Toward an Architectural Treatment of Parameter Variations

UNIV. OF VIRGINIA DEPT. OF COMPUTER SCIENCE TECH. REPORT CS-2005-16

SEPT. 2005

Eric Humenay*, Wei Huang†, Mircea R. Stan†, and Kevin Skadron*
Depts. of *Computer Science and †Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22904
humenay@virginia.edu, weihuag@virginia.edu, skadron@cs.virginia.edu

Abstract

This paper develops a new model of parameter variations for use in early-stage, pre-RTL architecture studies. It improves over prior models by extending the FMAX model to more faithfully model various microarchitecture structures, especially SRAM, which is dominant in contemporary superscalar processors. It also incorporates optical phenomena, which show strong spatial correlation but nevertheless cannot be ignored for large dies. Finally, it incorporates IR Vdd drop and temperature, and closes all these feedback loops to obtain converged estimates of frequency, leakage, voltage, and temperature. With this model, we explore PVT limitations on multi-core integration and the difficulties in obtaining matched cores.

1 Introduction

The 2004 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, both ITRS and some industry presentations, e.g. [1], have pointed out that computer architecture plays an essential role in mitigating parameter variations. Architectural analysis and design, however, is often carried out in the earliest design stages of a chip, before a physical design or even RTL description is available. This necessitates *pre-RTL* modeling capability.

Parameter variations encompass a range of variation types, including process variations due to manufacturing phenomena, temperature variations, and voltage variations. Process variations manifest as both die-to-die (D2D) and within-die (WID) variations, while temperature and voltage variations are primarily WID phenomena. Temperature variations stem from different activity factors among functional units, from different circuit structures and hence different power densities among functional units, and from non-uniformities in the 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, and activity-dependent IR drops due to switching activity and non-ideal decoupling capacitance.

While both D2D and WID effects can be addressed by architecture changes, this paper focuses on the WID variations, because temperature and voltage variations are primarily WID phenomena and interact with WID process variations in interesting ways. (D2D variations can be treated as a random offset.)

This paper describes a modeling methodology that is compatible with pre-RTL architecture analysis. It accounts for process, voltage, and temperature (PVT) variations, including both systematic and random process variations. The model for now focuses on phenomena that affect threshold voltage (V_{th}), leaving for future work additional phenomena like variations in interconnect properties. Changes in V_{th} in turn affect device speed and leakage. We also account for extrinsic temperature and voltage variations, which also affect device speed and leakage. Leakage in turn affects temperature, creating a feedback loop.

Our study uses this microarchitectural PVT model to explore integration of multiple CPUs into a multi-core processor and the impact of closing the PVT-leakage-performance feedback loops on operating frequency, voltage, temperature, and leakage.

2 Related Work

Historically, variations in gate length, L_{eff} (also referred to as critical dimension—CD—variations), have been modeled as independent random gaussian distributions [19]. Recent studies have concluded that a significant portion of WID variation is systematic in nature [7, 16], and that it

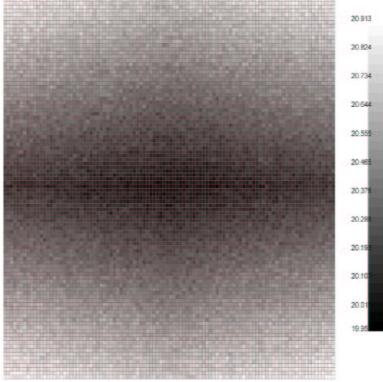

Figure 1. CD Map.

is necessary to take the device's on-chip location into account when estimating delay. In [16], systematic characterization of layout patterns were extracted from dies fabricated in a 180nm step-and-repeat exposure system. Their results showed that every die, regardless of its position on the wafer, had similar variation patterns across its surface. The resulting on-chip delay distribution was spherical in nature with the dies' corners being 25% slower than in the middle of the chip. In [7], CD maps were gathered from a more mature step-and-scan exposure system. While systematic variations were prevalent, more interestingly was how the average CD map differed from that of the previously mentioned process. In Fig. 1, a sample CD map for a step-and-scan system is shown. The CD map, derived from our own, lumped process variation model, is based on measurements and conclusions from [7]. An important observation that can be seen in the figure is that more variation occurs in the vertical direction than the horizontal direction. Notice that our lumped process variation model also takes into account the random variations as well, which can be seen from the random disturbs in the CD map.

In [2], a predictive model for estimating frequency distribution is presented. The "FMAX" model is comprised of a generic critical path model, GCP, that was validated with measured data from a .25um process. In [13], the authors extend the FMAX model by assuming number of critical paths per stage, N_{cp} , is proportional to the stages device count as well as introducing metrics and models to evaluate variability in the micro-architectural domain. The proposed model differs substantially from previous work by better modeling SRAM blocks, which dominate in modern superscalar processors. Other key differences in our work are inclusion of optical phenomena, which are spatially correlated but still affect large dies, and incorporation of voltage and temperature to close all the PVT feedback loops. The only other microarchitectural parameter-variation research of which we are aware, [5], presents a statistical methodology for pipeline delay analysis to show the importance of logic depth in variability studies.

3 Variation Model

3.1 Process Variation Model

To explore the impact of within-die variations we model the die as being an *i* by *j* matrix where *i* is the number of columns and j is the number of rows in the matrix. Initially, each cell in the matrix is assigned a value for L_{eff}

$$L_{eff_{ij}} = L_{eff_{nom}} + \Delta Rand_{ij} + \Delta Syst_{ij} \qquad (1)$$

Where Δ $Syst_{ij}$ is the deviations in L_{eff} that are systematic across the die and are a function of the illumination system. δ $Rand_{ij}$ are normally distributed random perturbations in L_{eff} that result from fabrication phenomenas such as line edge roughness, LER.

Threshold voltage, V_{th}, plays a major role in determining both leakage and delay therefore making it necessary to be included in variation models. V_{th} is modeled similarly to L_{eff} where

$$V_{th_{ij}} = V_{th_{nom}} + \Delta V_{th_{sys}} + \Delta V_{th_{rand}} \qquad (2)$$

Δ $V_{th_{rand}}$ is a result of fluctuations in dopant densities from and Δ $V_{th_{sys}}$ occurs due to the the dependency of threshold voltage on L_{eff}. In [4] the authors present us with a equation for determining V_{th} as a function of L_{eff}:

$$V_{th_{sys}} = V_{th0} - V_{dd} \cdot exp(-\alpha_{DIBL} \cdot L_{eff}) \qquad (3)$$

Where V_{th0} is the threshold voltage for long channel transistors, α_{DIBL} is the DIBL coefficient, and V_{dd} is the supply voltage.

3.2 Within-Die Temperature and Supply Voltage Variations

In addition to process variations due to manufacturing phenomena, there are also temperature variations and voltage variations. Investigations on micro-architecture variation-aware techniques are certainly incomplete if neglecting the effect of within-die temperature and voltage variations. This is because sub-threshold leakage power is exponentially dependent on operating temperature, transistor carrier mobility hence the delay is also dependent on temperature, and delay is proportional to supply voltage.

3.3 Temperature Variations

We investigate the effect of temperature variations using the existing HotSpot thermal model [18] that has been widely adopted in architecture research. HotSpot is able to model temperature distributions at the within-core functional unit level for arbitrary floorplan with power dissipations as the input. Because subthreshold leakage power is

exponentially dependent on temperature, and in turn temperature is also determined by the overall power dissipation, there is a feedback loop between leakage power and temperature [23]. In order to get accurate temperature and leakage estimations, this well-known loop has to be modeled as well. Considering the manufacture process variations in L_{eff} and V_{th}, the distribution of leakage current for each transistor is log-normal. This is because we have assumed V_{th} has a normal distribution, and it is well-known that the exponential of a normal random variable is a log-normal distribution. Based on the temperature dependent subthreshold leakage equation given in [20], we have a subthreshold leakage power model at the granularity of within-core functional unit as the following:

$$P = N_{tran} \times V_{dd} \mu C_{ox} \frac{W_{eff}}{avgL_{eff}} (m-1)(v_T)2 \times e^{\gamma + \beta (T - T_{ref})} \times (1 - e^{-V_{DS}/v_T}) \quad (4)$$

where N_{tran} is the approximated transistor count inside each functional unit and can be found from ITRS. V_{dd} is supply voltage, μ is the carrier mobility, C_{ox} is the gate oxide, capacitance, W is the average drawn width of each transistor, $avgL_{eff}$ is the average effective gate length for each functional unit based on the CD map of L_{eff} variations, m is a technology-dependent factor that is slightly greater than 1.0 [17], v_T is the thermal voltage equal, β is the temperature factor of subthreshold leakage current, which equals to 0.0085 (typical values of β can be derived from [8]). T is operating temperature, T_{ref} is the reference temperature at which β is derived. V_g is the gate voltage, V_th is the threshold voltage, and V_{DS} is the drain voltage. γ[1] is the equivalent exponent so that

$$e^{\gamma} = \Sigma_{i=1}^{N_{tran}} e^{(V_g - V_{th})/mv_T} \quad (5)$$

Values for V_{th} are extracted from the units underlying grid cells. Each unit has enough corresponding grid cells to obey by the rules of the Central Limit Theorem.

With Eq. 4 and the HotSpot thermal model, we are now able to close the loop between temperature and subthreshold leakage power with the underlying process variations. Usually, it only take about 5 iterations of the loop for temperature and leakage power to converge.

Temperature variations have significant impact on delay, because transistor's carrier mobility is dependent on temperature, and gate delay is approximately inversely proportional to carrier mobility. [2] A first-order model for the relationship between carrier mobility (μ) and temperature (T) can be expressed as the following:

$$\mu(T) = a * T^b \quad (6)$$

where a and b are fitting coefficient to measured carrier mobility for different technologies [21]. Here we use 1.15×10^4 for a, and -2.2 for b, mobility is in the unit of m^2/V-s.

From Eq. 6, it is clear that mobility decreases when temperature increases, thus the delay is worse at higher temperature. For example, delay is about 46% longer at 110C than at 50C. Without considering temperature dependency of carrier mobility, performance estimations is not accurate.

3.4 Voltage Variations

Due to the unfortunate fact that power supply is distributed by layers of metal wires, it is inevitable that there is some voltage IR drop on the power supply network, resulting in less-than-nominal V_{dd} for transistors. ITRS Roadmap requires at most 5% of V_{dd} drop for different technologies. We have developed a lumped power supply network model that can estimate voltage IR drop at the micro-architecture level. The model takes in power consumptions by each functional units and the technology-dependent power supply routing information from ITRS predictions, and output averaged actual V_{dd} estimations for each functional units. The resulting V_{dd} map is fed back to performance and leakage models to get more accurate results.[3]

3.5 Architectural Block-Delay Model

To account for the microarchitectural impact of PVT, we have chosen to use an FMAX-based delay model similar to the approach used for variability studies in [13]. It was necessary to slightly modify the FMAX model in order to take into account delay resulting from systematic variations. In [2], the equation for maximum critical path delay is given as

$$T_{cp,max} = T_{cp,nom} + \Delta T_{D2D} + \Delta T_{WID} \quad (7)$$

where $T_{cp,nom}$ is the nominal critical path delay, and ΔT_{D2D} and ΔT_{WID} are the deviations in block delay that are a result of D2D and WID variations. We have modified the orginial WID variation model in order to include systematic variations such that

$$\Delta T_{WID} = \Delta T_{rand} + \Delta T_{sys} \quad (8)$$

Where ΔT_{rand} and ΔT_{sys} are the deviations in the nominal critical path delay that arise from random and systematic variations.

[1] γ can be calculated using Wilkinson's method to match the first two moments of sum of log-normal transistor leakage current distributions. More details is shown in [22]

[2] The delay of a gate is inversely proportional to the saturation current of transistors, and the saturation current of a transistor is proportional to mobility, therefore, lower transistor mobility results in longer gate delay.

[3] Here, we find the fact that the temperature dependency of IR drop is negligible, firstly because the temperature dependency of the metal resistivity is not very strong, and secondly V_{dd} drop is mostly a very localized phenomenon, and cannot be fully captured at the granularity of functional units. Therefore, modeling IR drop is accurate enough at the microarchitecture level by assuming no temperature variations.

At the heart of the FMAX delay model is the generic critical path model,GCP. The GCP is comprised of a chain of 2 input CMOS Nand gates with a fanout of three and of depth n_{cp}. The delay of a Nand gate, T_{nand} is the average propogation delay through 2 series NFETS and the delay through one PFET as derived from the physical alpha power based law-model. The delay for the critical path is then given as:

$$T_{cp} = n_{cp} T_{nand} \tag{9}$$

In order to gain a better understanding of the relationship between n_{cp},ΔT_{rand}, and ΔT_{sys} we have chosen to model the critical path delay as a summation of the delay of n_{cp} different Nand gates.

$$T_{cp} = \sum_{i=0}^{n_{cp}} T_{nand_i} \tag{10}$$

As mentioned in [13], representing each block's critical path delay as a chain of Nand gates provides only a lower bound on propogation delay since many blocks will have some portion of their delay spent in local interconnects. We have developed a first order model that assigns a n_{cp} to each block based on the block's circuit type and area. In our model we have categorized all blocks as being of type *SRAM* or *logic*. An SRAM block's $n_{cp} \sim$ block area with the largest block in the pipeline, the L1 data cache, spending 25% of $T_{cp,nom}$ in the wires and the smallest SRAM block, cast-out queue, spending 50%. For logic blocks such as the execution units and decode unit, the amount of wire delay is held constant at 10% of $T_{cp,nom}$. We have extended the GCP critical path model to take into account 6T SRAM cells such that the critical path for an SRAM block is:

$$T_{cp,sram} = \sum_{i=0}^{n_{cp}} T_{nand_i} + T_{cellaccess} \tag{11}$$

where $T_{cellaccess}$ is the access delay of an sram cell. The equation to calculate cell access time is provided in [14]. Values for N_{cp} are assumed proportional to device count.

3.5.1 Simulation Design Flow

The previously mentioned temperature-leakage feedback loop is placed inside of a larger loop that calculates core frequency when PVT variations are considered. Initially. the critical path model uses default values for T and V when calculating the per-core frequency and dynamic power dissipation. The equation for dynamic power dissipation scaling is

$$Power_{dynamic} = F * V_{dd}^2 * C * A \tag{12}$$

where F is the frequency, V_{dd} is the supply voltage, C is the relative capacitance, and A is the functional unit's activity factor. A and V_{dd} are pre-determined input parameters, C is a function of the CD map, and F is a dynamic variable that converges at a final value at the end of the simulation.

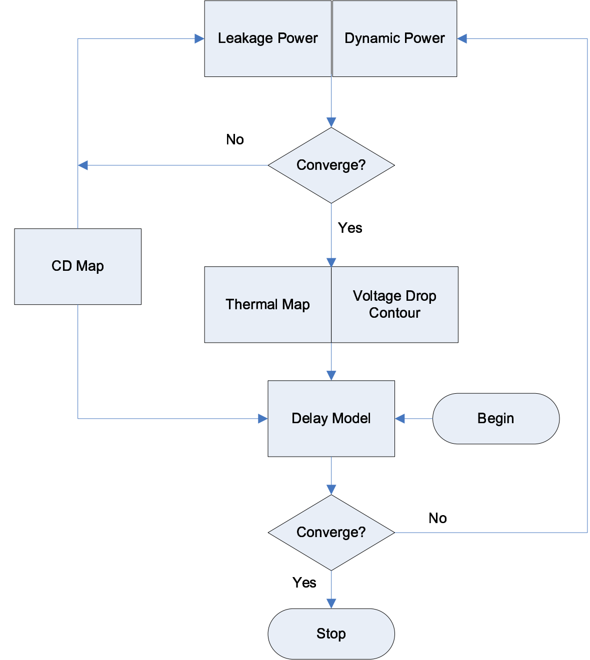

Figure 2. Delay-Temperature-V_{dd} Flow Chart

The newly calculated dynamic power dissipation and leakage is fed as input to the inner thermal-leakage loop. Once the thermal-leakage loop has converged the new value for T and V_{dd} are inputs into the delay model. The entire loop is iterated until all variables have converged.

In Figure 3 frequency convergence has been graphically illustrated. At iteration 1, the frequency is calculated with the inputs being the CD map and the default values for T and V. In iteration 2, the frequency is drastically reduced since the T and V values from the previous iteration adversely affect block delay. The slower frequency results in lower dynamic power dissipation, lower T (therefore lower leakage), and higher VDD values. In iteration 3, the frequency rises again since the previous iteration produced lower T values and higher VDD values. This oscilating pattern converges at final values at roughly 10 iterations.

4 Experimental Methodology

For our studies we have chosen to model a POWER4-like core similar to the one presented in [11] with the main difference being that we have chosen to model blocks at a per stage granularity. Are pipeline is 17 stages deep and is derived from the single-threaded pipeline model in [12]. Because different units execute in parallel in a pipeline there will always be more units that require their delay to be less than the cycle time than pipeline stages, and this number increases in a clustered micro-architecture. In our pipeline model we assume 33 different time-critical units. Core area

Figure 3. Average core frequency as the number of iterations in delay-thermal loop increases

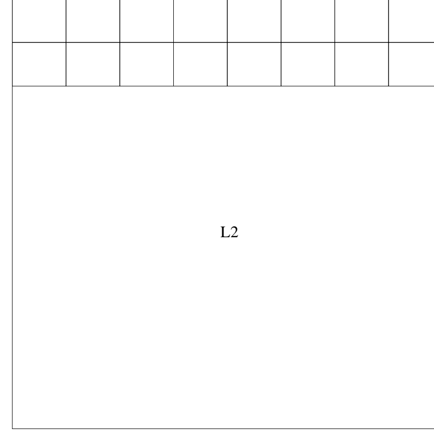

Figure 4. Floorplan for 16 core chip.The upper right hand corner of each core is the location of the hotspot,integer and load/store execution unit.

has been scaled to 50nm and we have assumed a die consisting of 16 cores. The activity factor in 12 was calculated using Turandot/PowerTimer [15, 9, 3]. For simplicity, we assumed all cores were executing the GCC benchmark, and only the average activity factor over the entire benchmarks lifetime was considered.

5 Results

In table 1 the impact of different variations are shown. P_s,P_r, and P_{rs} are results when considering only systematic, only random, and both random and systematic process variations. The results for these parameters as well as for no variations have had their delay calculated and fed into the thermal-leakage feedback loop. That is they have not progressed past iteration 1 in the thermal-delay feedback loop. Most interesting, is the dramatic difference in leakage that can be had from closing the thermal-delay loop. The reason for the drastic decrease in leakage is that dynamic power is scaled in proportion to frequency and this results in cooler-on-chip temperatures.

Systematic Variation studies have reported a wide range of values with the variation from the middle of the outer edge of the die being 3-4% of nominal CD values in [16, 6] to 25% in [7]. Since it is unclear how systematic variations will scale with technologies we have swept the variation amount from an optimistic 3% to a pessimistic 25%. We have assumed random variations for L_{eff} and V_{th} to be $\pm$10% and $\pm$30% respectively. The previously mentioned floorplan is compared to a floorplan that has its cores located in the middle of the die. As noted earlier, not only is it imperative that cores are placed in positions that will maximize their frequency but also in locations that will minimize core-to-core frequency variation. In order to do this, designers must have knowledge about the fabrication process's CD map. As can be seen in Figure 5 by positioning the cores in the middle of the die the frequency degradation is mitigated. Also, by placing each row of core an equal distance from the center core-to-core frequency variation will be minimized.

Taking a closer look at how the different types of variations affect the units in the in the pipeleline will provide us with better insight into the significance of the variations. The following experiment has its cores located in the middle of the die, but unlike the previously mentioned floorplan each row of cores is now a mirrored copy of the other row. With each set of assumptions about variation types the slowest unit in the pipeline changes. Figure 6 shows why delay models are incomplete if they do not consider all sources of PVT variation. When only random variations are considered the L1 Data cache is the slowest structure because it las the largets number of critical paths, N_{cp}, out of all the untis considered. If only systematic variations are considered the fixed unit arithmetic logic is has the greatest delay because it is located far away from the center of the die, and it is a unit of type *logic* so therefore it has a large n_{cp}. As logic depth increases the random variations will be averaged out leaving only systematic variation to impact delay. When both random and systematic variations are considered the floating point register file is slightly the worse performing unit because out of all the sram units in the core it is located farthest away from the center of the die. Finally, when the full temperature-delay loop has converged the Fixed Unit,FXU, register file is the slowest unit because it is the hottest.

Creating homogenous core performance in the presence of systematic and random variations is a daunting task;therefore, requiring strategies that can create hetero-

Variations	Ave. Core Frequency(GHz)	Leakage Power(W)	Dynamic Power(W)	Hottest Temp(k)
No Variation	3.0	155.95	181.2	400
P_r	2.75	202.15	173.13	408
P_s	2.85	140.1	179.5	396.5
P_{rs}	2.67	190	173.3	402
$P_{rs}VT$	2.5	110.5	96.62	367.8

Table 1. Comparisons of how different combinations of PVT variations affect frequency,power, and temperature

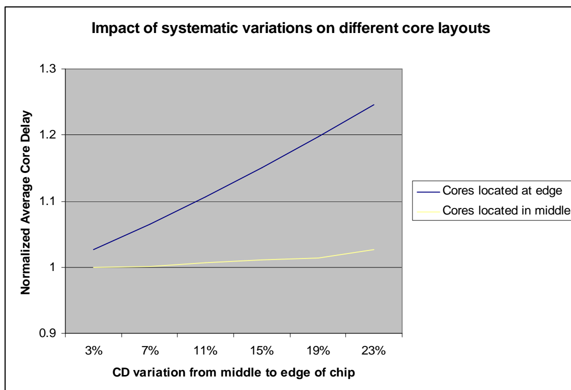

Figure 5. Comparison of systematic variation induced delay in two different core configurations.

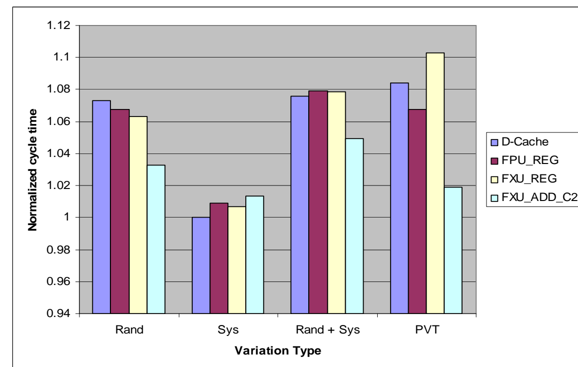

Figure 6. Change in the slowest microarchitecture unit considering different variations.

genity while sacrificing as little power and performance as possible. We have extended our grid model to examine the consequencs of using forward body biasing to speed up slower cores. Unfortunately, since the core's slowest units located in the core's hotspot only marginal gains in performance can be expected before thermal runaway occurs 7 . Figure 8 shows the importance of using the grid-based model in order to gain more accurate leakage estimations when variations are considered.

6 Conclusions and Future Work

An architectural model to estimate delay, power dissipation, and temperature as a function of within-die PVT variations allows early-stage architecture to consider limitations imposed by PVT and explore new architecture techniques to mitigate PVT effects. This enables a powerful set of optimizations that complement techniques in the circuit and manufacturing realms, because early architecture decisions define the specification that must subsequently be implemented.

Our results demonstrate the importance of including both random and spatially-correlated WID variations in an architectural model. The resulting model shows that PVT variations limit core placement for multi-core chips and hence may limit the degree of integration that can be achieved, especially if homogeneous performance is a requirement. In fact, our results call into doubt the ability to make all cores match the nominal frequency target without incurring prohibitive leakage in some cores. This suggests the use of a heterogeneous architecture, but also calls for architecture innovations that can reduce the impact of PVT. Redundancy is particularly attractive for coping with random variations. In fact, there is likely substantial synergy between redundancy techniques for reliability and the ability to mitigate PVT effects.

Finally, our results show the importance of closing the loop between process variations, leakage, temperature, voltage, and frequency. Current architecture power/performance methodolgies neglect these effects, with substantial inaccuracy a likely result.

Clearly the model can be improved in many ways by tying it more closely to the likely circuit implementations of the various blocks in question. Nevertheless, we propose the model in its current form because we believe it illustrates several important ways in which architecture decisions must take PVT into account, and because the model

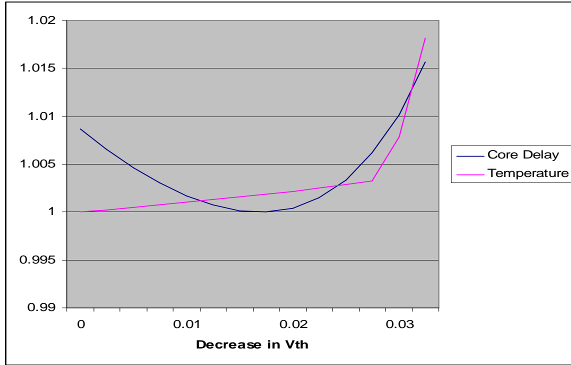

Figure 7. Normalized cycle time, leakage power, and hotspot temperature as threshold voltage is decreased

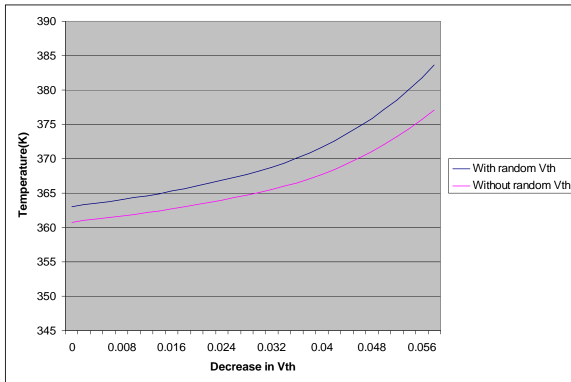

Figure 8. Underestimation of temperature during body biasing when random V_{th} is not modeled

enables new architecture research to cope with PVT variations.

Acknowledgements

This work is supported in part by NSF grants CCR-0133634 (CAREER), CCF-0429765, Army Research Office grant W911NF-04-1-0288, a Univ. of Virginia FEST Award, two research grants from Intel MRL, and an IBM Faculty Partnership Award. The authors would also like to thank Karthik Sankaranarayanan, Dee A. B. Weikle, and Yan Zhang for their assistance with the preparation of this report.

References

[1] S. Borkar, T. Karnik, and V.De. "Design and Reliability Challenges in Nanometer Technologies." 41st DAC, June 2004.

[2] K.A. Bowman, S.G. Duvall, J.M. Meindl. "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration." *IEEE J. Solid State Electronics*, 37(2), Feb. 2002

[3] D. Brooks and P. Bose and V. Srinivasan and M. K. Gschwind and P. G. Emma and M. G. Rosenfield. "New methodology for early-stage, microarchitecture-level power-performance analysis of microprocessors." *IBM J. R & D,* 47(5-6), 2003

[4] Y. Cao and L. T. Clark. "Mapping statistical process variations toward circuit performance variability: an analytical modeling approach." 42nd DAC, 2005

[5] A. Datta, S. Bhunia, S. Mukhopadhyay, N. Banerjee, and K. Roy. "Statistical Modeling of Pipeline Delay and Design of Pipeline under Process Variation to Enhance Yield in sub-100nm Technologies." DATE, Feb. 2005.

[6] P. Fernandez-Martinez, T. Marschner, C. Fulber, J.W. Bartha. "Characterisation of Lithography Performance in high end Semiconductor Manufacturing with Electrical Linewidth Measurements." 7th AEC, 2004

[7] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos. "Modeling Within-Die Spatial Correlation Effects for Process-Design Co-Optimization." ISQED, Mar. 2005.

[8] S. Heo, K. Barr, and K. Asanovic. "Reducing Power Density through Activity Migration." ISLPED, Aug. 2003

[9] Z. Hu, D. Brooks, V. Zyuban, and P. Bose. "Microarchitecture-level power-performance simulators: Modeling, validation, and impact on design." Tutorial, MICRO-36, Dec. 2003.

[10] W. Huang, M. R. Stan, K. Skadron, K. Sankaranarayanan, S. Ghosh, and S.Velusamy. "Compact thermal modeling for temperature-aware design." 41st DAC, June 2004

[11] Y. Li, D. Brooks, Zhigang Hu, and K. Skadron. "Performance, Energy, and Thermal Considerations for SMT and CMP Architectures." HPCA, Feb. 2005

[12] Y. Li, D. Brooks, Z. Hu, K. Skadron, and P. Bose. "Understanding the energy efficiency of simultaneous multithreading." ISLPED, Aug. 2004

[13] D. Marculescu and E. Talpes. "Variability and energy awareness: a microarchitecture-level perspective." 42nd DAC, June 2005.

[14] S. Mukhopadhyay, H. Mahmoodi-Meimand, and K. Roy "Modeling and Estimation of Failure Probability due to Parameter Variations in Nano-Scale SRAMS for Yield Enhancement." VLSI Circuits Symp., June 2004.

[15] M. Moudgill, J.-D. Wellman, and J. H. Moreno. "Environment for PowerPC microarchitecture exploration." *IEEE Micro*, 19(3):15-25, 1999.

[16] 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 Trans. Semiconductor Manufacturing, Vol. 17,No. 1, 2004

[17] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand. "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits " *Proc. of the IEEE*, 91(2):305–27, Feb. 2003.

[18] K. Skadron, M.R. Stan, W. Huang, S. Velusamy; K. Sankaranarayanan, and D. Tarjan. "Temperature-aware microarchitecture." ISCA-30, June 2003.

[19] B. Stine, D. S. Boning, and J. E. Chung. "Analaysis and Decomposition of spatial variation in integrated circuit processes and devices." IEEE Trans. Semiconductor Manuf., 10:24.

[20] Y. Taur and T.H. Ning. *Fundamentals of Modern VLSI Devices* Cambridge Univ. Press, 1998

[21] R. Pierret. *Semiconductor Device Fundamentals* Prentice Hall, 1995

[22] H. Chang and S. Sapatnekar "Full-chip analysis of leakage power under process variations, including spatial correlations." 42nd DAC, June 2005.

[23] H. Su, F. Liu, A. Devgan, E. Acar, and S. Nassif "Full chip leakage estimation considering power supply and temperature variations." ISLPED, Aug. 2003.