Impact of Parameter Variations on Multi-Core Chips

Eric Humenay, David Tarjan, Kevin Skadron
Dept. of Computer Science
University of Virginia
Charlottesville, VA 22904
humenay@virginia.edu, dtarjan@cs.virginia.edu, skadron@cs.virginia.edu

ABSTRACT

Increasing variability during manufacturing and during runtime are projected for future generation microprocessors. This paper introduces a pre-RTL, architectural modeling methodology that incorporates the impact of manufacturing and runtime temperature variations on delay and power for both combinational logic and SRAM structures. The model is then used to show that frequency variations among microarchitectural functional units and among cores are relatively small in a high-performance microprocessor design. However, the impact of within-die systematic process variations on leakage power will result in major leakage variation across multiple cores on a single chip. WID leakage variation can cause core-to-core leakage to differ by as much as 45%.

1. INTRODUCTION

The 2005 International Technology Roadmap for Semiconductors projects that parameter variations will present critical challenges for manufacturability and yield. While process, circuit-design, and statistical CAD techniques can mitigate the impact of some parameter variations, the roadmap and some industry observers [3] have claimed that computer architecture will play an important role in mitigating the effects of parameter variations.

At the same time, multi-core designs have become the dominant organization for future microprocessor chips, as high-frequency single cores run into power, thermal and complexity limitations that will only be exacerbated by future technology trends. The inclusion of multiple cores—of the same or different types—allows continued exponential performance scaling for applications that can take advantage of the parallelism that CMPs offer. Multi-core organizations, however, also multiply the ways in which parameter variations can affect a processor. Although some have speculated that this will yield significant variations among units in a single core, this paper argues that instead the most important phenomenon will be core-to-core (C2C) leakage variations at the 45nm technology node and beyond.

Parameter variations encompass a range of variation types, including *process* variations due to manufacturing phenomena, *voltage* variations due to manufacturing and runtime phenomena, and *temperature* variations due to varying activity levels and power dissipations—in fact, these three main sources are often referred to as PVT (process-voltage-temperature) variations. Process variations are static and manifest themselves as die-to-die (D2D), within-die (WID) variations, and wafer-to-wafer variations (W2W), while temperature and voltage variations are a dynamic phenomena. Temperature variations stem from different activity factors among cores, functional units, from different circuit structures, and from non-uniformities in the thermal interface material (TIM) that bonds the chip to its package. Voltage variations stem from IR drops that result from non-ideal voltage distribution, which in turn are exacerbated by activity-dependent IR drops. These are exacerbated by temperature-dependent leakage-current variations (i.e., varying the I term) or switching activity that causes voltage droops due to circuit inductance and possibly insufficient decoupling capacitances. These three variation sources exhibit a number of feedback loops. Process variations affect leakage, which affects both voltage and temperature. Temperature then affects leakage forming a feedback loop between the two parameters [11].

This paper focuses on WID variations. D2D variations cause each die on a wafer to have different mean values for a particular parameter. Gate length is the most common parameter to exhibit D2D variation, and is typically modeled by assigning a normally distributed offset to each die. D2D variations can be dealt with by sorting chips into different product bins or chip-wide techniques to compensate for a parameter's offset, such as adaptive body biasing [17]. W2W variations primarily affect the shape of the WID systematic pattern as well as across wafer systematic patterns that in chip-to-chip variations similar (but larger in magnitude) to D2D variations. In short, D2D variations determine the variance of the frequency distribution while WID variations determine the mean of the distribution [4].

WID process variations can further be divided into *random* and *systematic* variations. Random variations will affect each transistor differently, while systematic variations cause transistors to be spatially correlated. Systematic variations may be caused by a variety of different sources. Most notably, variation in optical intensity across the exposure field and non-uniform chemical-mechanical polishing that occurs due to different pattern densities.

This paper argues that the WID variation phenomenon of chief interest in the *computer architecture* domain will occur at a C2C granularity, rather than at a unit-to-unit granularity. While unit-to-unit variations in delay will occur, the WID frequency distribution will likely be dominated by large SRAM structures. This occurs because of the nature in which existing critical path models determine worst-case delay. The final result is that large SRAM units will have a mean delay that is much greater than the rest of the units. We find that overall impact of random variations on clock frequency to be fairly mild when reasonable assumptions were made about each parameter's variance. At a per-unit granularity, random variations in leakage are even milder than frequency variations since a stage/unit's worst-case delay is the unit's maximum critical path delay, while leakage is merely an aggregate sum across all the transistors in the unit. WID systematic variation, WID_{sys}, play an important role, because at the 45nm generation and beyond, reduced core areas will cause parameters within a core to be highly spatially correlated, while the amount of variation that can occur across a chip can be large.

Systematic variation will result in both C2C frequency and leakage variation. C2C frequency variations will be modest in comparison to leakage variation. This is because the amount of WID_{sys}

In the 2006 Workshop on Architectural Support for Gigascale Integration (ASGI'06), held in conjunction with the 33rd International Symposium on Computer Architecture

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Impact of Parameter Variations on Multi-Core Chips	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Virginia,Department of Computer Science,151 Engineer's Way,Charlottesville,VA,22904-4740**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**9**	

that occurs across a chip—10–15% variation in gate length—has only a linear impact on frequency. Instead, leakage—which has an exponential dependence on the gate-length variation (because of its impact on threshold voltage)—shows the most important architectural WID variation.

Specifically, the main contributions of this paper are:

- A top-down model for studying parameter variations at the architectural level. This model accounts for random and systematic WID device variations. D2D and W2W variations are easily added but not discussed further here. The chief requirement is that the model not require detailed circuit implementations, because early-stage, pre-RTL studies, especially for a multi-core chip, require an ability to explore the design space before detailed circuit implementations are available.
- An improved critical path model is used to analyze the likely impact that each functional unit will have in determining the processor's maximum clock frequency distribution. We determine that SRAM units are likely to determine the processor's clock speed, not logic dominated stages. In particular, the L1 caches will be the primary limiter, unless variation-aware techniques are applied.
- Using a 14mm by 14mm die as our baseline chip model, we show that in a multi-core architecture, C2C leakage variation can be as much as 45% when the thermal-feedback loop between leakage and temperature has been closed.

The rest of the paper is structured as follows. Section 2 gives an overview of parameter variation phenomena and discusses related work, Section 3 introduces the architectural PVT model, Section 4 looks at frequency variation at both the functional unit level and the core level, Section 5 looks at across chip leakage variation, and Section 6 concludes.

2. BACKGROUND AND RELATED WORK

2.1 Background

Parameter variations cause chip characteristics to deviate from the uniform, ideal values desired at design time. Three major sources of variation are often discussed: *process* variations, which consist of deviations in the manufactured properties of the chip, such as feature size, dopant density, etc.; *voltage* variations due to non-uniform power-supply distribution, switching activity, and IR drop; and *temperature* variations due to non-uniformities in heat flux of different functional units under different workloads as well as the impact of non-uniformities in the chip's interface to its package. These comprise the classic "PVT" variations.

Process variations occur because specific steps in the fabrication process, such as lithography, ion implantation, and chemical-mechanical polishing, are vulnerable to imperfections, noise, and imperfect control across time and locations. Process variations present a problem because they can make a given circuit exhibit different delay or power characteristics than intended during design. Since the operating frequency in high-performance chips is typically determined by the expected delay of the slowest path, variation in the delay of the slowest path can make a single, fixed clock frequency too fast (causing errors) or too slow (incurring an opportunity cost). Post-manufacture testing is therefore used to characterize chips and "bin" them according to their maximum clock frequency giving perhaps a 30% variation among chips. Unfortunately, the fastest chips are usually the leakiest, because both frequency and leakage are affected by one of the main victims of process variations, the threshold voltage. Threshold voltage is affected by both fluctuations in the channel doping (which gets worse as smaller channel lengths mean fewer dopant atoms are in the channel region) and the effective gate length (which affects threshold voltage through Drain Induced Barrier Lowering (DIBL)). In fact, subthreshold leakage is exponentially dependent on threshold voltage, and this produces large D2D leakage variation. The fastest chips often cannot operate at their peak sustainable frequency because they would overheat, and a suitable cooling solution is too expensive. Per-chip adaptive body biasing (ABB) [17] can reduce these spreads and boost the yield of high-quality parts at the cost of some additional testing and calibration.

Until recently, W2W and D2D variations were the main source of concern, and these could be addressed through bin splits and ABB. However, as transistors scale in size, small, WID variations in feature size and doping density —once imperceptible relative to the large features sizes in older technologies—have become important as their impact becomes larger in relative terms. As mentioned in Section 1, two forms of WID variations are present. *Random* variations are small changes from transistor to transistor which do not show any correlation across larger distances on the chip. *Systematic* variations, on the other hand, exhibit high degrees of spatial correlation. Random variations stem primarily from two main sources. Non-uniform dopant implantation in the channel depletion region affect threshold voltage, and imperfect control of the lithographic process result in non-deterministic gate lengths. Systematic variations in gate length stem primarily from the the lithographic exposure process. Non-uniform exposure intensity, lens aberrations, defocus errors, and mask errors, as well as many other factors, may all contribute to the final systematic variation pattern.

While systematic variations are modeled as affecting all circuits in a critical path in the same fashion, random WID variations can affect the same circuit in a myriad of different ways. Analyzing all possible permutations is usually prohibitive, requiring statistical treatments which have become a major research topic in the CAD community. These variations are exacerbated by runtime effects like temperature and noise. To account for the possible slowdown due to PVT variations, voltage margin must increased to compensate. The concern is that as technology scales PVT variations will increase in severity, resulting in worse required design margins.

2.2 Modeling

While there has been a great deal of work on statistical approaches to modeling and compensating for variation in the CAD community, there has been little work on modeling variations in the architecture community. Yet parameter variations are important enough that architectural mitigation techniques need to be explored *before* or at least in parallel with circuit design. This requires a pre-RTL modeling capability that does not depend on detailed circuit designs to estimate the impact of parameter variations on different microarchitectural units.

Perhaps the most relevant prior work is the "FMAX" model introduced by Bowman et al. [4]. FMAX is a predictive model for capturing the maximum frequency distribution. It is comprised of a generic critical path model (GCP) that is based on canonical NAND gates. The NAND gate's delay is derived from the RC delay equation. The delay distribution can then be determined with Monte Carlo analysis by varying the delay equation's inputs. Results from the GCP model were compared to measured data from high volume industrial 0.25um and 0.13um processes. While the GCP model did not perfectly recreate the measured frequency distribution, it did provide insight into what the frequency distribution would look like. In the FMAX model there are two parameters of concern to microarchitects: number of independent critical paths, N_{cp}, and the

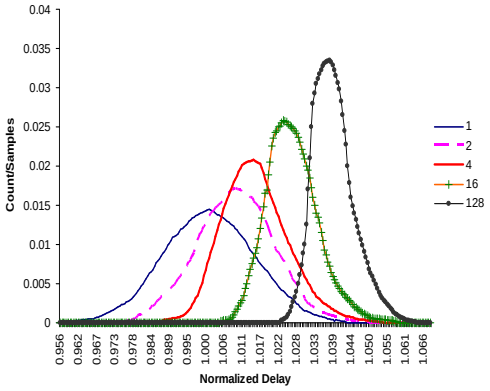

Figure 1: Plot showing delay distribution's dependency on the number of N_{cp}.

depth/length of the critical path, L_{cp}.[1]

The statistical relevance of N_{cp} is that the worst case delay of a unit is the *maximum* delay across all critical paths. As N_{cp} increases, the stage's mean delay will also increase. The reason this occurs is that when a larger sample size is considered, the probability increases that the worst-case is an extreme outlier. Similarly, as N_{cp} increases the distribution's variance will decrease since the maximum delay is likely to be determined by an outlier. Fig. 1 illustrates the dependency between the delay distribution and N_{cp}. Logic depth determines L_{cp}. Path delay is determined by taking an aggregate sum of each gate's delay in the path. Since a sum operation is performed, the path's ratio of variance to mean will decrease as L_{cp} increases.

This paper primarily focuses on leakage variation that occur as a result of systematic effective gate length (L_{eff}) variations. Zhang et al. [18] showed that it is necessary to consider systematic L_{eff} variations when estimating full-chip subthreshold-leakage. Ashouei et al. [2] developed a model that addresses systematic WID_{sys} leakage variation at the circuit level. Systematic variations are modeled as circular areas with highly correlated L_{eff} values. The correlated areas may vary in their area, location, and magnitude. Our modeling methodology differs from this since we base our WID systematic variation pattern off of measured data reported in [6, 14].

It is necessary to emphasize that the pattern of the WID systematic variation, (WID_{sys}), is highly dependent upon the fabrication process. They can manifest themselves as being either deterministic or random in nature depending on the particular fabrication process. Deterministic systematic variations can be mitigated with a combination of optimal proximity correction, phase-shift masking, as well as other mask-level techniques. Since masks cost are already burdensome and increasing with every technology node, design-for-manufacture (DFM) techniques that simplify mask complexity with variation-tolerant designs are desired.

The main advantage of modeling a measured deterministic systematic pattern is to better understand at what granularity the systematic change will occur at, and how this will affect decisions in the architectural domain.

2.3 Architectural Implications

In [13], Marculescu and Talpes propose to apply the FMAX model in the microarchitecture domain by assuming that N_{cp} is proportional to the stage's device count. While this assumption often times holds true, it is not always the case since not all paths are critical [1, 9]. Also, the authors do not consider that a large portion of a stage's delay will be spent in wires when estimating L_{cp}. While these assumptions provide a simplistic way to reason about how variations affect the FMAX distribution, it may be misleading when analyzing the impact that each particular unit's delay distribution will have on the final FMAX distribution. The author's proceed to show how a GALS architecture can mollify the impact of process and temperature variations.

Ernst et al. [9] also use a GALS architecture, but use shadow latches on critical paths to dynamically correct and detect circuit timing errors. With this added functionality, the authors show that significant power savings can be obtained by performing per stage DVS in order to reclaim design margin. The focus of this work, however, is on reclaiming excess design margins in single-core chips, including D2D variations and not just unit-to-unit and run-time voltage-temperature variations.

Neither paper considers the impact of C2C variations. Our paper present a more detailed modeling methodology and shows the importance of C2C phenomena. The main contributions of the model are: (i) Stage-specific N_{cp} and L_{cp} characteristics, most importantly the differences between SRAM and combinational logic; (ii) Systematic L_{eff} variations; (iii) The importance of leakage, as opposed to frequency, as a consequence of WID variations and resulting design driver.

3. VARIATION MODEL

3.1 Critical Path Model

To model the impact of parameter variations upon delay, we observe that the clock frequency is dictated by the worst-case delay for any pipeline stage. Similarly, the delay of each pipeline stage is determined by the worst-case delay across all the stage's critical paths. Frequency is therefore given by MAX(T_{cp}), i.e. the worst-case delay of all critical paths. The delay of each critical path, in turn, can be decomposed into D2D,WID-random, and WID_{sys} variations:

$$T_{cp} = T_{cp,nom} + \Delta T_{D2D} + \Delta T_{WID-rand} + \Delta T_{WID-sys} + \Delta_{Temp} + \Delta_V \quad (1)$$

where $T_{cp,nom}$ is the nominal critical path delay without variations, ΔT_{D2D} is the contribution of D2D variation, which is a fixed offset per die; $\Delta T_{WID-sys}$ is the WID contribution of systematic variations; and $\Delta T_{WID-rand}$ is the *accumulated* contribution of random variations across that critical path. Δ_{Temp} and Δ_V in turn represent the additional impact of temperature and voltage variations.

To understand the role that each pipeline stage plays in determining the processor's final frequency distribution, we model delay variations at a per functional unit granularity. For a critical path model to be useful for architectural studies, the model should be able to recognize the inherent differences between different functional units' circuit structures. With this information, the model should then be able reason about the processor's frequency distribution.

A main assumption in our model is that all stages can be loosely categorized as being dominated by either SRAM or combinational logic. SRAM-dominated stages include not only cache/TLB stages, but those that involve large buffers, queues, or lookup tables. Combinational logic will have a much larger L_{cp} than an SRAM stage since a large portion of an SRAM device's total delay is spent in

[1]In [4], the authors use the notation n_{cp} to represent logic depth. In order to avoid confusion between n_{cp} and N_{cp}, we refer to logic depth as L_{cp}

wires (bitlines, wordlines, etc.) rather than transistors. Therefore, in an SRAM device, transistors will only contribute to a small fraction of the stage's total delay. Even in logic-dominated stages such as an ALU, it is expected that a significant portion of the overall delay will be spent in the interconnects, but wire-friendly circuit implementations can be used to minimize the amount of interconnect delay [12]. Prior critical path delay models did not consider the ratio of wire delay to transistor delay, causing their conclusions to be overly pessimistic.

It should be noted that, while wire delay is not exempt from manufacturing variations, it is the general consensus that their impact will pale in comparison to that of transistor-level variation. The reason for this is that wire geometries are not as aggressively scaled as gate length. For simplicity, we have chosen not to model wire variation in this study. Future work includes analyzing the interaction between WID_{sys} L_{eff} and WID_{sys} wire variations.

The other main difference between logic and SRAM stages is the value of N_{cp}. The critical path in any SRAM is in accessing the actual cell through a wordline and sensing the voltage difference on the bitline with the help of the sense amplifiers. Since both wordline and bitlines have to be brought back to their initial state before a new access can begin, this critical path forms a loop with itself. As a consequence, this critical path determines not just the access time, but also the minimum cycle time in a pipelined cache—pipelining this critical path will be increasingly difficult as variations worsen. We model the number of critical paths in an SRAM as the number of bits in the cache multiplied by the number of read ports per bit.[2] Prior models did not consider that each read port is equally critical, but rather treated each SRAM cell as being one critical path.

Identifying the value of N_{cp} in a logic dominated stage is more complicated than in an SRAM. In a standard circuit design, only a subset of the total paths are actually critical. However, circuit designers increase the delay of non-critical paths in order to reduce dynamic and static power dissipation, potentially causing non-critical paths to become critical.

The inherent differences between SRAM and logic circuitry necessitate different critical path models. In order to estimate the impact of process variations on SRAM structures, we have modified a beta version of CACTI 4.0 to incorporate the effects of process variations on delay. More detail will be provided about this model in the following section.

For simplicity, the logic critical path model is based on conventional static adder circuitry. Although representing all logic stages with an adder is an idealistic assumption, we feel that this simplification still provides important insights that allow architects to draw useful conclusions.

The combinational logic critical path model is based off of a Sklansky adder. The Sklansky adder is not as heavily impacted by wire delay in comparison to similar prefix adders, such as a Kogge-Stone [12]. We assume the critical path in an adder is determined by the time required to pass the carry-bit from the least significant bit to the most significant bit. The entire delay for the adder is the carry-bit propagate delay as well as the delay of the initial carry generate and the sum logic. Fig 2 illustrates the critical path in a Sklansky adder. For simplicity, only the carry-bit's path is shown. One drawback of the Sklansky adder is that the number of fanouts double at each level. The high fanouts make it important to properly size gates on the critical path, or else high performance would not be obtainable. Transistor widths were chosen such that a 64 bit adder's nominal delay fell in accordance with data extrapolated from [12], and curve fitted to a 45nm technology node. We assume that 35% of the total delay in a 64 bit Sklansky adder can be attributed to interconnect delay.

[2]We have neglected write ports on the assumption that they are not on the critical path.

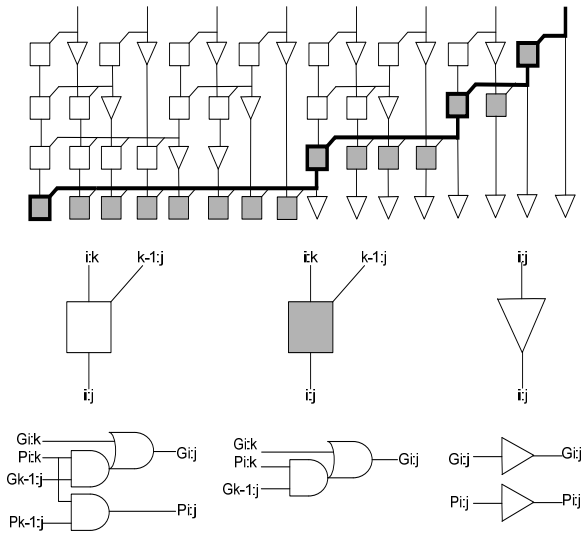

Figure 2: Critical path in a Sklansky adder is highlighted. The critical path is assumed to be the delay required to propagate the carry bit from the least significant bit to the most significant bit.

In order to estimate delay, we used the same delay model with which CACTI models decoder logic. More information about the delay model can be found in [10]. One advantage of using this delay model is that it takes into account that transistor delay is dependent on the load of the input signal. By properly modeling this, the correlation in delay between adjacent gates is accounted for, which some prior models have neglected. All gates in the critical path, except for buffers and inverters, require two input signals, one from the previous gate in the critical path and the other from the bit slice (white squares in Fig. 2). The critical path model only considers variations on the input signal from gates in the critical path. The reason variations on inputs received from bit slice logic is not factored into the delay model is because the bit slice path is not critical and the bit-slice path's result will be computed well before the carry-bit signal will have arrived. However, this assumption may be idealistic since it is common for circuit designers to increase the delay of non-critical paths in order to save power. Prior variation models treat each gate's delay as being independent of one another. While greatly simplifying the analysis, a model intended for more thorough comparative analysis should consider this.

The improved critical path model does have its limitations. Due to the characteristics of the delay model, the critical path model cannot account for delay variations that occur in transistors in series. For this reason, the critical path model cannot traverse paths that flow though the NOR pull-up and NAND pull-down networks. Also, we neglect the impact that variation in gate capacitance has on the previous gate's output signal. Finally, SRAM delay calculations do not take into account bitline leakage.

3.2 Random Process Variations

Random variations are modeled as being normally distributed parameters. In this study we consider C_{ox}, L_{eff},W_{eff}, and V_{th}. Both SRAM and combinational logic critical path models use a first order RC model for all elements. For such a model, the impact of

varied parameters on gate resistance can be expressed as:

$$R = \frac{1}{\mu C_{ox}} \frac{L_{eff}}{W} \frac{1}{V_{dd} - V_{th}} \quad (2)$$

(3)

This is performed with a brute-force Monte-Carlo analysis on N_{cp} critical paths to determine the unit's delay distribution.

Transistor width becomes a very important parameter when comparing different function units. The reason for this is because the magnitude of random dopant fluctuations in V_{th} is proportional to transistor width,W_{eff}.

$$\sigma V_{th} \sim \frac{1}{\sqrt{W_{eff}\dot{L}_{eff}}} \quad (4)$$

For SRAM functional units, we assume the L1 cache to have minimal W_{eff}. For all other SRAM units, we assume W_{eff} to be 5 times the minimal value. This fact, together with the large size of caches, makes them most likely to exhibit the worst variation.

Table 1 shows our baseline assumptions for the variability of a minimum size transistor. These values were extrapolated from ITRS and academic predictions [15, 8].

Name	$3\sigma/\mu$
L_{eff}	12%
V_{th}	30%
C_{ox}	10%
W_{eff}	4%

Table 1: Default $3\sigma/\mu$ for parameters varied.

3.3 Systematic Process Variations

Across chip L_{eff} variations arise from imperfections in the fabrication process. The optical component that we model is chiefly due to lens aberrations that can be modeled as a simple polynomial function of position within the field of exposure [6]. The equation can be approximated by

$$L_{eff} = a{\cdot}x^2 + b{\cdot}y^2 + c{\cdot}x + d{\cdot}y + e{\cdot}xy + intercept \quad (5)$$

where x and y are the coordinates on the chip's surface. Baseline values derived from [6] and scaled to 45nm are given in Table 2. They were chosen under the assumption that the proportion of WID_{sys} to mean L_{eff} will stay constant with scaled dimensions. In our model systematic variations will cause there to be a 12% difference between nominal L_{eff} and the area of the chip having the largest L_{eff}.

Parameter	Value
a	5.37×10^{-4}nm/mm^2
b	1.829×10^{-3} nm/mm^2
c	-1.06×10^{-2}nm/mm
d	-.458 nm/mm
e	-1.67×10^{-3}nm/mm
intercept	28.0 nm

Table 2: Constants for our 2nd order polynomial equation for modeling WID systematic variations

Our model assumes that all circuit types within a core are affected uniformly by WID_{sys}, neglecting the impact of pattern density , orientation, and sizing. This is justified both by the high-level, pre-RTL architectural treatment and the fact that within each core, SRAMs dominate both the core's operating frequency and its leakage, exhibit a regular pattern density, and have near minimum-size features. Also, Orshansky et al. [14], measure WID_{sys} for various circuit layouts, and show the majority of circuit layouts will have a similar bowl-like pattern across the chip.

Ultimately, variations in gate length matter because they affect threshold voltage, which determines both switching speed and leakage. In [7], the authors present an equation for determining V_{th} as a function of L_{eff}:

$$V_{th_{eff}} = V_{th0} - V_{dd} \cdot exp(-\alpha_{DIBL} \cdot L_{eff}) \quad (6)$$

where V_{th0} is the threshold voltage for long channel transistors, 0.22; α_{DIBL} is the DIBL coefficient, 0.15; and V_{dd} is the supply voltage, 1V. The default values for V_{th0} and α_{DIBL} were provided in [5]. This equation highlights an important concept: as L_{eff} increases V_{th} will also increase. This is why leakage has an exponential relationship on L_{eff}.

4. FREQUENCY VARIATION

In Fig. 3 the delay distribution of several of the more interesting SRAM functional units is shown. The delay distribution only considers the SRAM cell and the delay variation in the decode logic is not taken into account. The figure illustrates an important concept: not all units/stages will directly contribute to the final WID frequency distribution. Table 3 shows the parameters corresponding to each unit's delay distribution. In Sec. 2, it was mentioned that variance decreases as N_{cp} is increased; however, the 64KB L1 cache has a greater variance than the other two units even though it has a much larger N_{cp}. The reason for this is that the L1 has minimum sized W_{eff}, and according to Eq. 4, this causes the L1 to have greater σV_{th}. As can be seen, either the L1 D-cache or I-cache is likely to be the slowest SRAM unit because of the large N_{cp}. The reason that variation in the SRAM cell only results in a 5% performance degradation is that SRAM access time is a combination of bit line delay, wordline delay, and sense-amp delay. According to our modeling methodology, process variations will only significantly impact bit-line delay. In conclusion, even though the variation in bit-line delay can be relatively large, it will not have a great impact on overall access time since only a fraction of the overall access time is susceptible to process variations.

Name	Entries	Line Size(bits)	ports	N_{cp}	W_{eff}(nm)
RF	120	64	6	46080	375
TLB	1024	64	1	65536	375
L1	512	1024	1	524288	75

Table 3: Description of functional units plotted in Fig. 3.

In Figure 4, the delay distribution of the combinational logic model is compared to the slowest SRAM stage (64KB L1 cache). Mean combinational logic delay is significantly less than the L1 caches' mean delay since N_{cp} is equal to 1 in the combinational logic model. For this same reason, the logic delay distribution also has a much greater variance since variance decreases as N_{cp} increases.

The simplistic critical path model shows that *the WID frequency distribution of the processor core will solely be determined by the L1 caches*. The primary reasons for this is that the L1 cache's have greater N_{cp} than all other SRAM structures, causing the delay distribution's mean to increase. If caches are removed from consideration (e.g., by allowing multi-cycle access), then TLBs and other SRAM structures dominate With nominal WID_{sys}, combinational

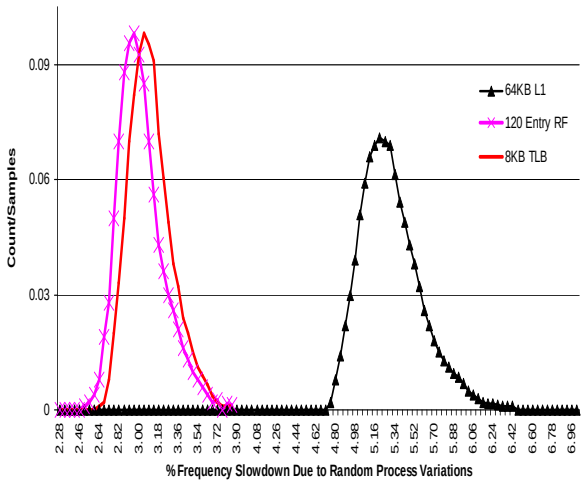

Figure 3: Cacti generated delay distribution for several different SRAM functional units

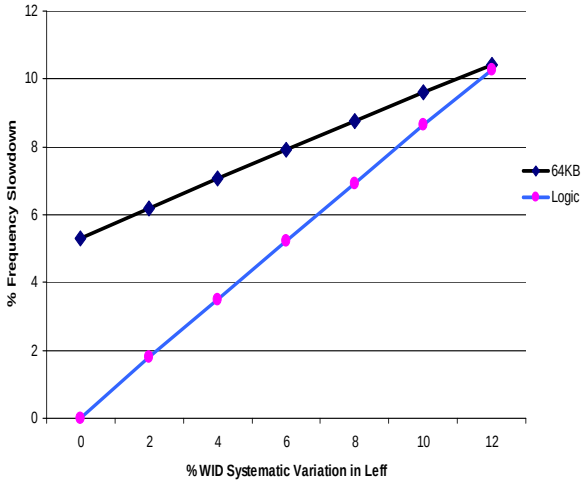

Figure 5: Distributions of a logic stage and a 64KB cache when random variations and 12% systematic variations are considered.

logic stages will be much faster than SRAM because of their low N_{cp}. However, when D2D variations are considered it is possible for combinational logic to have delay greater than the L1 since logic stages are more sensitive to changes in L_{eff}. Unit-to-unit delay variations will contribute to clock skew, which does have an impact on the maximum frequency. Determining the impact of unit-to-unit variations on clock skew is more of a circuit-design rather than a pre-RTL architectural-modeling issue, and hence is beyond the scope of this paper.

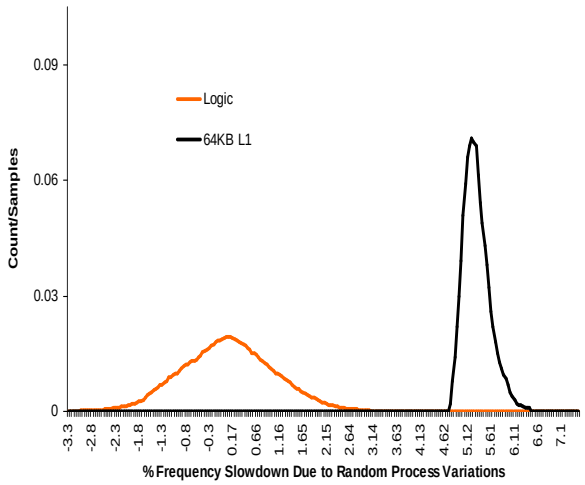

Figure 4: Comparison of logic stage's delay distribution to that of the slowest SRAM stage. Logic stage is modeled as having N_{cp} of 1 and L_{cp} of 14.

Systematic L_{eff} variations will have a detrimental impact on delay since resistance is dependent on both L_{eff} and V_{th}. The magnitude of the effect depends upon the functional unit's ratio of logic delay to wire delay, the change in L_{eff}, and how problematic the DIBL effect is in the particular process. Intuitively, logic dominated stages will be more impacted by systematic L_{eff} variations than an SRAM stage since logic stages are more transistor dominated than an SRAM. This is illustrated in Fig. 5. Data in this figure was gathered by a Monte-Carlo analysis with the mean L_{eff} value being varied from 0% to 12% in order to represent WID_{sys}. As the figure shows, WID_{sys} will more severely affect the performance of logic stages than SRAMs. Fig. 5 also shows that WID_{sys} will result in C2C frequency variation with the mean difference between the fastest core location and the slowest core location being less than 5%.

In summary, when considering only WID_{rand} variations, the L1 cache will determine the processor's FMAX distribution, resulting in an average 5% performance loss. The degradation is likely to affect all cores on a chip to a similar degree, because the variance in this value will be small ($< 1\%$). Since delay exhibits a linear dependency on systematic variations, a 12% degradation in L_{eff} will result in an average frequency degradation of roughly 5%. The combination of WID_{rand} and WID_{sys} results in a 10% frequency degradation for the slowest location on the die. A 12% change in L_{eff} is a worst case assumption, so according to our model, the difference between the frequency of the fastest core location and the slowest core location will be less than 5%. It is worth noting that the C2C variation is likely to be less than 5% if the relationship between leakage and SRAM access time were considered in our delay model. The reason for this being that higher leakage slows down SRAM access time, and the fastest cores will have the most leakage.

5. LEAKAGE VARIATION

In the previous section we showed why WID variations will not play a large role in determining the C2C frequency distribution. When turning to leakage, this is not the case. As mentioned previously, WID_{rand} leakage variation will not be significant at a course enough granularity to concern microarchitects since variation is averaged out when a sum operation is performed. Fig. 6 illustrates this. The distributions of the leakage summation across 1, 2, and 4 transistors is showed. Each distribution is normalized to its smallest value in order to compare the variance of each distribution.

On the other hand, WID_{sys} L_{eff} variations will *shift the threshold voltage of all transistors in a core* by an offset. V_{th} has an exponential effect on the overall leakage of a core, as opposed to the linear effect of threshold variations on frequency. The magnitude of the across chip leakage variation is dependent on both L_{eff} and the DIBL coefficient. Fig. 9 illustrates the relationship between

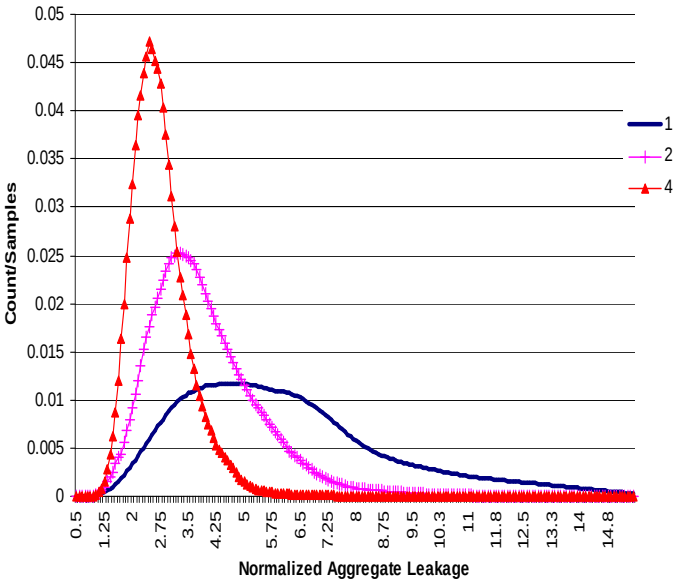

Figure 6: Normalized aggregate leakage across 1, 2, and 4 transistors. Each distribution is normalized to the smallest value in the distribution so that variances can be compared.

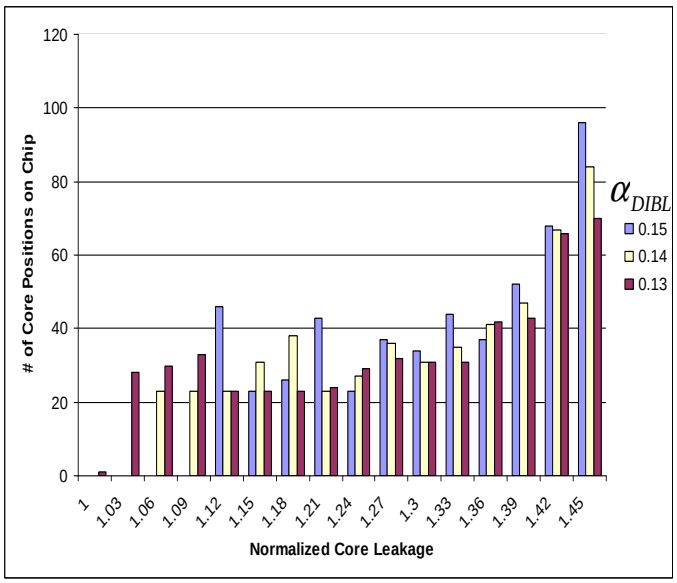

Figure 7: Leakage distribution of all possible core positions on a die's surface for different α_{DIBL} values.

these parameters and leakage. In all leakage calculations, the feedback loop between temperature and leakage has been closed. Thermal calculations were performed using Hotspot [16]. By modeling the thermal-leakage feedback, more accurate leakage estimations can be obtained, since leakier cores will have a higher power density, and therefore, higher temperatures than cores with less leakage. Since leakage is exponentially dependent on temperature, the feedback loop will exacerbate C2C leakage variation. It is also important to recognize that, because of the characteristics of the polynomial equation used to model systematic variation, the worst-case leakage value is independent of the DIBL coefficient: worst-case leakage occurs in area of the die having nominal L_{eff}. Changing the DIBL coefficient results in more C2C leakage variation because this parameter determines the leakage of the core that is located in the area of the die having the largest L_{eff} value. The leakage in this area of the die will increase with the α_{DIBL}. Simply put, parameter values that are good for performance (small L_{eff} and V_{th} values) are worse for leakage.

The situation that we analyze is one in which the entire exposure field is 28mm by 28mm, with the reticle being comprised of 4 identical 14mm by 14mm dies. The WID_{sys} L_{eff} pattern is transposed onto a grid and the resulting across chip systematic L_{eff} pattern is depicted in Fig. 8. This was derived using Eq. 5, and the baseline constants in Table 2. We consider a POWER4-like core scaled to 45nm dimensions. Assuming constant scaling, the core area will be 2.5mm by 2.25mm. In order to gather the leakage distribution, all possible core positions on the chip's surface are considered. Sub-threshold leakage is determined by taking the aggregate sum of the leakage in the core's underlying grid cells. The C2C leakage distribution for all possible core positions on a die is shown in Fig. 7. The skewed C2C leakage distribution occurs because of the polynomial nature of the systematic equation variation the resulting distribution is negatively skewed. As mentioned earlier, closing the thermal feedback loop exacerbates C2C leakage.

In contrast, random variations in leakage will not be of particular interest in the architectural domain. The reason for this is that the leakage in a core and unit is an aggregate sum of the leakage in the underlying transistors. When a summation is taken across a large enough sample, very little variation in the mean will occur because of the "averaging effect" that occurs when a sum operation is performed on random variables.

6. CONCLUSIONS AND FUTURE WORK

This paper presents a model that allows microarchitects to reason about how WID process variations may affect a multi-core environment. The model is based on an abstract representation of combinational logic and SRAM structures, and accounts for the way logic depth (L_{cp}) and the number of independent critical paths (N_{cp}) affect delay distribution. Using the model, this paper shows that:

- Unit-to-unit variations within a single core are likely to be dominated by SRAM structures.
- WID random variations will not materially affect the C2C distributions—all each core is likely to be impacted by random variations to a similar degree.
- The impact of WID systematic variations on the C2C frequency distribution will be minimal.
- The exponential relationship of leakage on V_{th}, and of V_{th} on L_{eff}, means that WID_{sys} variations will produce C2C leakage variations up to 45%.

These results suggest that pre-RTL PVT modeling is important for future multi-core designs. The goal of this work is not to dismiss the importance of random variations within individual cores, but rather to argue that the impact of *random* variations chiefly manifests at a *circuit* level of abstraction, where optimizing the length and number of critical paths will be most fruitful. The impact of *systematic* variations, on the other hand, chiefly manifests at an *architectural level*. Our results suggest that the real focus of architectural techniques for addressing WID process variations should therefore explore variation-tolerant integration of cores, rather than

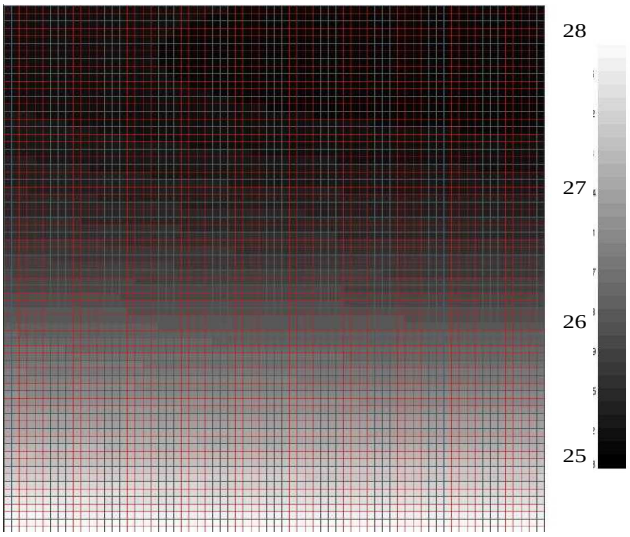

Figure 8: 2D contour map of across chip L_{eff} variation in nm

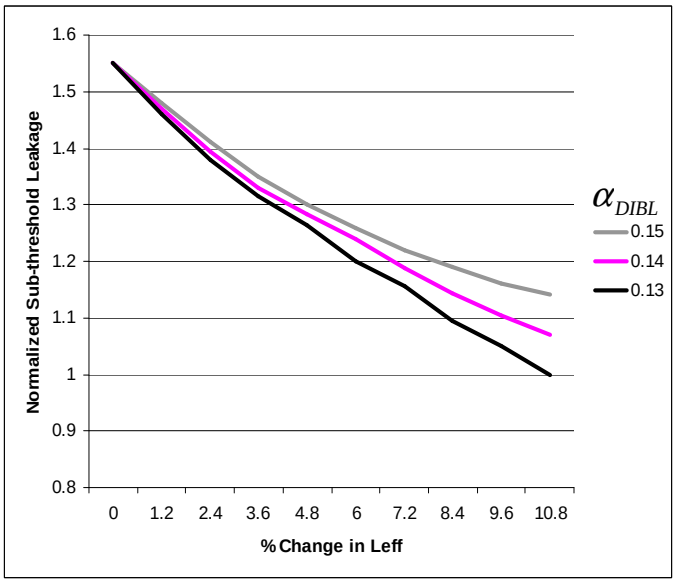

Figure 9: Plot shows the dependency between sub-threshold leakage, L_{eff}, and the DIBL coefficient. For each α_{DIBL} value, the value of V_{th0} in Eq.6 was modified so that the nominal value of V_{th} is always .2V. By doing this, across chip leakage variation comparisons can be made for different α_{DIBL} values.

within-core techniques for balancing out variation among units. This might involve novel leakage- and temperature-aware scheduling techniques, the capability for multiple cores to operate at independent voltage and frequency, variation-aware per-core leakage-mitigation techniques, and so forth. Note that techniques like Razor [9] may still be needed to reclaim excess design margins. Razor-like techniques also provide the opportunity to design for typical-case variations, relying on error-recovery support like Razor shadow latches for exceptional runtime conditions and unusual input values. Our conclusions are predicated on an "FMAX" assumption, namely that the clock speed is determined by the worst-case delay through any critical stage (or nearly so, e.g. 3σ). The opportunities for within-core variation mitigation are larger if the clock speed is in fact determined by average- or common-case delay, which will cause many paths to violate timing integrity (sometimes intermittently, for paths where temperature and voltage are the determining factor). In addition to Razor, a variety of other fault-tolerance techniques may be helpful.

Improving the model's fidelity is an obvious direction for future work. Exploring the relationship between WID_{sys} correlation distance and core size is an especially important aspect. A sensitivity study on the impact of different magnitudes of the random and systematic variation phenomena is also needed. Extending the model account for D2D and W2W variations may be valuable too, as architectural techniques may be able to mitigate these effects.

Acknowledgments

This work has been supported in part by NSF grant nos. CCR-0133634 (CAREER), CCF-0429765, Army Research Office grant W911NF-04-1-0288, a research grant from Intel MTL, and an IBM Faculty Partnership Award. The authors would like to thank Wei Huang for his valuable assistance on the paper , and also the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, and Y. Ismail. Statistical static timing analysis: how simple can we get? In *Proceedings of the 42nd Annual International Conference on Design and Automation*, June 2005.

[2] M. Ashouei, A. Chatterjee, A. D. Singh, V. De, and T. M. Mak. Statistical estimation of correlated leakage power variation and its application to leakage-aware design. In *Proceedings of the 19th International Conference on VLSI Design (VLSI Design 2006)*, Jan. 2006.

[3] S. Borkar, T. Karnik, and V. De. Design and reliability challenges in nanometer technologies. In *Proceedings of the 41st Annual International Conference on Design and Automation*, June 2004.

[4] K. Bowman, S. Duvall, and J. Meindl. Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration. *IEEE Journal of Solid State Electronics*, 37(2), February 2002.

[5] Berkeley predictive technology model. http://www-device.eecs.berkeley.edu.

[6] J. Cain. Characterization of spatial variability in photolithography. Master's thesis, Univ. of California, Berkeley EECS Dept., Nov. 2002.

[7] Y. Cao and L. T. Clark. Mapping statistical process variations toward circuit performance variability: an analytical modeling approach. In *Proceedings of the 42nd Annual International Conference on Design and Automation*, June 2005.

[8] Y. Cao, P. Gupta, A. B. Kahng, D. Sylvester, and J. Yang. Design sensitivities to variability: Extrapolation and assessments in nanometer vlsi. In *Proceedings of the IEEE ASIC/SoC Conf.*, pages 411–415, Sep. 2002.

[9] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge. Razor: A low-power pipeline based on circuit-level timing speculation. In *Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture*, page 7, Dec. 2003.

[10] M. A. Horowitz. Timing model for MOS circuits. In

Technical Report SEL83-003, 1983.

[11] W. Huang, E. Humenay, K. Skadron, and M. R. Stan. The need for a full-chip and package thermal model for thermally optimized ic designs. In *Proceedings of the 2005 ACM/IEEE International Symposium on Low Power Electronics and Design*, pages 245–250, New York, NY, USA, 2005. ACM Press.

[12] Z. Huang and M. D. Ercegovac. Effect of wire delay on the design of prefix adders in deep-submicron technology. In *Proceedings of the 34th Asilomar Conference on Signals, Systems, and Computers*, Oct. 2000.

[13] D. Marculescu and E. Talpes. Variability and energy awareness: a microarchitecture-level perspective. In *Proceedings of the 42nd Annual International Conference on Design and Automation*, June 2005.

[14] M. Orshanksy, L. Milor, and C. Hu. Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction. *IEEE Transactions on Semiconductor Manufacturing*, 17(1), Feb. 2004.

[15] SIA. *International Technology Roadmap for Semiconductors*, 2004. http://public.itrs.net.

[16] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture. In *Proceedings of the 30th Annual ACM/IEEE International Symposium on Computer Architecture*, pages 2–13, New York, NY, USA, 2003. ACM Press.

[17] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and V. De. Die-to-die and within-die parameter variations on microprocessor frequency and leakage. *IEEE Journal of Solid-State Circuits*, 37(11), Nov. 2002.

[18] S. Zhang, V. Wason, and K. Banerjee. A probabilistic framework to estimate full-chips subthreshold leakage power distribution considering within-die and die-to-die P-T-V variations. In *Proceedings of the 2004 ACM/IEEE International Symposium on Low Power Electronics and Design*, pages 156–161, Aug. 2004.