as follows:

$$Power \times Energy = P \times P \times D$$

In this metric, which is referred to as the power-energy product (PEP), power has a higher geometric weighting than delay and thus produces a lower power solution than the other two metrics.

This leads very naturally to the notion of a generalized set of metrics based on power and delay as follows:

$$Generalized\ Power\text{-}Delay\ Metric = P^m D^n$$

When $m=0$ and $0 < n < \infty$ then the entire emphasis is on the reduction of delay. For $0 < m < \infty$ and $n=0$, emphasis is only on power reduction. For all other values of m and n, as long as $m>n$, power remains the primary concern, and for $m<n$, delay becomes the primary concern. The metric to be chosen should depend upon the overall design optimization goal. To balance the two, clearly the standard power-delay metric should be used.

3.3 Comparison of Three Metrics

While it is instructive to generalize the notion of power-delay metrics, only a few such metrics are actually of practical value to the designer. Ideally, a useful metric should lead to a non-trivial optimal solution. Furthermore, the usefulness of any given metric may increase or diminish, depending on the technology characteristics. To illustrate how technology scaling may affect the quality of a metric, consider the circuit of Figure 5.

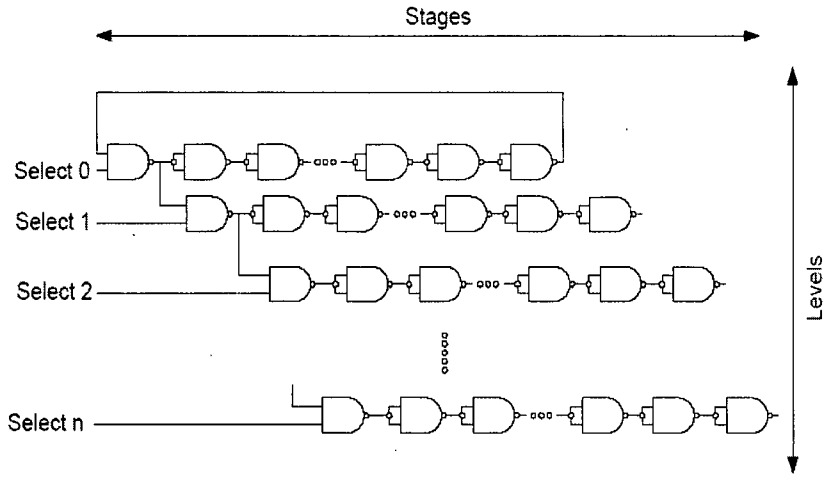


Figure 5. Circuit Used to Evaluate Metrics.

It is taken from [42] and will be used to compare the energy, energy-delay, and power-energy metrics. It has a NAND-based ring oscillator at the top level and NAND chains in subsequent levels. Its select inputs are used to control the amount of activity in the overall circuit. For a circuit that comprises of m stages and n levels if *Select j* is "*0*", where $0 \leq j \leq n$, then $m \times j$ gates would be active and the remainder of the gates would be inactive. Thus to have the minimal activity other than the ring oscillator only, *Select 0* is set to "*1*" and all the other selects are set at "*0*".

The circuit is simplified and consider only a single level for the moment by setting *Select 0* = "*1*" and all other *Select* lines to 0. This is equivalent to a ring oscillator circuit. The total power of a ring oscillator with N stages is of the form:

$$Power = C_L V_{DD}^2 f + (N-1) I_0 e^{\frac{-V_T}{nV_{th}}} V_{DD} \qquad (12)$$

and the frequency of the ring oscillator can be written as:

$$f = \frac{k_1 (V_{DD} - V_T)^2}{N V_{DD} (V_{DD} - V_T + k_2)} \tag{13}$$

where k_1 and k_2 are constants that depend on the specific technology. The delay through the circuit provides the maximum allowable clock frequency of the ring oscillator. HSPICE simulations were performed using this circuit to validate the results of these equations.

First consider the metrics when dynamic power dominates; assuming static power is zero by setting $I_o = 0$ in Equation (12). Next, V_T is fixed and then the metrics are optimized for V_{DD}. In that case, the graphs of the three metrics as shown in Figure 6 are produced. Clearly, both PEP and PDP are not meaningful since the optimal value of V_{DD} is zero in both cases. However, EDP has a non-zero optimal value and therefore serves as the only useful metric.

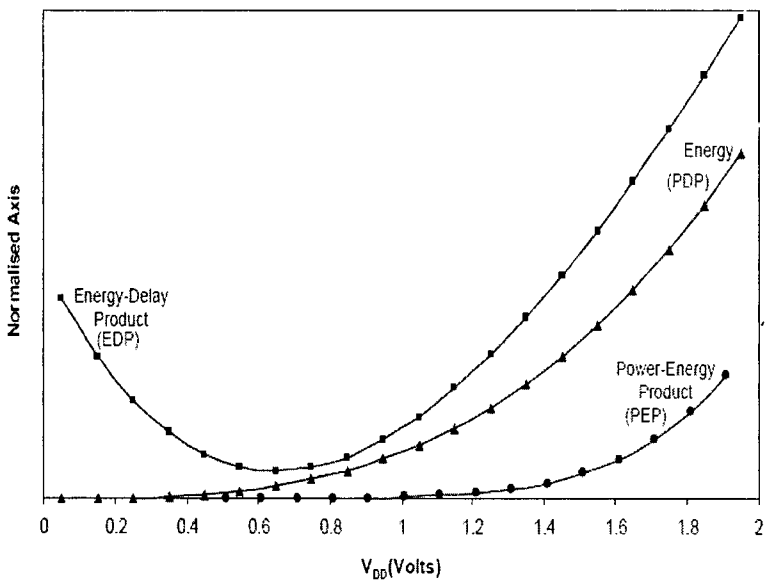


Figure 6. PDP, EDP and PEP Metrics when Dynamic Power Dominates.

Next, 90nm CMOS technology is considered where both static and dynamic power components contribute to the total power. For the given design, Figure 7 shows the three different metrics, namely, power-energy, energy, and energy-delay, for different V_{DD} and fixed V_T. Unlike the previous situation, there is now an optimal V_{DD} for each of the metrics. Each one provides a different tradeoff between power and delay. Thus it can be concluded that any one of the metrics may be used for design optimization, depending on the objectives. However, the actual power and delay values will change depending on which metric is used.

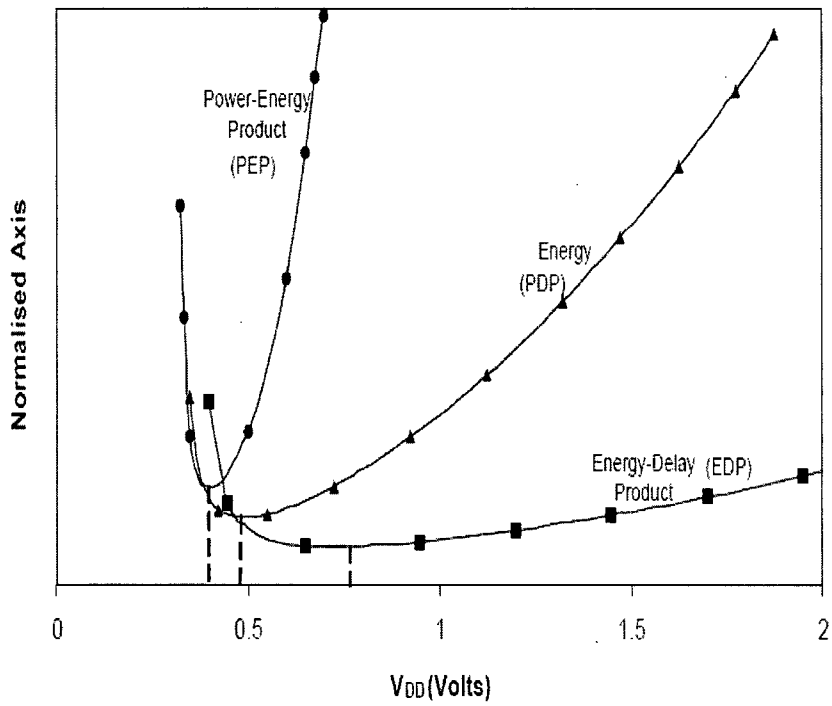


Figure 7. PDP, EDP, PEP Metrics with Static and Dynamic Power.

Figure 8 indicates the delay and total power (both static and dynamic) for different metrics at their optimal V_{DD} point for a fixed V_T. The delay of the system decreases as one makes a transition from PEP to PDP to EDP. On the other hand, the total power consumption steadily decreases in the opposite direction. The PDP values remain somewhere in between.

Of the three metrics, EDP prioritizes delay over power while PEP optimizes power over delay, and PDP optimizes both equally. So for EDP, the delay will be the lowest of the three cases, while for PEP, total power consumption would be the minimum of the three cases. In case of PDP, the results lie in between the two extremes, as neither has a higher priority over the other. It is to be noted that for a power-centric design paradigm, EDP may not be as effective as PDP or PEP.

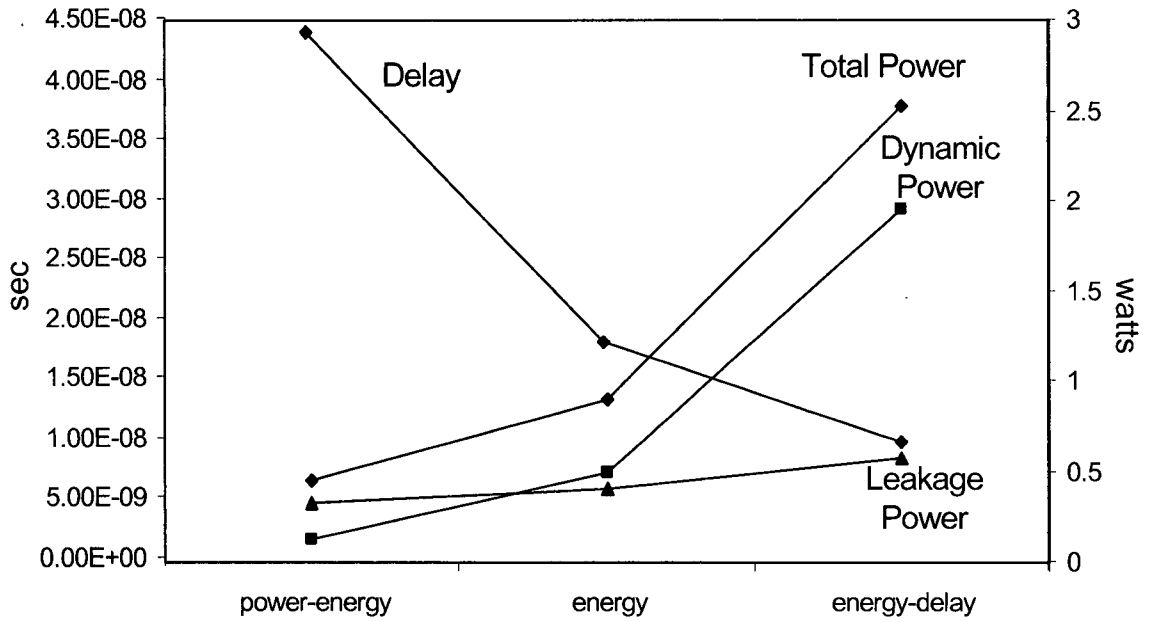


Figure 8. Dynamic and Leakage, Total Power and Delay for the Three Metrics.

An unexpected result of the analysis was encountered regarding the percentage of dynamic and leakage power in the total power. The dynamic power decreases more rapidly than leakage power from EDP to PDP. Surprisingly, the percentage of leakage power is *higher* in the case of the PEP metric and is almost equal to dynamic power in PDP metric. But for EDP, the dynamic power is much higher than the leakage power. A logical conclusion from this trend is that, for a system where low power is of primary concern, it may have a higher leakage

percentage than other systems. It is contrary to the belief that a high percentage of leakage is always detrimental to the system. As shown in the graph, it all depends upon what is being optimized. If delay is of greater importance, then leakage power is a much smaller percentage of the total power, but is larger in magnitude than the other two cases.

When power minimization is needed, then leakage power may actually be larger than dynamic power. Hence, higher leakage current, in relative sense, may be experienced to produce the desired result. Finally, it must be noted that dynamic power and leakage power are about equal in the PDP case. Intuitively, a balance between the two components is expected at the optimal point. It should also be noted that these results assume that the optimal V_{DD} and V_T values are used throughout the entire design.

Hence, in systems where low power is of more importance, the PEP metric is a useful choice. Systems where performance has higher priority, the EDP metric can be used. And systems that consider power and delay as equally important should use the PDP metric. The relationship between the various metrics with respect to low power design is explored in the next chapter.

Chapter 4

Application of Metrics to V_{DD} and V_T Scaling

With industry now focusing on power, researchers have been exploring the power-delay tradeoff as a function of V_{DD} and V_T [3, 17]. It is clear from previous discussion that a lower V_{DD} decreases dynamic power quadratically, but also causes an increase in delay. To reduce the delay, it is possible to decrease V_T, thereby increasing the gate overdrive term. Of course, this increases the static power exponentially, according to Equation (7). Therefore, determining the optimal (V_{DD}, V_T) pair to satisfy both power and frequency specifications is an important issue. In the rest of this chapter, the power and delay metrics are compared on a V_{DD} vs. V_T plane.

It is assumed that the optimal values of V_{DD} and V_T are used throughout a given design, whether it is a chip, a block or sub-block of logic. DVS and DVTS are the two methods used to obtain these optimal values. Moreover it is assumed that all values of V_{DD} and V_T are

possible, although in practice, only certain discrete values are permissible. Since various metrics are compared here relative to one another, this is a valid assumption.

To explore the metrics on a representative design, the circuit of Figure 5 is used and the inputs are adjusted to provide roughly 10% activity [43]. Moreover the analysis is carried out on the circuit for 90nm technology. The power and frequency values are computed using all combinations of V_{DD} and V_T.

4.1 Effect on Feasible Operating Region

Given minimum frequency and maximum power dissipation specifications, a feasible operating region can be defined on the V_{DD} vs. V_T plane. The frequency line determines one boundary of the region while the power line determines the other boundary. Figure 9 shows the contours of constant power and constant frequency in this plane for the case where the power is comprised *only of the dynamic component.*

There are two shaded regions, indicated as region A and region B. In region A, the system guarantees a frequency of operation of 400MHz and does not consume more than 10W of power. Similarly region B guarantees a power budget of 5W and a frequency budget of 25MHz. The feasible regions of operation are relatively large in Figure 9 since only dynamic power is considered.

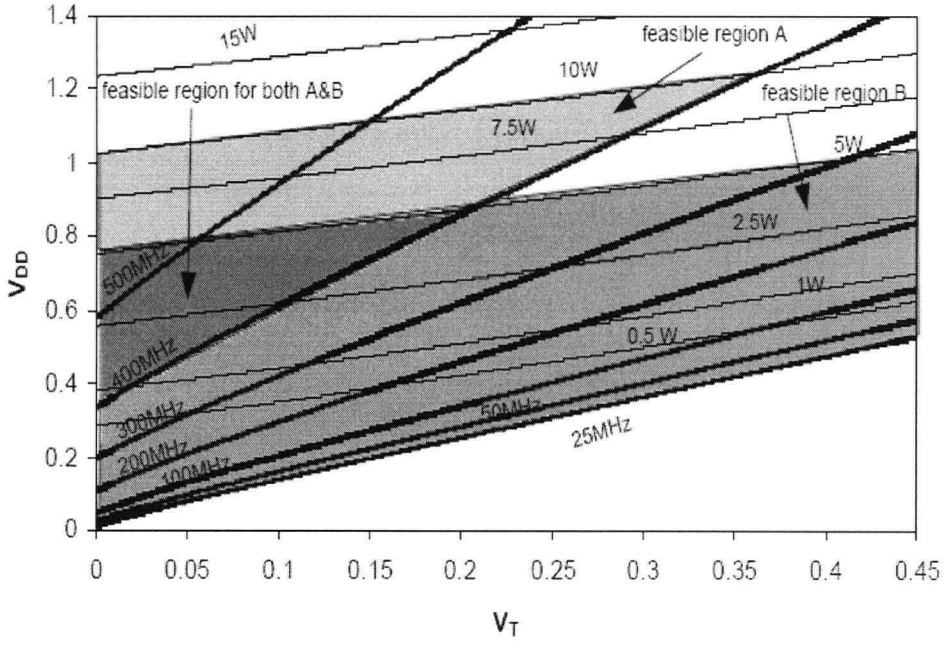


Figure 9. Dynamic Power vs. Frequency.

The same set of plots can be generated when *static power due to leakage is introduced.* Figure 10 shows constant value contours of both total power (static and dynamic) and frequency on the V_{DD} vs. V_T plane. The constant frequency contours are again linear whereas the constant power contours are different in this case. Due to the effect of leakage power, the lines of total power have become "S" shaped. The exponential nature of the power curves in the lower V_T regions is due to the dominance of the leakage power.

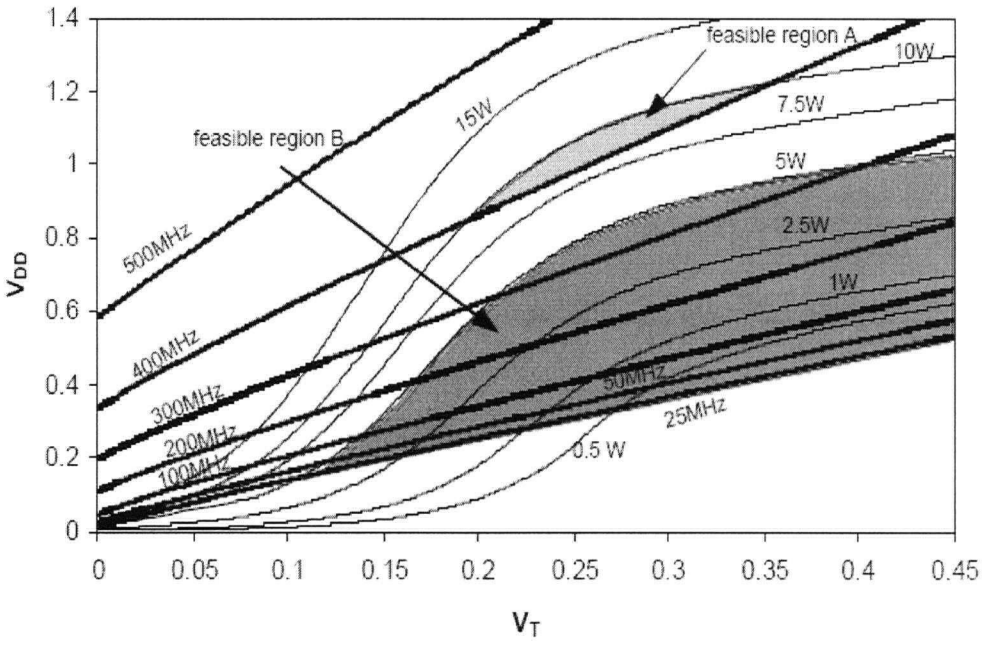


Figure 10. Total Power vs. Frequency.

Comparing Figures 9 and 10, if the system has only dynamic power, then the feasible region of operation is much larger than in the case where leakage power is significant. From the designer's perspective, the region of feasible (V_{DD}, V_T) pairs is shrinking with increasing leakage.

Given the power and frequency contours, the next step is to find the optimal operating point in this plane. From previous discussion, the two boundaries of a design space are set by power and frequency specifications. If the power budget is specified then optimality would be achieved when a certain (V_{DD}, V_T) pair is chosen, such that maximum frequency of operation is realized while corresponding power budget is not violated. For example, if the allowable maximum power dissipation is 10W, the V_{DD} and V_T may be chosen anywhere on the line (Fig. 9) or "S" shaped curve (Fig. 10) labeled as "10W". But there exists only one optimal point on

that line or curve, which provides the maximum frequency possible at the 10W power specification.

In Figure 9, the point (1.0V, 0V) lies on the Y-axis while for Figure 10 it is (1.0V, 0.27V). Similarly, for a given frequency specification, the optimal (V_{DD}, V_T) pair would provide the required frequency of operation but consume the minimum power. For example, if the frequency specification is 400MHz, the V_{DD} and V_T may be chosen anywhere on the line (Fig. 9 and Fig 10) labeled as "400MHz". Again there is a single optimal point for which the design would consume minimum power but operate at 400MHZ. That point in Figure 9 is (0.37V, 0V), which also lies on the Y-axis, and in Figure 10 is (0.98V, 0.26V). For different power or frequency specification, the optimal operating points for both cases can be found. Interestingly, all these optimal operating points define a single line, the "*optimal operating line*". Such an optimal line for the later case is shown in Figure 11.

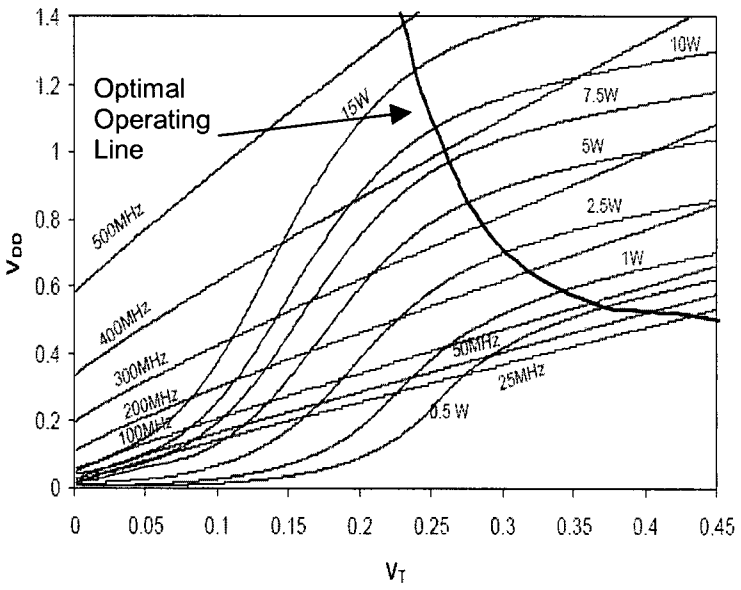


Figure 11. Optimal Operating Line.

When the (V_{DD}, V_T) pairs are chosen from this line, power and frequency optimality is achieved simultaneously. If the design operates with (V_{DD}, V_T) pairs chosen from any point other than the optimal operating line, then it is sub-optimal.

4.2 Optimal Points of Operation

Another perspective can be obtained by examining optimality from a metric point of view. Consider first the contours of constant EDP, PDP and PEP on the same plane. This is shown in Figures 12, 13 and 14.

The contour values decrease in value towards the center in each case. The optimal point is indicated with a "*", except for the PEP case where the optimal is somewhere to the right of the graph. Thus when the design operates on the points marked by "*", optimality in terms of a given metric is achieved.

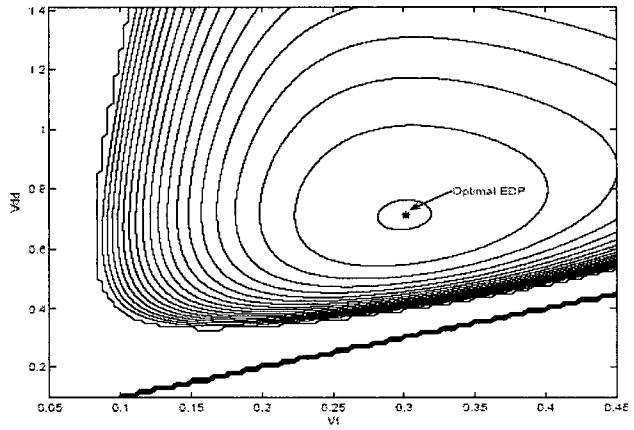


Figure 12. Energy-Delay Product.

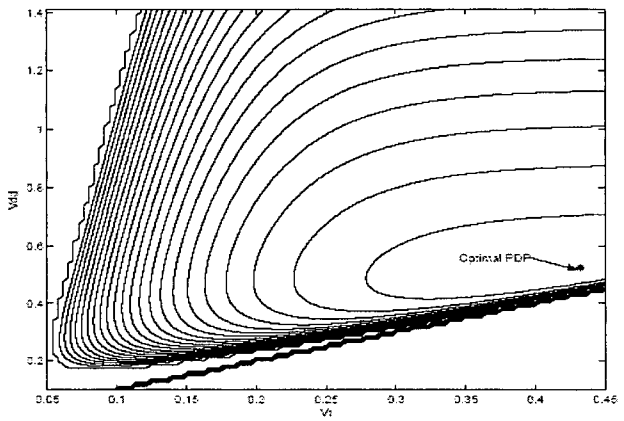


Figure 13. Power-Delay Product.

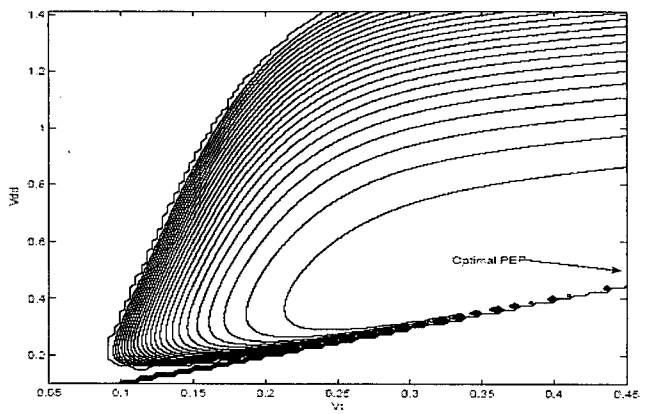


Figure 14. Power-Energy Product.

As discussed previously, each of these metrics have some relation to optimality in terms of power and delay. In fact, there exists a relation between the metric optimality and the optimal operating line that was derived in the previous section. To illustrate this, Figure 11 was superimposed on Figures 12, 13 and 14 to derive Figure 15.

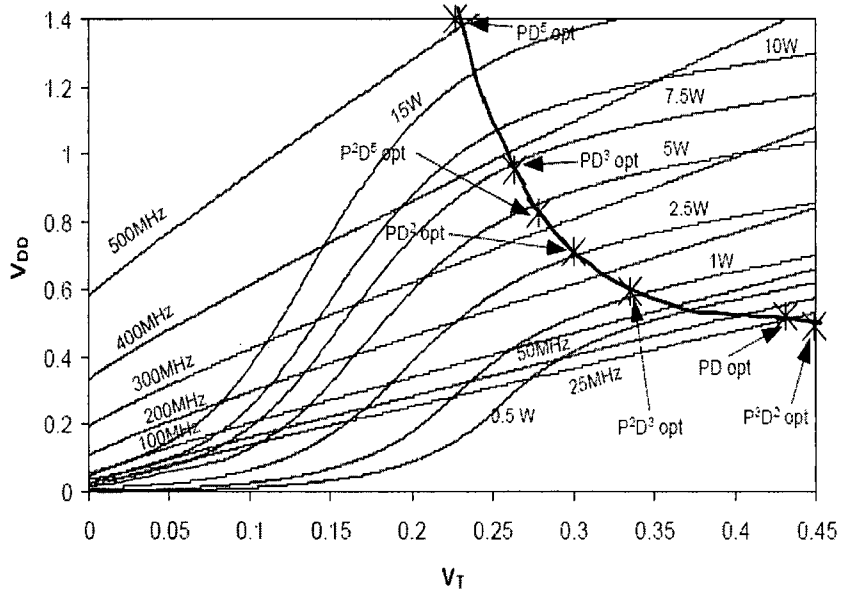


Figure 15. Metric Optimality and Optimal Operating Line.

Surprisingly, the optimal values of the PDP and EDP metrics lie on this optimal operating line. More generally, the optimal operating line is essentially the curve defining the *optimal values of all metrics* of the form P^mD^n. To provide evidence of this, some of the optimal values of other metrics have been plotted on the same graph. Moving up along the curve provides optimal values for metrics that provide greater emphasis on delay, which is reflected by the faster increase of n than m in the metrics. On the other hand, moving down along the curve places more emphasis on the power, which is evident from faster increase in m over n in the metrics.

Given that this optimal operating line defines the optimal values of the metrics, they can be correlated to the feasible operating regions. Figure 15 is superimposed on Figure 10 to produce Figure 16. It is observed that the optimal operating line passes through both the feasible regions. For optimal operation, the choice of the (V_{DD}, V_T) pairs is defined by two boundaries of the feasible region of operation. Though the design can operate with any (V_{DD}, V_T) pair within the feasible regions, for optimal operation the solution should lie on the optimal operating line.

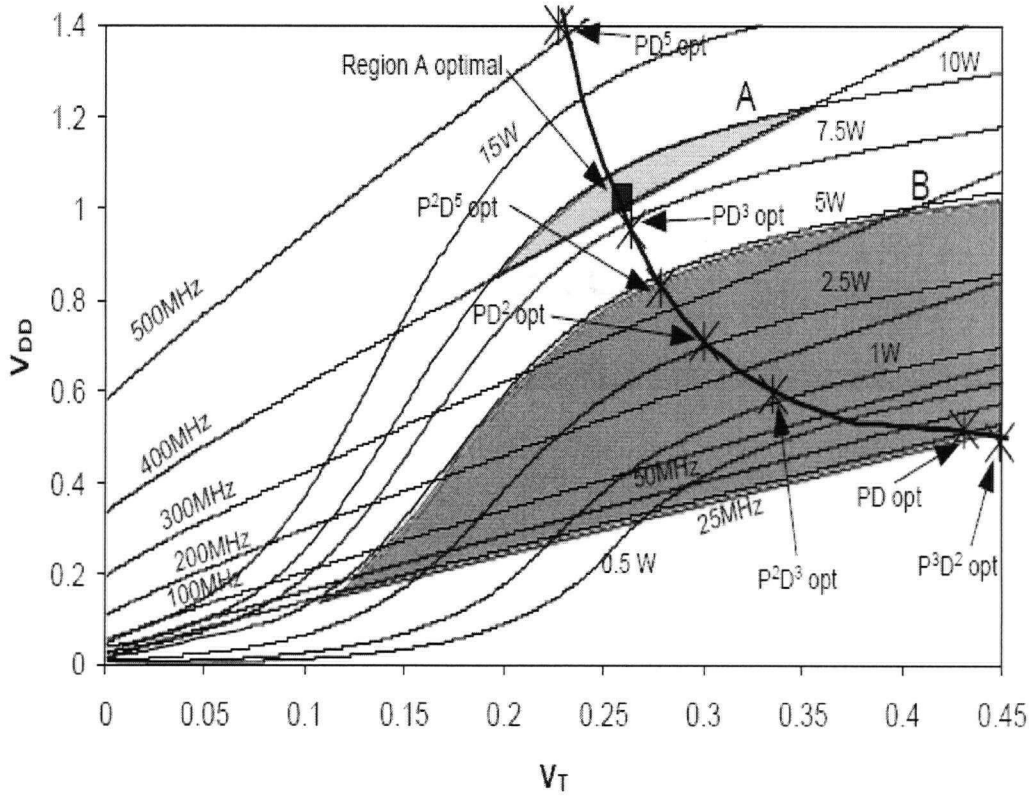


Figure 16. Feasible Region and Optimal Operating Line.

As shown in Figure 16, not all metric optimal values achieve their optimal solution in each feasible region. For example, the EDP, PDP or PEP optimality is unachievable in region A. In fact, region A has the same optimal point for all three metrics indicated by the square symbol.

In case of region B, the system has a larger operating line of operation. As such, it encompasses the optimal points of both EDP and PDP, again marked by "*" in Figures 12 and 13. As the optimal value of PEP does not lie in feasible region B, the optimal PEP is simply the bottom, right-hand corner of region B where the optimal operating line intersects the feasible region. So depending upon the metric, a different value for (V_{DD}, V_T) is produced.

Summarizing, there is an interesting and perhaps fundamental relationship between the optimal frequency-power curve and the optimal operating line, which is defined by the optimal operating points of general metrics of the form P^mD^n.

Chapter 5

Process, Temperature and Voltage (PVT) Effects

In the analysis of the previous chapters, it was assumed that the supply and threshold voltage can be set to precise values. But in reality there are uncertainties that move the operating point away from ideal settings. Parameter variations, especially *intra-chip variations*, pose a major challenge in the design optimization of high performance VLSI circuits, especially for sub-100nm technologies [44]. These intra-chip variations arise either from temperature variations (*T*) and supply voltage variations *(V)* or from process variations (*P*). This produces an uncertainty in the power and frequency [45]. Moreover circuits that use DVS and DVTS also experience supply and threshold voltage variations. Process, temperature and voltage variations are individually analyzed and their effect on the power and frequency contour is investigated. Based on these contours, the effect on the optimal operating line as well as the metrics is further looked into. As a case study, the 0.18μm CMOS technology is used and HSPICE simulation performed on the 8-bit adder to illustrate the effects.

5.1 Process Variation

As designers move into the nanometer regime, process variation is an important priority in manufacturing for the proper operation of a circuit. Process variability has a huge impact on the power consumption of the chip. Due to transistor parameter variation there can be 20X variation in the leakage power and 30% variation in chip operating frequency [44]. Typically yield is calculated by characterizing chips by their frequency specification. Recently, it was found that out of the good chips that meet the frequency specification, a significant percentage have to be discarded due to unacceptable power dissipation [46]. Both saturation as well as leakage current fluctuation may be attributed to the variation in channel length and variation in threshold voltage. The threshold voltage fluctuations are mainly due to:

1) random fluctuation of dopants underneath the gate, and
2) variation of gate channel length due to V_T roll off due to short channel effects [47].

In addition to threshold voltage fluctuation, body biasing also causes threshold voltage fluctuation. Typically body bias voltage is considered to vary about 20% of its nominal value [47]. Given that the circuit is robust to such fluctuation, the power and delay contours will shift with the threshold voltage variations causing a shift in the optimal operating line. This leads to an *optimal operating region* in place of a single *optimal operating line*. To justify this claim, HSPICE simulations were performed on an 8-bit adder for two extreme process corners and noted the shift in the optimal operating line is shown in Figure 17.

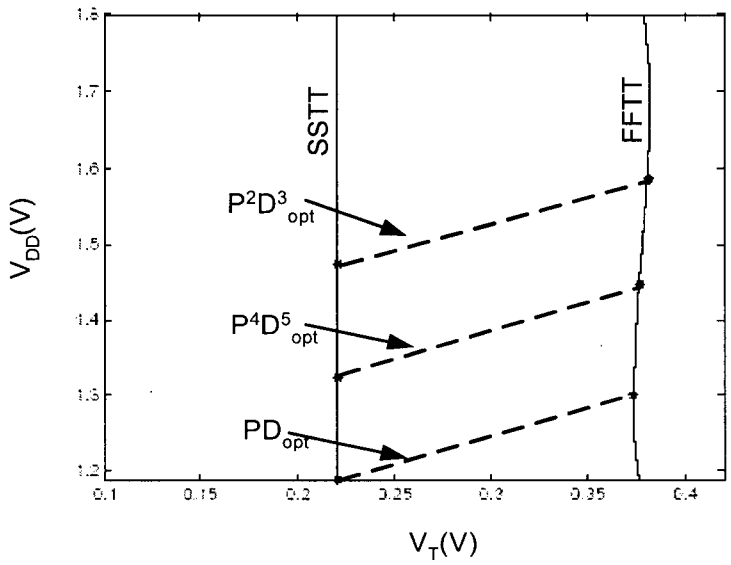


Figure 17. Optimal Operating Line shifts with Process Variation (HSPICE).

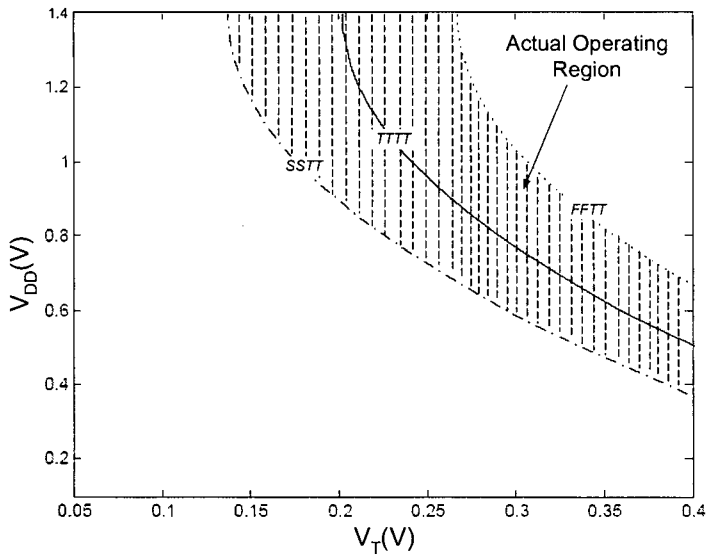


Figure 18. Optimal Operating Region with Process Variation (Analytical).

The region between the two process corners defines the optimal operating region for the adder circuit. The variations of the metrics have also been plotted here (where S = Slow, T = Typical and F = Fast; a process corner is defined in the order: NMOS, PMOS, V_{DD}, Temperature).

Based on the above results, the percentage of variation of the optimal line with process change was calculated and mapped to the 90nm technology node. In Figure 18, the *actual operating region* is shown. Thus, considering the existence of process variation within the chip, the design's optimal operating point would lie somewhere in the shaded operating region.

5.2 Temperature Variation

Temperature variation is another important issue for system design. The changes in performance vs. temperature will depend to some degree on the details of CMOS technology, since the MOSFET performance can improve as much as T^{-1} to $T^{-0.5}$ depending upon the process and operating electric field details [6]. In previous analysis, the temperature of the system was implicitly considered to be constant for the entire V_{DD} vs. V_T plane. But in reality the thermal effect causes shift in the operating point of the system. Temperature variation between 0°C to 100°C is considered here. Since the threshold voltage changes by −0.8 mV/°C [13], the 100°C variation in T introduces approximately 80mV of threshold voltage variation. This would cause the effect of temperature to be more pronounced in the leakage power than the dynamic power. To understand the effect of temperature variation on the optimal line, HSPICE simulation on the 8-bit adder with varying temperature of the circuit was performed.

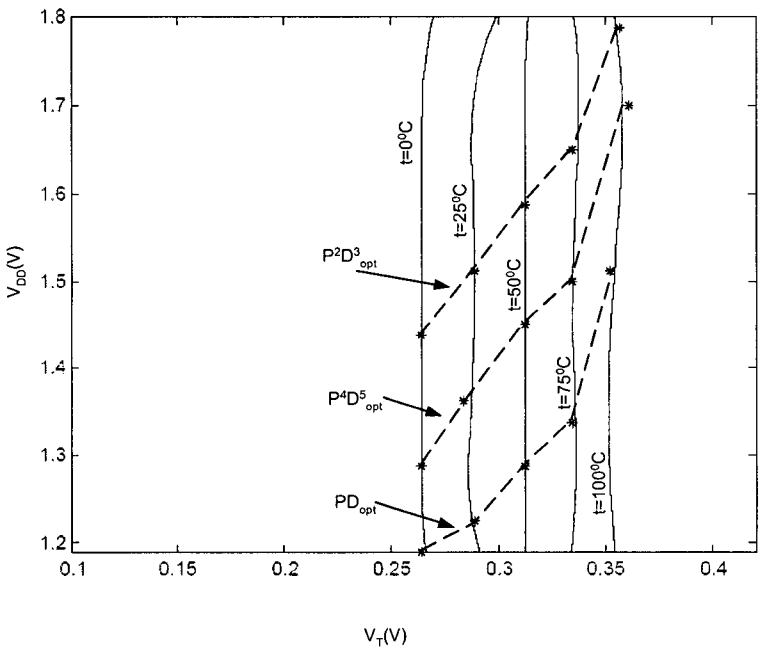


Figure 19. Optimal Operating Line Shift of 8-bit Adder with Temperature Variation (HSPICE).

From Figure 19 it is seen that the optimal operating line moves towards higher threshold voltage with higher temperature. Based on the above results, the model was extended to incorporate the thermal effect on the chip. Figure 20 shows how the optimal operating line shifts with the change in temperature.

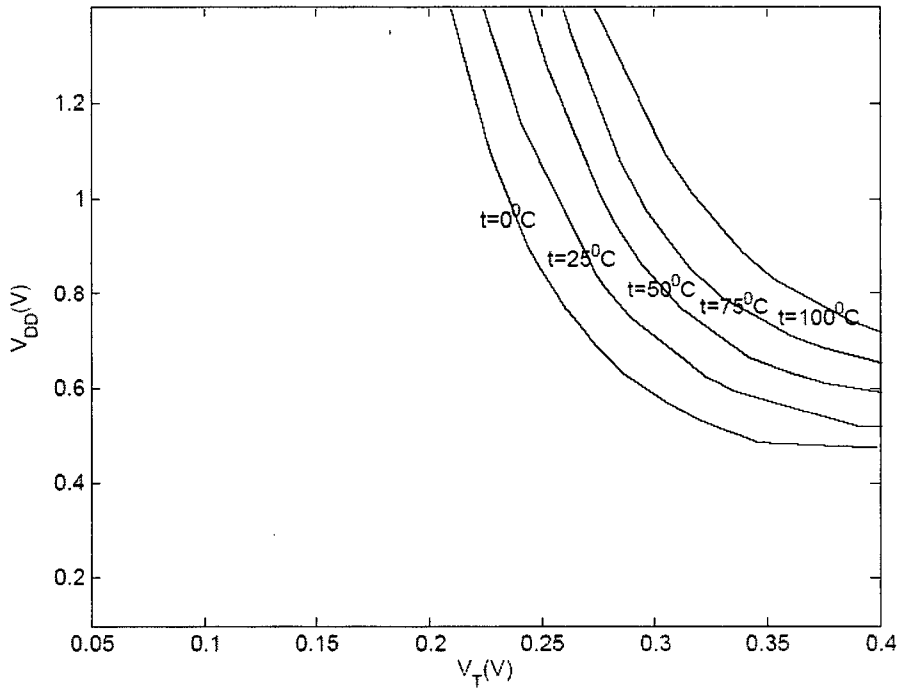


Figure 20. Optimal Operating Region shift of a Chip with Temperature Variation (Analytical).

Temperature increase causes higher leakage power consumption thereby having more influence on leakage power than on dynamic power within the V_{DD} vs. V_T plane. For a given frequency specification, a higher temperature in the circuit would shift the optimal line towards higher (V_{DD}, V_T) values. This is because a higher V_T would reduce the leakage power but would also decrease the frequency of operation. An increase in the V_{DD} value would then cause the frequency to rise.

From analytical and HSPICE analysis, it is understood that temperature changes cause a shift in the optimal operating line of operation. Incorporating the effect of temperature would also define the *actual operating region* for the design rather than a fixed optimal operating line. In

Figure 21 the actual operating region of operation from temperature perspective is illustrated for 90nm CMOS.

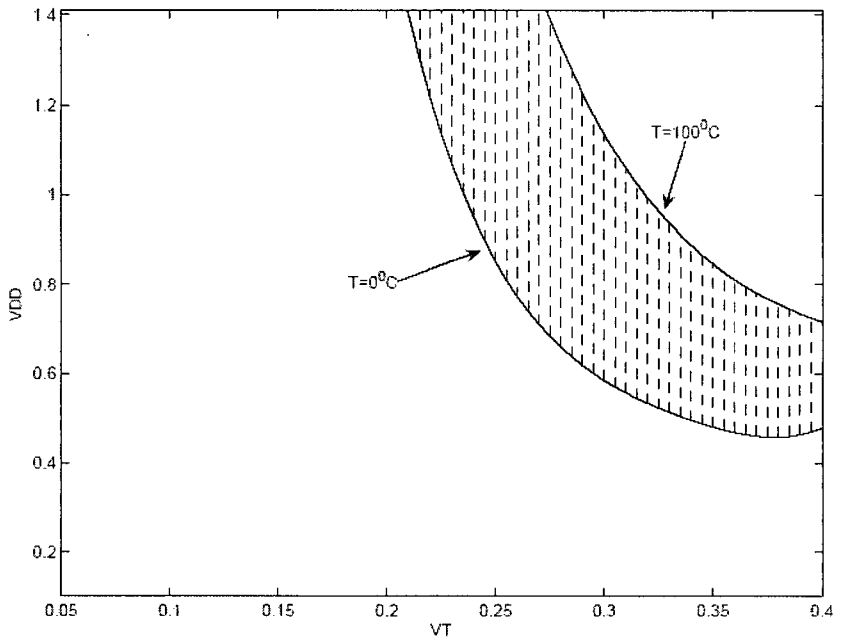


Figure 21. Optimal Operating Region due to Temperature Variation.

The design would ideally be operating within the region of optimal operating lines for the maximum and minimum temperature variation.

5.3 Voltage Variation

Power supply fluctuations across the design should also be taken into account to understand its effect on the optimal operating points. Maximum tolerable power consumption and reliability would determine the maximum allowable supply voltage while minimum performance specification to be met determines the minimum allowable supply voltage. In

dynamic voltage scalable systems, about 5-10% supply voltage variation may occur from the nominal supply voltage [48]. Thus there lies a window of supply voltage within which the design goals can be met. In other words, there is a spread in the optimal operating line over this window forming a band rather than a single line.

5.4 PVT Effects

From all the above analysis of PVT variation, it is evident that the circuit functions in a band in the *optimal operating region*. HSPICE result shows that the individual regions due to variation of only process or temperature or voltage overlap considerably. Given a frequency and power specification, the supply and threshold voltage would be defined by a rectangle in the V_{DD} vs. V_T plane within the optimal operating region whose width defines the threshold voltage variation, due to process variation, while the height would define the supply voltage variation. The supply and threshold voltage would be changing within that window only, for a fixed power and frequency specification. This has been depicted for optimal EDP of the design in Figure 22.

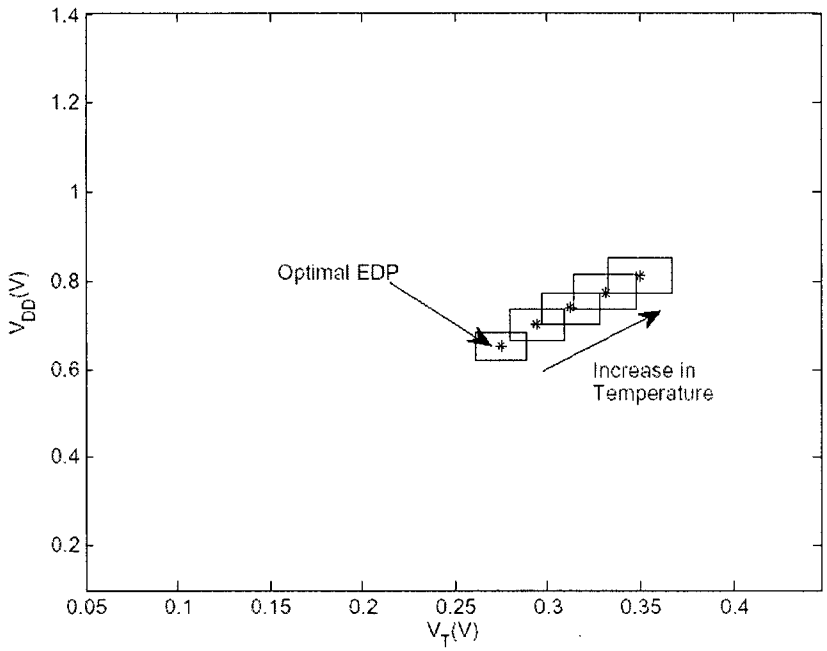


Figure 22. Optimal EDP Shift and Region of Operation.

Including the temperature variation, this rectangle would shift accordingly with the optimal operating region. This would then guarantee optimal power or frequency constraints.

In other words, the PVT variation replaces the notion of *optimal operating line* by *optimal operating region* and thereby provides a much more realistic design space from the optimality standpoint.

To translate these results to real design, power and delay modeling of logic blocks that vary with V_{DD} and V_T is needed. The next chapter shows how these models can be derived.

Chapter 6

Power and Delay Modeling

As discussed in Chapter 2, power estimation is an extremely important issue for making early decisions on design block optimization. Moreover delay modeling must be carried out to make an estimate of how fast a circuit can operate for some particular supply and threshold voltage value. In this chapter, techniques for both power and delay modeling are described where it is assumed that both supply and threshold voltage can be altered. A VHDL description of the circuit and a random data set is used for modeling purposes. A data-dependent dynamic power model, based on Hamming distance, is illustrated for an adder and a multiplexer. The method can be extended to other logic blocks and can be integrated in a simulation tool with other blocks to form a larger circuit. Moreover, using the proper choice of *Training Set*, the accuracy of the model can be improved. The advantages and limitations of the approach are also identified using the two examples. HSPICE simulation at the transistor level is used for static power modeling. A ring oscillator based delay model that captures the effects of V_{DD} and V_T variations is also described.

To understand the types of circuits for which power and delay are to be modeled, consider the simplified block diagram of a single cycle MIPS datapath and control unit in Figure 23 derived from [49]. The overall goal would be to determine the power dissipation of this design while executing a task defined by a software routine.

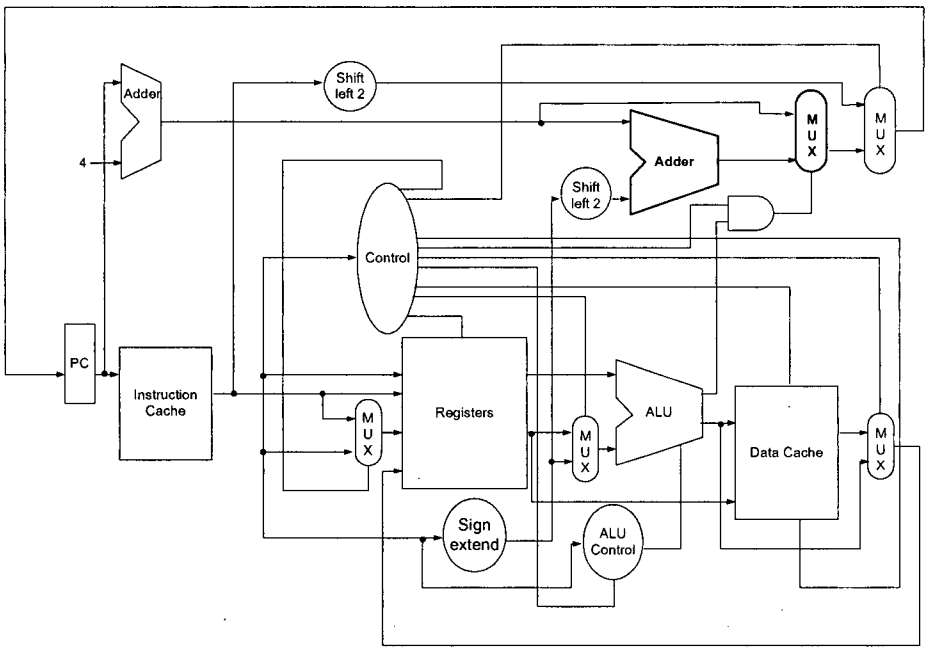


Figure 23. Single Cycle MIPS Datapath and Control Unit [49].

Typically the power consumption can be split into two subsections: the combinational logic such as adders, ALUs, controllers and multiplexers, and memory modules such as data cache, instruction cache and registers. Cache memory modeling for delay and energy consumption has been extensively analyzed in [50, 51, 52]. Their methods can be further extended in order to incorporate the V_{DD} and V_T parameters. Here, two of the basic logic blocks are considered for power modeling: the adder and the multiplexer. The adder circuit has been considered for delay modeling.

6.1 Power Modeling

In order to model the power of different logic blocks, the power components of the blocks are split into two parts, namely the dynamic component and the leakage component. First, dynamic power modeling is described, followed by leakage power modeling.

6.1.1 Hamming Distance Modeling

For dynamic power/energy modeling, the use of transition-based estimation has been quite a popular approach in the past [53, 54, 55]. As in Equation (1), the dynamic power can be expressed as:

$$P_{dynamic} = \alpha C V_{DD}^2 f_{clk}$$

The clock frequency (f_{clk}) is $1/T$, where T is the clock period as determined by the critical path delay of the circuit. As shown in Chapter 3, the energy is given by:

$$E_{dynamic} = \alpha C V_{DD}^2 \quad (14)$$

To compute the energy, reasonable estimates of the capacitance and toggle count for each node are needed [54]. For this purpose, the level of abstraction of a circuit may be at the block level, sub-block level or at the transistor level with the effort of computation increasing as one moves towards finer granularity of the design. Here, the Hamming distance based approach of the energy estimation has been used which is a data-driven approach. That is, the power dissipation depends on the input and/or the output data. The difference in the number of bits of two consecutive bit patterns is defined as the Hamming Distance. For example, if the input

vector to a circuit is "11100011" followed by "11111111" then the Hamming distance at the input would be 3.

In this modeling approach, the energy dissipation is directly correlated to the Hamming distance of the circuit's input/output in the form of equation as:

$$E_i = f(H(input/output_i, input/output_{i-1}))$$

where E_i is the energy consumed in the i^{th} cycle due to flipping of inputs/outputs from the previous $i\text{-}1^{st}$ stage to the i^{th} stage and $H()$ relates to Hamming distance. In this way, the total energy consumed by any logic block over n cycles can be computed as:

$$E_{total} = \sum_{i=1}^{n} E_i$$

Next the use of this approach for dynamic energy modeling for two circuits is illustrated.

For each block PrimePower™ simulation is carried out for various Hamming distances. A polynomial equation can be used to model the energy as a function of Hamming distance. Using curve fitting for different Hamming distance values, the coefficients can be extracted using MATLAB™ to generate the models. This provides a modeling method with certain bound of accuracy. While models may not be accurate for a small set of input or output patterns, it is expected to be more and more accurate as the number of patterns increase due to averaging effects.

6.1.2 Adder Example

An 8-bit ripple carry adder is considered as the first case study. Figure 24 shows the circuit. It has two 8-bit inputs, *A* and *B*, an 8-bit output *Sum*, a *Carryin* input and a *Carryout* output.

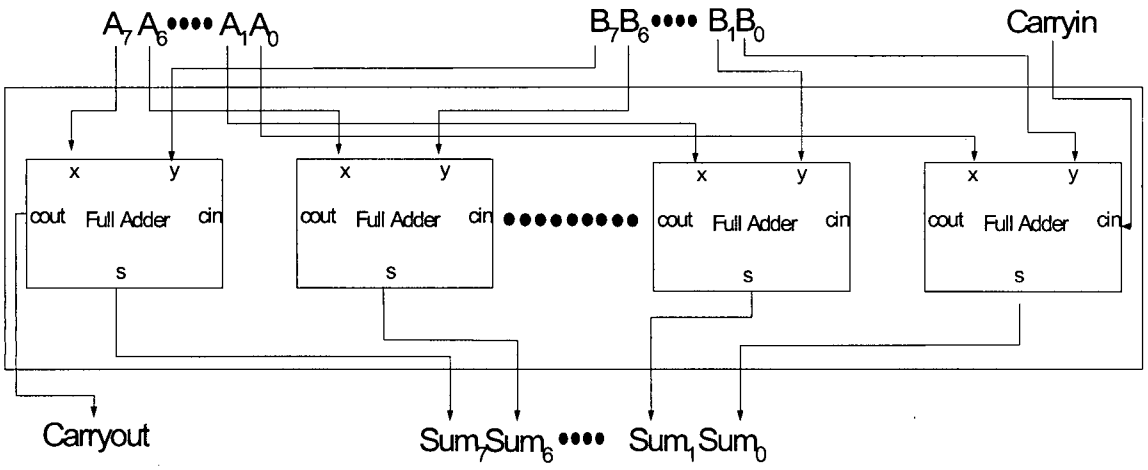


Figure 24. 8-bit Ripple Carry Adder.

Figure 25 shows the energy consumption of the adder based on the results of the PrimePower™ tool. The Hamming distance at the input is plotted on the X-axis and the energy is plotted on the Y-axis.

Each "*" on this graph corresponds to the energy consumption derived from simulation for a particular Hamming distance at the input of the adder. For simplicity, only the two 8-bit inputs are considered to be flipping while the carry input is at logic zero. As it is evident from the graph, for each Hamming distance, different amounts of energy may be consumed. Further, there is a rough trend of higher energy consumption for higher Hamming distance, which

reflects typically larger amounts of activity within the adder with larger number of bit flips at the input of the adder circuit.

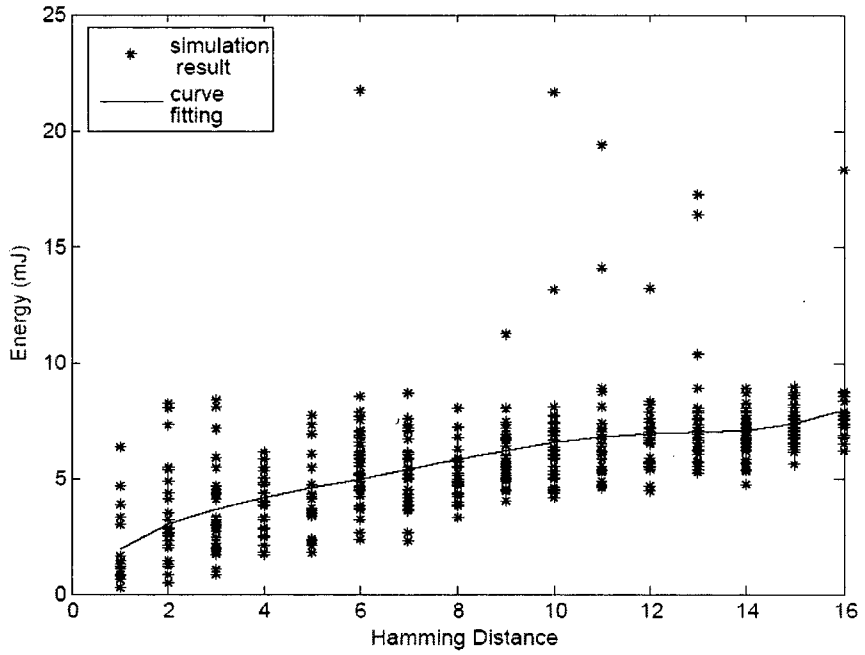


Figure 25. Curve Fitting of Energy based on Hamming Distance.

The energy consumption for a fixed Hamming distance number varies in case of an adder. To explain the reason for this let us consider two different cases. Consider all the inputs of the adder to be initially set to "0". Let only one bit of input *A* flip which corresponds to a Hamming distance of *one*. Consider another case where initially all the bits of input *A* are held to logic "1" state and all other inputs are held at "0" state. Next the LSB of input *B* is flipped to "1" state. According to the Hamming distance method, the energy expended should be equal to previous case with Hamming distance of *one*, as only one input bit has flipped. But, in reality, the energy consumed in this case will be significantly more than that the previous case. This is because the number of internal flips in the first case is less than the second case, where carry is propagated through the circuit.

As seen in Figure 25, most of the energy points are close together except for a few, which have much larger values than others for the same Hamming distance. These points reflect the internal flipping within the adder due to carry propagation. But, as discussed earlier through the example, these flips cannot be modeled using Hamming distance at the input of the adder. So, whenever inputs at the adder arrive with minimal Hamming distance but cause considerable flipping inside the adder, then that energy consumption goes unnoticed by this methodology, which is definitely a drawback. However, after a large number of cycles, the relative error due to the averaging is reduced.

One way to deal such a case is to consider the average energy consumed per Hamming distance value. The consequence of using an average is that sometimes there would be over estimation of energy while at other times there would be under-estimation of energy. For purely random data, it is found that these over-estimations and under-estimations cancel out, giving improved results from the model as compared to simulation results as discussed next.

To develop a Hamming distance based model of the energy consumption, first a small subset of vectors is chosen called the *Training Set*. With that subset, the energy consumption corresponding to each Hamming distance are gathered using PrimePower™ and its values are used for curve fitting to derive an equation of the form:

$$E_i = a_1 x_i + a_2 x^2{}_i + a_3 x^3{}_i + \ldots\ldots\ldots + a_k x^k{}_i \qquad (15)$$

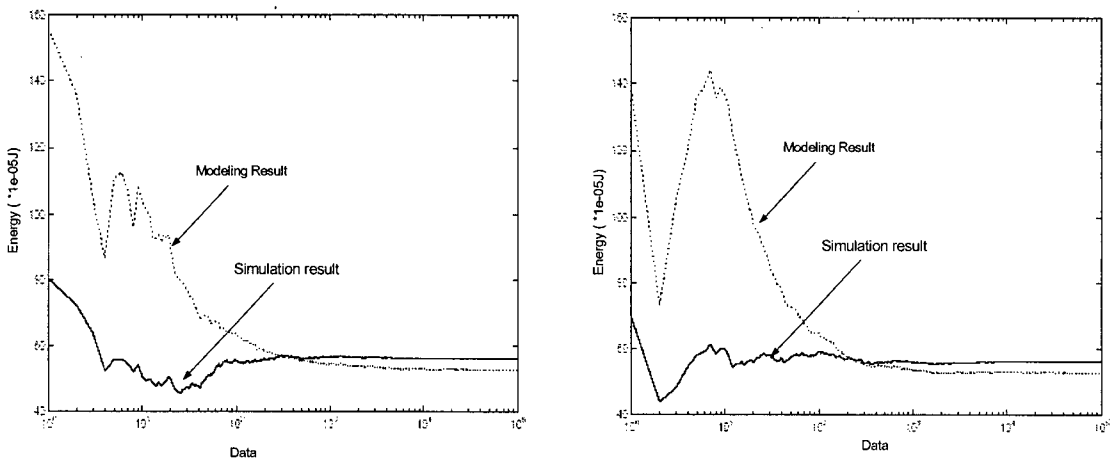
where E_i is the energy consumed in the i^{th} cycle, $a_1, a_2, \ldots .., a_k$ are the coefficients of the curve and x_i is the Hamming distance between $(input_{i-1})$ and $input_i$. In the Figure 25, 500 vectors were used as training set and the solid line is the curve-fitting model using Equation (15). This

equation can now model the energy consumption of larger data sets by computing the number of flips between two consecutive data inputs.

To understand the relative error in this data-driven approach the average energy consumption is considered. The average energy consumption $(E_{average})$ in the i^{th} cycle is defined as:

$$E_{average_i} = \frac{\sum_{j=0}^{j=i} E_j}{i} \tag{16}$$

where E_j is the energy consumed in the j^{th} cycle. Two randomly generated data sets of size 10^5 are used as a case study. First, simulation was performed using PrimePower™ and the average energy was computed as in Equation (16). Based on the input data and the curve-fitting model (as described above) estimate of the energy consumption were made, followed by average energy calculation using Equation (16). Figure 26 shows how the actual average energy values and the modeled average energy values track. The number of random data used for simulation is plotted on the X-axis and the Y-axis defines the cumulative average energy consumption.



(a) **(b)**

Figure 26. Comparison of Model and Simulation Result.

One important point to emphasize here is that the choice of the number of vectors in the *Training Set* is crucial to the level of accuracy that can be achieved. In the above case, 500 vectors have been used to derive the equation. With the increase of number of vectors used, the relative accuracy is improved. Figure 27 shows the comparison of the average energy estimation for the same set of simulation vectors but different curve fitting equation generated from 500, 1000, 2000 and 5000 training set vectors. Based on these results, it is clear that larger training sets render higher accuracy.

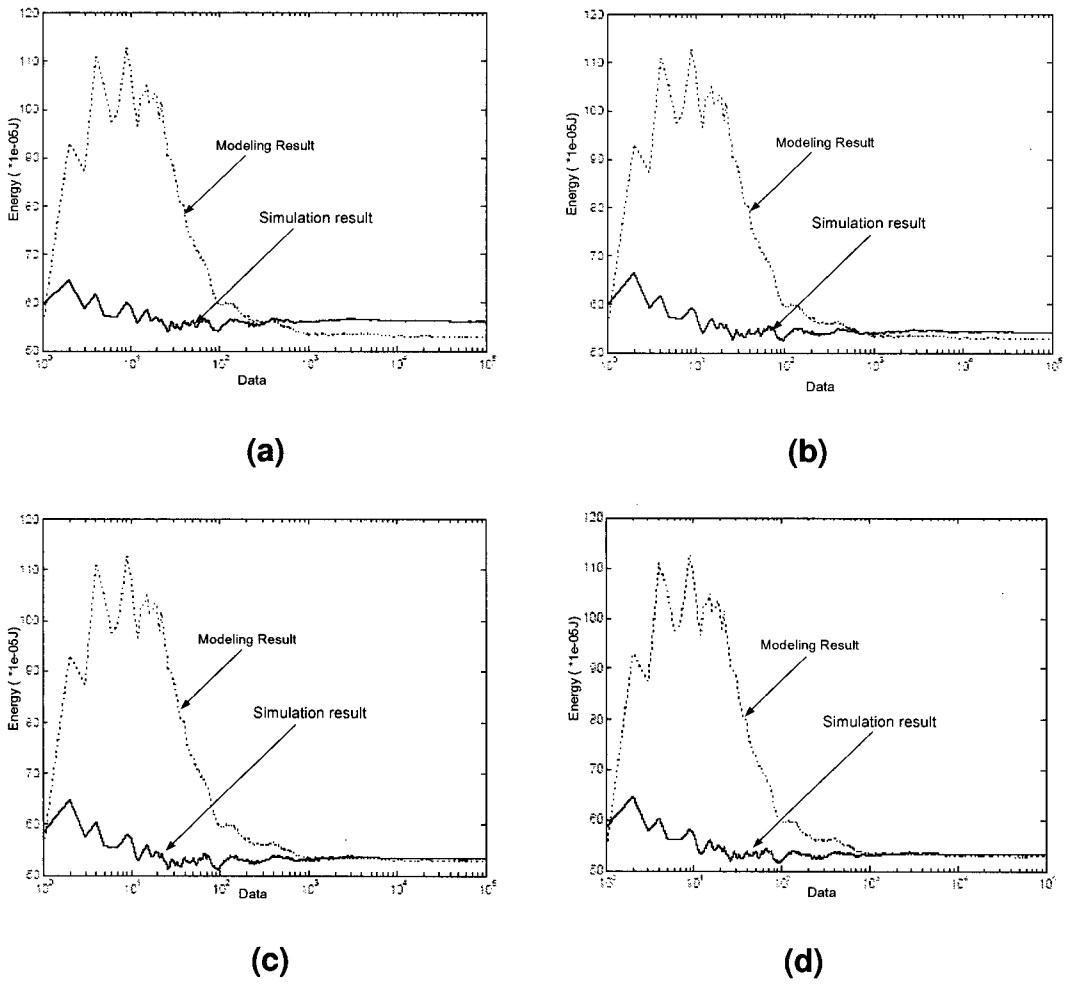


Figure 27. Average Energy Modeling (a) Training set=500, (b) Training set=1000, (c) Training set=2000, (d) Training set =50000.

Such an analysis should be performed for logic blocks like subtractors, multipliers, etc., where same Hamming distance produces different amount of energy dissipation. As long as a large enough Training Set is used in each case, it is expected that the average energy levels will be accurate as the number of execution cycles of the processor increases.

6.1.3 Multiplexer Example

There is another class of logic where a fixed amount of energy is always dissipated corresponding to Hamming distance. To illustrate this simpler case, consider a 2:1 multiplexer which has two 8-bit data input and an 8-bit data output with a one bit select input as shown in Figure 28. The inputs I_0, I_1 and the output O are 8-bits and the select input S is single bit.

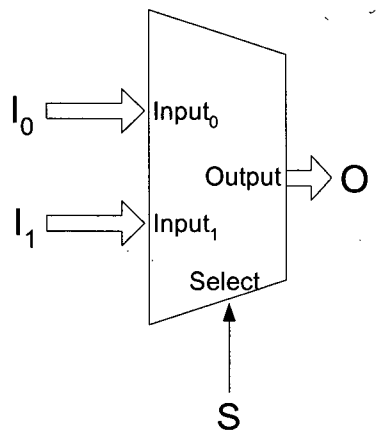


Figure 28. 2:1 Multiplexer.

The energy consumption for the multiplexer not only depends upon the data inputs but also on the Select input. There are two cases to be considered: if the selected data changes, and if the data input not selected changes. Energy consumption is found to be minimal if the data input that is not selected flips. To track the energy dissipation, unlike the adder case, here the

output flip and the select flip are considered. The methodology of computing the energy consumption is similar to the adder. The only difference is that the energy dissipation for flipping at the output for a single bit flip as well as that of select bit flip is to be simulated. That would be sufficient to model the energy dissipation of the multiplexer. In Figure 29, the actual energy dissipation of multiplexer versus the Hamming distance at the output as well as select flip is shown. The Hamming distance is plotted on the X-axis and the corresponding energy consumption on the Y-axis.

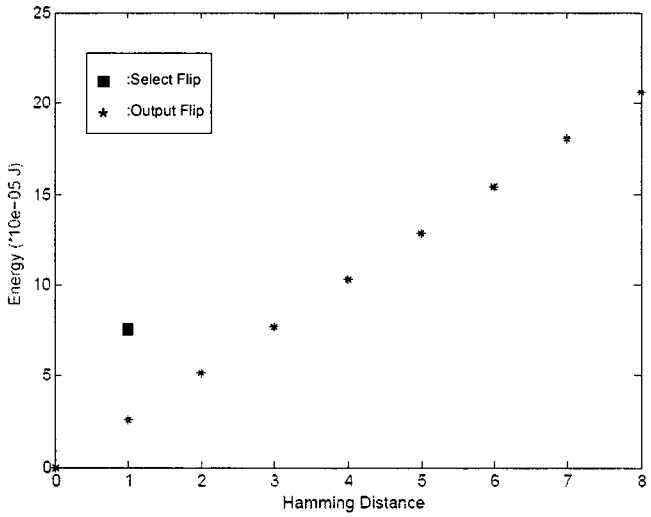
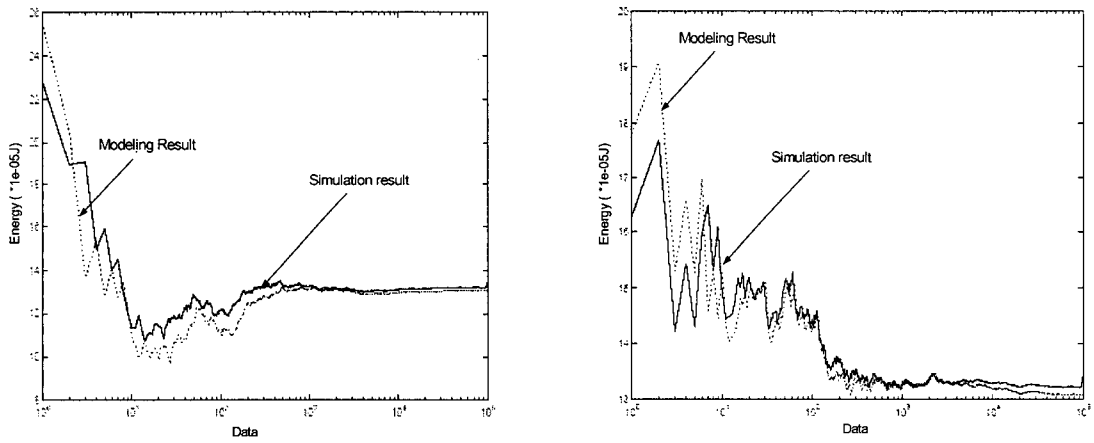


Figure 29. Hamming Distance versus Energy Consumption of 2:1 Multiplexer.

The energy consumption for each Hamming distance at the output as well as the select flip is noted for the purpose of modeling. For two different randomly generated data sets, the energy consumptions were noted using PrimePower™ and the average energy consumptions were computed using Equation (16). With the same sets of data, and by noting the output and select flip at each cycle, the energy consumption was modeled (using values from Figure 29) and average energy consumption was also computed. In the Figure 30, the simulated result is

compared against the modeled result. The average energy consumption is plotted in Y-axis and the number of data in the X-axis.



(a) (b)

Figure 30. Comparison of Hand Calculation and Simulation Result of Average Energy Consumption.

It is clear from the results that the average energy consumption of the multiplexer is in much closer agreement with the simulation result. Such a data-driven approach can be used for circuits such as de-multiplexer, decoders, etc.

This method of active power modeling has some drawbacks. It was observed that the energy dissipation for any fixed Hamming distance *may or may not* be always equal. In other words, different amounts of energy dissipation may be observed for the same Hamming distance. In other cases, fixed energy dissipation is always observed for a fixed Hamming distance value. A ripple-carry adder is an example of the previous case while a multiplexer is the example of the later case. The primary reason of having different energy consumption values is that the Hamming distance based method is not always sufficient to model the internal switching that

happens in a given circuit. That is, the Hamming distance at input or output may not always reflect the capacitance switching within the circuit properly. However, the averaging effect due to a larger number of cycles act to reduce these errors.

Based on the above discussion, it is evident that dynamic energy modeling for logic circuits can be done in two different ways. In the first case, better modeling can be performed with a greater accuracy of the training set data. For example, if a certain known pattern of input data is more likely to occur, then using a subset of such data in the training set would give a greater accuracy in terms of modeling. The methods of dynamic energy consumption modeling as well as their combination can be extended to derive a data-driven model of an entire processor core and be analyzed in Simplescalar using the resulting models. These models should also be a function of V_{DD} (Equation (14)) and thus is capable of modeling the change in energy consumption due to DVS. This is a simple matter of using the appropriate V_{DD}^2 term in the energy calculation. However, it is somewhat more complicated in the static power case. Discussion on leakage power modeling follows in the next section.

6.1.4 Leakage Power Modeling

In order to capture the effects of V_{DD} and V_T on leakage power, the logic circuits can first be simulated in HSPICE. As in Equation (2) and (6), the leakage power may be modeled as:

$$P_{leakage} = I_{leakage} \times V_{DD} = I_0 e^{\frac{-V_T}{nV_{th}}} \times V_{DD} \quad (17)$$

The leakage current can be measured in HSPICE for the circuit. The change of leakage current for different V_{DD} can also be noted by changing the supply voltage of the circuit in HSPICE. The body bias voltage is used to change the threshold voltage of all the transistors of the circuit. The 8-bit adder is considered here for illustrative purposes. HSPICE simulations were performed while varying the threshold voltage for some fixed supply voltages. The inputs to the adder were all kept at "0" (to have a non-active circuit) and the leakage current was noted. It was then used to fit Equation (17) to model for leakage power for different supply and threshold voltages. Figure 31 shows the comparison of the HSPICE simulation result vs. modeling result for four different supply voltages at 0.18μm technology. The threshold voltage is plotted on the X-axis and the leakage power on the Y-axis. The model results are fairly close to the simulation result. This methodology can be used and extended for any (V_{DD}, V_T) pair and for any other circuit.

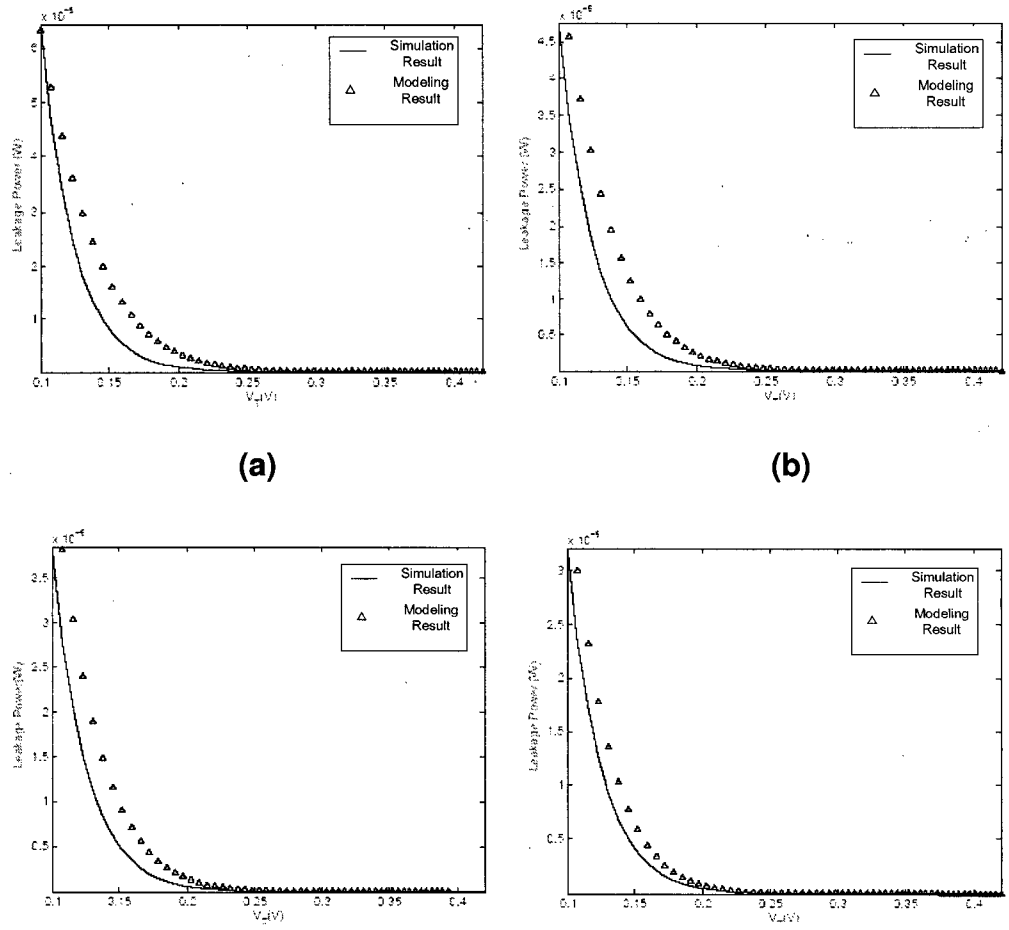


Figure 31. Comparison of Modeling Result and Simulation Result of Leakage Power, (a) V_{DD}=1.8V, (b) V_{DD}=1.65V, (c) V_{DD}=1.5V, (d) V_{DD}=1.35V.

6.2 Delay Modeling

Delay models are needed since the critical path timing will be a function of V_{DD} and V_T, which may vary during the analysis. Therefore, enhancements to the existing models are needed to incorporate these effects. The goal of the delay model is to represent the critical path of each block in some way as to capture the effects of V_{DD} and V_T variations accurately.

6.2.1 Ring Oscillator based Delay

A chain of inverters can be used to model the critical path in a circuit. However, to provide a representative loading on the last inverter, its output can be fed back to the first inverter to form a ring oscillator. It must have an odd number of stages to induce oscillation. Hence, delay modeling can be done using ring oscillator as shown in Figure 32.

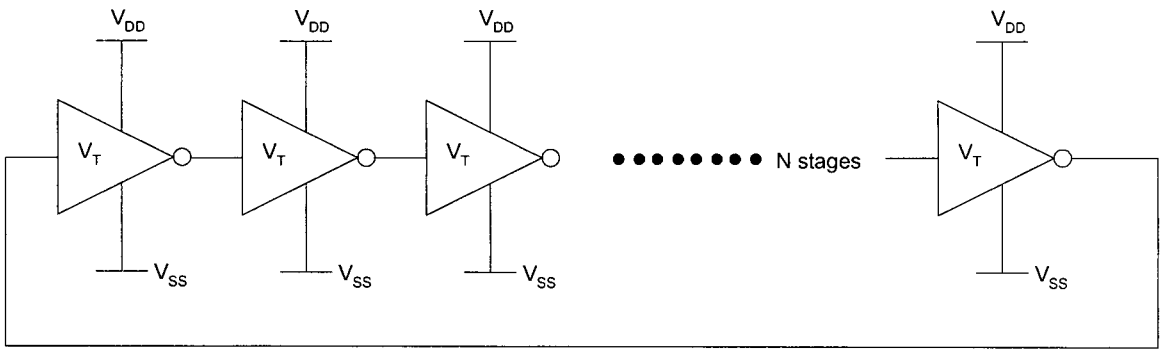


Figure 32. N-Stage Ring Oscillator.

The steps involved in modeling the delay of the circuits are as follows. The circuit's critical path is first identified and abstracted, and the vectors that would cause this critical delay are used to measure the delay of the circuit with HSPICE. Initially the threshold voltage is kept

constant and the supply voltage is varied to determine the delay variation. Next, a ring oscillator with N stages is used to model the delay of the critical path (where PMOS of each inverter is 2X in width compared to the NMOS and minimum size inverters are used for modeling). Once the path delay of the arbitrary circuit is modeled via an N-stage ring oscillator, delay variations due to V_{DD} and V_T are based on the ring oscillator only.

For a single inverter with load capacitance C_L, the delay for a ramp input can be expressed as [14]:

$$\tau_{inv} = \frac{t_r}{4} + \frac{C_L V_{DD}/2}{I_{Dsat}} \tag{18}$$

where t_r is the rise/fall time, I_{Dsat} is the saturation current of the NMOS or the PMOS depending upon the signal transition. The delay of the N-stage ring oscillator can be expressed as:

$$f = \frac{1}{2N\tau_{inv}} \tag{19}$$

where f is the oscillation frequency. The selection of N is important in determining the overall accuracy, as described below.

Let us first consider the last term of Equation (18). It is dependent on the saturation current. The saturation current is a direct function of supply and threshold voltages and used to compute the propagation delay. The saturation current here is the average of the NMOS and the PMOS saturation currents. The rise time/fall time delay is determined using HSPICE and modeled as a function of supply voltage. In this fashion, the delay of the ring oscillator is able

to capture the variations due to changes in V_{DD} and for a fixed value of V_T, as long as proper N is used.

As an example the 8-bit adder circuit of Figure 24 is considered. The critical path of the adder is from the *Carryin* input to the *Carryout* output. The threshold voltage is first fixed at 0.45V and HSPICE simulation was performed for 0.18μm technology. The delay is observed with supply voltage varying from 1.8V to 0.9V. The delay of the adder was then modeled with ring oscillator and a 65-stage ring oscillator was found to model the delay of adder quite accurately at the specified threshold voltage. Once the HSPICE result of the ring oscillator was obtained, this delay was modeled using Equations (18) and (19). Figure 33 depicts the delay of the adder, the ring oscillator in HSPICE as well as the model. V_{DD} is plotted on the X-axis and delay on the Y-axis.

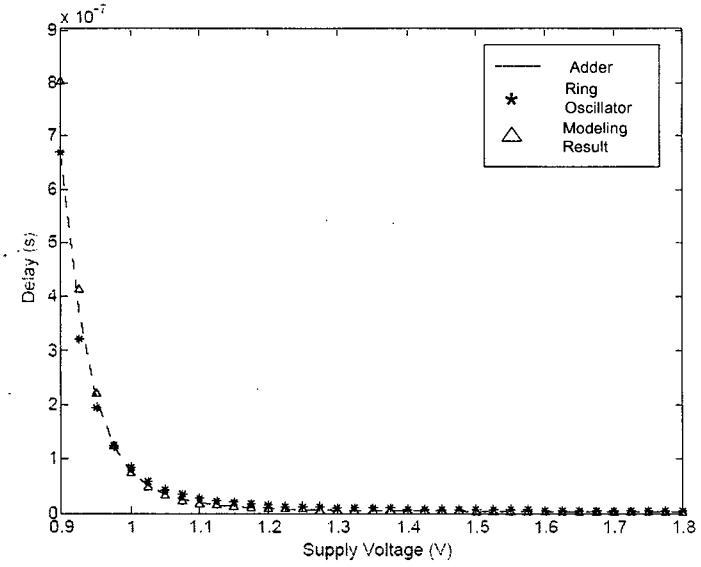


Figure 33. Comparison of Delay of Adder, Ring Oscillator and Hand Modeling.

The figure shows how the model tracks the ring oscillator, which again tracks the delay in the adder circuit. However, the accuracy of the delay modeling actually depends on the number of stages used. In fact, the *number of stages of ring oscillator* needed for accurate results is *different* for *different threshold voltage* values. To illustrate this point, consider the 8-bit adder example with threshold voltage changed to 0.1V. The delay of the 65-stage ring oscillator (*) as well as that of the adder (-) is plotted in Figure 34. The supply voltage is scaled between 1.8V and 0.9V. It is quite clear that the delay of the 65-stage ring oscillator is much larger than the adder. In fact, for V_T=0.1V, a 25-stage ring oscillator models the delay of the adder better than the 65–stage ring oscillator. The delay of the 25-stage ring oscillator (□) is also plotted in the same figure.

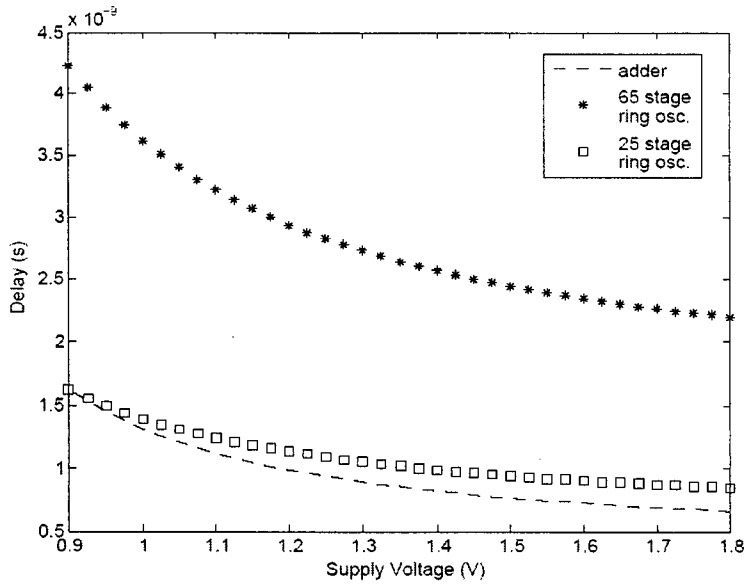


Figure 34. Comparison of Delay of Adder, 65 and 25-Stage Ring Oscillator.

To accurately capture the delay of the adder for different threshold voltages, the number of stages of the ring oscillator needs to be changed. The principle reason behind the need for a variable stage ring oscillator is due to the large difference in the delay when V_T is shifted. To

compare the delay variation of the adder as compared to different stages of ring oscillator, consider that the supply voltage ranging between 1.8V to 0.9V and the threshold voltage ranging between 0.1V to 0.45V. Considering V_T=0.1V, 0.3V and 0.45V, the delay of the adder was found to be modeled accurately using 25-stage, 41-stage and 65-stage ring oscillator respectively. To illustrate that none of these ring oscillator stages can accurately model the delay of the adder in the entire V_{DD} and V_T plane two extreme points in this plane are considered. One of these points is where delay would be maximum, V_{DD} = 0.9V and V_T=0.45V, and the other is where delay would be minimum, V_{DD} = 1.8V and V_T=0.1V (see Figure 36). For these two cases, the different delays for the adder, a 65-stage ring oscillator, a 41-stage ring oscillator and a 25-stage ring oscillator are shown in the Figure 35. It is found that none of the ring oscillators are suitable to cover both cases. For the minimum delay situation, the 25-stage ring oscillator is the best while for the maximum delay situation the 65-stage ring oscillator is the proper choice. The 41-stage ring oscillator is also not a suitable choice these two cases. However, it is better for some other intermediate supply and threshold value.

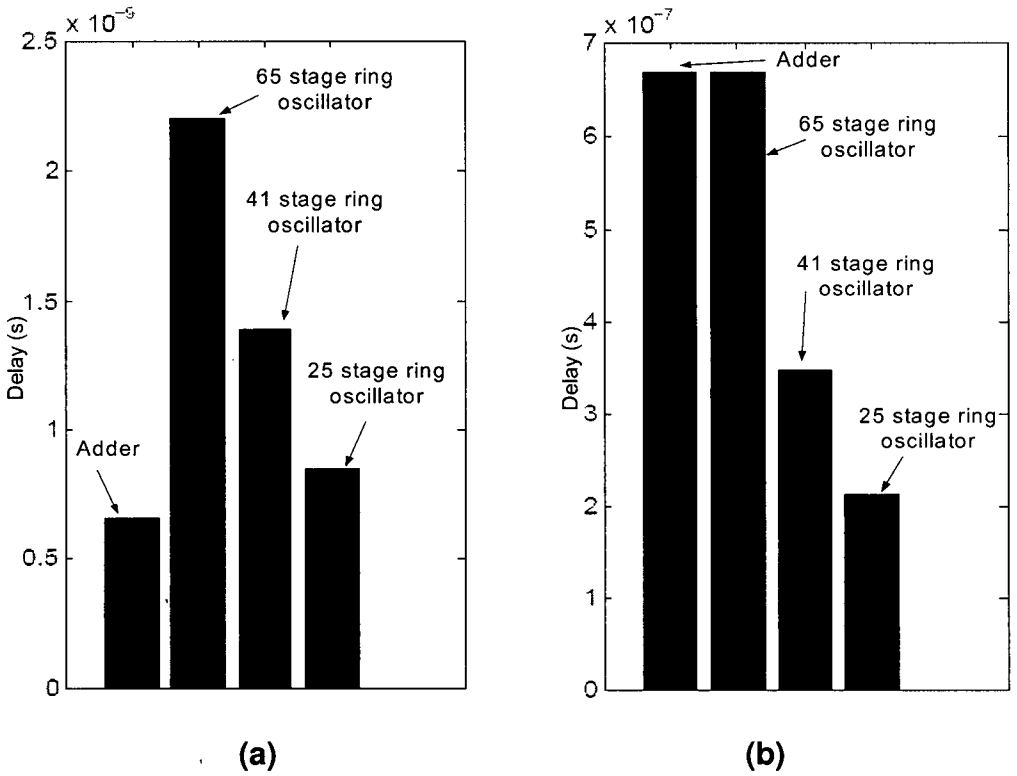


Figure 35. Comparison of Delay of Adder and Ring Oscillators for (a) V_{DD} = 1.8V, V_T=0.1V and (b) V_{DD}=0.9V, V_T=0.45V.

From a modeling perspective, each threshold voltage would require a different number of stages of the ring oscillator. As shown in Figure 33 and 34, *fixed* stage ring oscillators can accurately model delay of adder with *variable* supply voltage and *fixed* threshold voltage. Thus, the entire threshold voltage range needs to be compartmentalized with distinct values of N. Then, for each threshold voltage, the required N is to be calculated. Figure 36 depicts the modeling approach for delay of the adder. The number of ring oscillator stages needed is specified for 3 distinct cases of threshold voltage. Depending upon the accuracy required, *N* can be calculated for other threshold voltage values by interpolation.

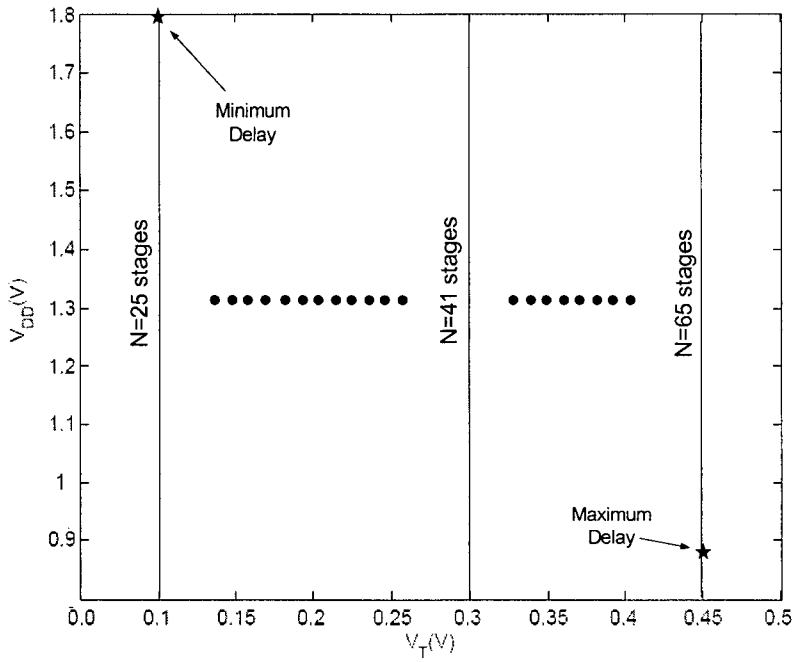


Figure 36. Template of Delay Modeling of the 8-bit Adder.

Following the above method, the delay of other circuits can be modeled in the same way provided the critical path is known.

6.3 Summary

A Hamming distance based model produces reasonable estimates of dynamic energy. Two distinct cases have been shown and the methods of modeling have been described in both the cases. In the first case the data-driven approach requires a training set. Realistic data sets, in place of random data sets, would give higher accuracy but that would require more information about the system being designed. Thus, there is a trade off between accuracy and modeling effort. For modeling leakage power, leakage current flow through any logic block

using HSPICE is needed. Delay modeling with V_{DD} and V_T is accomplished by translating the delay of the circuit to the delay of ring oscillator. It has been demonstrated that variable stage ring oscillator is needed to model the worst-case delay of an adder circuit for a V_{DD} vs V_T plane. The modeling has to be done for some discrete threshold voltage values. This type of modeling approach can be extended for any other logic circuit. The modeling effort would depend upon how many threshold voltage levels are available for typical circuit operation.

Though only two logic blocks have been used for illustrative purposes, a generalized method of modeling can be developed and automated on this approach. Using any of the data-driven active energy modeling methods or using a combination of both discussed in this chapter, active power of other logic blocks of Figure 23 like the ALU or controller can be developed. The data-driven approach would help to model the active power whereas the V_{DD} and V_T parameterized approach would reflect changes in the active and static power when both DVS and DVTS are used. Delay modeling method for the adder can also be used for delay modeling of the other blocks in the MIPS processor and be integrated to model the delay of the entire processor. Once such modeling is done in terms of V_{DD} and V_T, it would be possible to observe the trade off between power saving versus performance at the processor level as well at the system level.

Chapter 7

Conclusion and Future Work

7.1 Conclusions

In today's power-centric design paradigm, the trade-off between power and performance must be properly managed. For this purpose, a useful set of design metrics must be established, and new models for design blocks must be developed to capture the variations of power and delay as functions of V_{DD} and V_T. These new models can then be used in simulation tools such as SimpleScalar to determine suitable values of V_{DD} and V_T to optimize the efficiency of a given computational task. In this thesis, power-delay design metrics in the presence of leakage currents have been revisited. EDP has been compared with PDP and a new metric called PEP has been introduced. Comparing the three metrics, it is clear that EDP places a higher weight on delay reduction, PEP places a higher weight on power reduction and PDP tries to strike a balance between the two. It turns out that a design may be required to have a higher leakage percentage over dynamic power to achieve an overall lower power level. From a broader

perspective, these three metrics are members of the generalized family of metrics of the form P^nD^r.

The relationship between the various metrics in the V_{DD} vs. V_T plane has also been studied. The manner in which leakage currents affect the power and frequency contours in this plane has been illustrated. The effect is to reduce the feasible regions of operation. Next the optimal operating points for the three metrics have been identified. A fundamental result was established between the optimal value of a given metric and its connection to the best operating point (highest frequency, lowest power) in the region of interest. It appears that all the optimal values of the metrics lie along the trajectory of the best operating point.

Next, the effect of PVT on the metrics as well as the optimal operating line was investigated. It was found that there exists an operating region within which a circuit could function optimally. Though the discussion has been from ideal point of view where any supply and threshold voltage pair is possible, a discrete set of values is used in practice. In that case, the closest realizable point to the optimal solution should be used. But the general principle for the choice of (V_{DD}, V_T) pair for optimizing a design and the generalized metric based choice of optimization would still hold in the feasible regions.

Lastly, methodologies of power and delay modeling for digital blocks have been illustrated. Dynamic power was modeled as a function of Hamming distance of the input/output of the logic. Two different scenarios of the model were discussed, one where fixed energy consumption for every Hamming distance is found and the other where there can be a range

of values for a particular Hamming distance number. Delay modeling has been carried out with ring oscillators. It was found that variable stage ring oscillator is needed for modeling the delay of any logic block within a certain V_{DD} vs. V_T plane.

7.2 Future Work

In this work, thin-oxide gate leakage has been ignored. With technology scaling, direct tunneling through the oxide would have significant effect [17]. In all the above analysis, this form of gate leakage can be added. One would expect that this would further shrink the feasible region of operation as well as the optimal operating region. In other words, the choice of supply and threshold voltage values will be reduced.

The power and delay modeling can be used for a more realistic design evaluation in order to decide how the metrics and optimal operating region affect the overall performance and power consumption of the system. The VHDL models for the rest of the logic blocks of the MIPS processor can be developed, followed by power and delay analysis. These models can then be integrated into the SimpleScalar tool and the power and delay must be analyzed at the block level dynamically. The metrics can be used to study different types of design optimization. The relative overhead of supply and threshold voltage scaling also needs to be compared to the saving one achieves in terms of power or delay. Further the simulation time in SimpleScalar to obtain accurate result can be studied. Eventually, the design of an on-chip power management unit can be enabled using this approach.

REFERENCE

[1] International Technology Roadmap for Semiconductors, 2001.

[2] S. Hua, G. Qu and S. Bhattacharya, "Energy Reduction Techniques for Multimedia Applications with Tolerance to Deadline Misses", DAC 2003, June 2003.

[3] L. Benini, A. Bogliolo and G. DeMicheli, "A Survey of Design Techniques for System Level Dynamic Power Management", IEEE Trans. on VLSI Systems, Vol. 8, No. 3, June 2000.

[4] Christian Piguet, Jacques Gautier, Christoph Heer, Ian O'Connor and Ulf Schlichtmann, "Extremely Low-Power Logic", DATE 2004, pp: 656-663.

[5] Shekhar Borkar, " Design Challenges of Technology Scaling", IEEE Micro, Vol. 19, Issue 4, July-Aug 1999.

[6] E.J. Nowak, " Maintaining the benefits of CMOS scaling when scaling bogs down", IBM J. Research and Development, March/May 2002, Vol. 46, No. 2/3.

[7] Thomas D. Burd et al., " A Dynamic Voltage Scaled Microprocessor System", IEEE Journal of Solid -State Circuits, Vol. 35, No. 11, November 2000.

[8] C. Kim and K. Roy, " Dynamic V_{TH} Scaling Scheme for Active Leakage Power Reduction", Proceedings of the DATE 2002, March 2002.

[9] David E. Lackey, " Managing Power and Performance for System-on-Chip Designs using Voltage Islands", IEEE/ACM Conference on Computer Aided Design, pp. 195-202, November 2002.

[10] http://www.mips.com/ProductCatalog/P_MIPS324KEFamily/ProductCatalog/P_MIPS324KEFamily/productBrief

[11] Bowman, K., et al., " Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration", IEEE Journal of Solid-State Circuits, Volume 37, February 2002, pp 183-190.

[12] Borkar, S., " Parameter variations and Impact on Circuits and Microarchitecture", C2S2 MARCO review, March 2003.

[13] R. Gonzalez, B. M. Gordon and M. Horowitz, "Supply and Threshold Voltage Scaling for Low Power CMOS", IEEE Journal of Solid-State Circuits, Vol. 32, No. 8, August 1997.

[14] D. A. Hodges, H. G. Jackson and R. A. Saleh, *Analysis and Design of Digital Integrated Circuits In Deep Submicron Technology*, Third Edition, McGraw-Hill, 2004.

[15] H. Veendrick, " Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits", IEEE Journal of Solid-State Circuits, vol. SC-19, no. 4, pp. 468-473, 1984.

[16] James Kao et al., "Subthreshold Leakage Modeling and Reduction Techniques", IEEE/ACM Conference on Computer Aided Design, pp. 141-148, November 2002.

[17] Kaushik Roy, " Leakage Power Reduction in Low-Voltage CMOS Design", Proceedings of IEEE International Conference on Electronics, Circuits and Systems, September 1998.

[18] Steven Martin et al., " Combined Dynamic Voltage Scaling and Adaptive Body Biasing for Lower Power Microprocessors under Dynamic Workloads" Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on, November 2002, pp. 721 – 725.

[19] V. Tiwari et al., " Power Analysis of Embedded Software: A First Step Towards Software Power Minimization," IEEE Trans. VLSI Systems, Vol.2, N.4, pp.437-445, December 1994.

[20] V. Tiwari et al., " Instruction level Power analysis and Optimization of Software," International Conference on VLSI Design, Bangalore, India, January 1996.

[21] M. T. C. Lee et al., " Power Analysis and Minimization Techniques for Embedded DSP Software", IEEE Transactions on the VLSI Systems, pp. 123-135, March 1997.

[22] S. Nikolaidis et al., " Instruction-Level Power Consumption Estimation Embedded Processors Low-Power Applications", International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Foros, Ukraine, July 2001.

[23] D.Brooks et al., " Wattch: A Framework for Architectural-Level Power Analysis and Optimizations", Proc. 27th Int. Symp. On Computer Architecture (ISCA27), May 2000.

[24] N. Vijaykrishnan et al., "Energy-Driven Integrated Hardware-Software Optimizations Using SimplePower", Proc. 27th Int. Symp. On Computer Architecture (ISCA27), May 2000.

[25] G.Cai et al.," Architectural Level Power/Performance Optimization and Dynamic Power Estimation", Cool Chips Tutorial in conjunction with the 32nd Int. Symp. on Microarchitecture, November 1999.

[26] http://www.simplescalar.com

[27] Soraya Ghaisai et al., " A Comparison of Two Architectural Power Models", Proceedings of First International Workshop on Power Aware Computer Systems, pp. 137-152., 2000.

[28] Nam Sung Kim et al., " Microarchitectural Power Modeling techniques for Deep Sub-Micron Microprocessors", ISLPED'04, August 2004.

[29] Sandeep Dhar et al., " Closed-Loop Adaptive Scaling Controller for Standard-Cell ASICS", ISLPED'02, August 2002.

[30] Krisztian Flaunter et al., " IEM926: An Energy Efficient SoC with Dynamic Voltage Scaling", Proceedings of DATE'04, February 2004.

[31] Anantha Chandrakashan, William J. Bowhill, Frank Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.

[32] Krisztian Flaunter et al., " Automatic Performance Setting for Dynamic Voltage Scaling", International Conference on Mobile Computing and Networking, pp. 260-271, 2001.

[33] Jan M. Rabaey, Anantha Chandrakasan and Borivoje Nikolic, *Digital Integrated Circuits*, Prentice Hall Publication, 2003.

[34] http://www.arm.com/products/CPUs/cpu-arch-IEM.html

[35] Krisztian Flaunter et al., "A combined hardware-Software Approach for Low-Power SoCs: Applying Adaptive Voltage Scaling and Intelligent Energy Management Software", DesignCon 2003, January 2003.

[36] Ashish Srivastava et al., " Power Minimization using Simultaneous Gate Sizing, Dual - V_{DD} and Dual-V_{th} Assignment", Proceedings of DAC 2004, June 2004.

[37] Koichi Nose et al., " V_{TH} hopping Scheme for 82% Power Saving in Low-voltage Processors", IEEE Custom Integrated Circuits Conference 2001, pp.93-96, May 2001.

[38] S. Mutoh et al., "1-V Power Supply High-Speed Digital Circuit Technology with Multi-threshold Voltage CMOS", IEEE Journal of Solid-State Circuits, vol. 30, pp. 847-854, August 1995.

[39] S. Narendra et al., "Impact of Using Adaptive Body bias to Compensate Die-to-die Vt variation on Within-die Vt Variation", IEEE Symposium on Low-Power Electronics Design, August. 1999, pp. 229-232.

[40] Johan Pouwelse et al., " Dynamic Voltage Scaling on a Low-Power Microprocessor", Proceedings of 7th Annual International Conference on Mobile Computing and Networking, 2001, pp. 251-259.

[41] Dipanjan Sengupta and Resve Saleh, "Power-Delay Metrics revisited for 90nm CMOS technology", in Proceedings of ISQED '05, San Jose, USA, March 2005.

[42] A. Wang, A. Chandrakasan and S. V.Kosnocky, "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits", in Proc. of the IEEE Computer Society Annual Symp. on VLSI, 2002.

[43] J. Burr and J. Shott, "A 200m V encoder-decoder circuit using Stanford Ultra Low-Power CMOS", ISSCC Digest of Technical Papers, February 1994, pp. 84-85.

[44] S. Borkar et al., "Parameter variations and impact on circuits and microarchitecture", in Proceedings 2003 Design Automation Conference, 2003, June 2003, pp. 338-342.

[45] Anirban Basu et al, "Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era", in Proceedings of Design Automation Conference, 2004, June 2004.

[46] Anirudh Devgan and Sani Nassif, "Power Variability and its Impact on Design", Proc. of 18th International Conference on VLSI Design, January 2005.

[47] Tom Chen et al., "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 (VLSI) Systems, Vol. 11, No. 5, October 2003.

[48] Sandeep Dhar et al., "Switching Regulator with Dynamically Adjustable Supply Voltage for Low Power VLSI", 27th Annual Conference of the IEEE Industrial Electronics Society, November-December 2001.

[49] John L. Hennesy and David A. Patterson, *Computer Organization and Design*, Morgan Kaufmann Publishers, Inc., 1998.

[50] Tomohisa Wada et al., "An analytical Access Time Model for On-Chip Cache Memories", IEEE Journal of Solid-State Circuits, August 1992.

[51] Glen Reinman et al., " CACTI 2.0: An Integrated Cache Timing and Power Model", Western Research Laboratory Report, February 2000.

[52] W.T. Shiue and C. Chakrabarti, " Memory Design and Exploration for Low Power, Embedded systems", ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 6, issue 4, pp. 553-568, October 2001.

[53] Paul E. Landman et al., "Architectural Power Analysis: The Dual Bit Type Method", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 3, No. 2, June 1995.

[54] Ricardo Gonzalez et al., " Energy Dissipation In General Purpose Microprocessors", IEEE Journal of Solid-State Circuits, Vol. 31, No. 9, September 1996.

[55] Farid N. Najm, " Power Estimation Techniques for Integrated Circuits", ICCAD-95, November 1995, pp. 492-499.