Addressing Process Variations at the Microarchitecture and System Level

Addressing Process Variations at the Microarchitecture and System Level

Siddharth Garg

University of Waterloo
Canada
s6garg@ecemail.uwaterloo.ca

Diana Marculescu

Carnegie Mellon University
USA
dianam@ece.cmu.edu

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Electronic Design Automation

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is S. Garg and D. Marculescu, Addressing Process Variations at the Microarchitecture and System Level, Foundations and Trends® in Electronic Design Automation, vol 6, no 3, pp 217–291, 2012

ISBN: 978-1-60198-658-0

Foundations and Trends® in Electronic Design Automation
Volume 6 Issue 3, 2012

Editorial Board

Editorial Scope

Foundations and Trends® in Electronic Design Automation will publish survey and tutorial articles in the following topics:

- System Level Design
- Behavioral Synthesis
- Logic Design
- Verification
- Test
- Physical Design
- Circuit Level Design
- Reconfigurable Systems
- Analog Design

Information for Librarians
Foundations and Trends® in Electronic Design Automation, 2012, Volume 6, 4 issues. ISSN paper version 1551-3939. ISSN online version 1551-3947. Also available as a combined paper and online subscription.

Foundations and Trends® in
Electronic Design Automation
Vol. 6, No. 3 (2012) 217–291

DOI: 10.1561/1000000031

now
the essence of knowledge

Addressing Process Variations at the Microarchitecture and System Level

Siddharth Garg[1] and Diana Marculescu[2]

[1] *University of Waterloo, 200 University Ave. West, Waterloo, Ontario N2L 3G1, Canada, s6garg@ecemail.uwaterloo.ca*
[2] *Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA, dianam@ece.cmu.edu*

Abstract

Technology scaling has resulted in an increasing magnitude of and sensitivity to manufacturing process variations. This has led to the adoption of statistical design methodologies as opposed to conventional static design techniques. At the same time, increasing design complexity has motivated a shift toward higher levels of design abstraction, i.e., micro-architecture and system level design. In this survey, we highlight emerging statistical design techniques targeted toward the analysis and mitigation of process variation at the system level design abstraction, for both conventional planar and emerging 3D integrated circuits. The topics covered include variability macro-modeling for logic modules, system level variability analysis for multi-core systems, and system level variability mitigation techniques. We conclude with some pointers toward future research directions.

Contents

1

Introduction

The relentless scaling of transistor dimensions in every technology generation has enabled unprecedented levels of on-chip integration. Increased integration density reduces the cost per transistor and enables greater functionality to be packed within the same silicon area. At the same time, transistor scaling also results in lower switching energy per transistor and reduced intrinsic delay.

However, the benefits of transistor scaling are accompanied by a number emerging challenges. One of the most important challenges that accompanies transistor scaling is the problem of manufacturing process variations, referred to simply as process variations in the rest of this survey. In the broadest sense, process variations refer to the mismatch between a transistor's physical and electrical parameters after manufacturing on the one hand, and the parameters specified by the designer on the other. Since smaller transistors are more difficult to manufacture precisely, the magnitude of process variations has been increasing with technology scaling [10].

At the full chip level, process variations result in a discrepancy between the power and performance of the fabricated integrated circuits (IC) and the power and performance desired by the designer.

In fact, each fabricated chip has a different power and performance profile compared to every other chip, although these should ideally be identical. Process variations can be addressed using conservative design techniques, i.e., designing for the worst case impact of process variations. However, as the magnitude of process variation increases, conservative design techniques become increasingly pessimistic and result in increased power dissipation and reduced performance. An alternative approach is to ignore process variations altogether and optimize the design assuming the nominal process parameters. Unfortunately, this approach can result in significant yield loss, where yield is defined as the percentage of chips that meet the desired power and performance specifications.

Mindful of the increasing impact of process variations, semiconductor IC designers have recently embraced *statistical design techniques* as an alternative to traditional conservative design techniques. This approach models process parameters as random variables with known probability density functions (*pdf*) instead of modeling them as deterministic values. Given the distributions of process parameters, designers can then perform both statistical analysis, i.e., determine the distribution of important metrics such as the chip's clock frequency and power distribution, and statistical optimization, i.e., optimizing for yield instead of optimizing for nominal values.

Another important trend emerging with technology scaling is an increasing emphasis on design at higher levels of abstraction, i.e., viewing an IC at the abstraction of logic modules, processing cores and memory arrays as opposed to the transistor or gate level abstraction. This is commonly referred to as system-level design [59]. The basic building block in a system-level design flow is a module that could represent, for example, intellectual property (IP) purchased from a vendor or developed in-house, a processing core or its sub-components, an on-chip memory, and on-chip communication routers or buses, to name a few.

This design methodology has a number of advantages: (i) it allows designers to view a system as a relatively small number of interacting modules (compared to the number of transistors or gates on the dice) thereby greatly simplifying the design process; (ii) the designer typically

has a more intuitive understanding of how the functionality and interaction between these modules affects the functionality and performance of the entire system; and (iii) design re-use at the system-level is particularly beneficial because re-using a module implies significant savings in design effort.

Statistical design techniques have conventionally been deployed at the transistor-level and gate-level abstraction. For example, gate-level statistical static timing analysis (SSTA) has begun to replace static timing analysis in industrial design flows [1, 14, 15]. However, with the increasing emphasis on system-level design, it has become important to migrate statistical analysis and optimization techniques toward higher levels of design abstraction. The challenges in doing so can be broadly categorized as:

(1) *Macro-modeling.* The challenge is to develop parameterized macro-models that can be used to determine the distribution of clock frequency and power consumption of a module as a function of process variation and micro-architectural parameters. This would enable designers to ascertain, for example, what happens to the distribution of critical path delay (or clock frequency) if the magnitude of process variations increases or decreases or if the micro-architectural parameters of the module are changed.
(2) *Compositional Analysis.* The challenge is to develop techniques that allow designers to predict the distribution of system performance and power that would result from a composition of modules, given that the distribution of clock frequency and power consumption for each module is given. This is particularly useful in the context of emerging multi-core and many-core systems since it would allow designers to determine the performance and power characteristics of the entire system under process variations, given the characteristics of each core.
(3) *Optimization.* The challenge is to determine the best design time and run time decisions that maximize the system yield in the presence of process variations. At the system

level, these decisions can include which modules to use in the system (design time), the partitioning the system into voltage/frequency islands (design time), or variation-aware dynamic power management (run time).

The goal of this survey is to provide the reader with an introduction to recently proposed techniques that address one or more of the challenges highlighted above. Contemporary system-level design practices encompass a wide range of product categories from general purpose computing to application specific ICs (ASIC) and multi-processor systems-on-chip (MPSoC). While the design flow, performance metrics and optimization objectives in each domain can be different, we will attempt to provide the reader with a broad sampling of variation-aware, system-level techniques from all three domains.

We begin with a primer on process variations focusing on the important variability sources, variability models, and some mathematical preliminaries.

1.1 A Primer on Process Variations

1.1.1 Process Variation Sources

Process variations impact, to various degrees, all parameters of the manufacturing process through two primary mechanisms. *Systematic variations* refer to variation sources where the discrepancy between the desired process parameters and their realizations can, in theory, be predicted using physics based models, although in practice these models may be too complex or poorly understood. For example, detailed optical models of the lithographic process can predict precisely how a desired shape will print after lithography, but running full-fidelity optical simulations for an entire mask can be too time consuming [83]. On the other hand, *random variations* arise from sources that are inherently unpredictable and cannot be modeled at design time. Random variation sources include, for example, variations in the number of dopant atoms implanted in the channel of a transistor.

The critical difference between systematic and random variations is that systematic variations are repeatable, while random variations

are not. In other words, if systematic variations were the only variation source, each manufactured chip would be identical although they would all be different from the designer's intent. On the other hand, random variations result in differences from one chip to another. This survey focuses primarily on random and not systematic variations.

As mentioned before, all parameters of the manufacturing process are subject to variations, but we highlight here the impact of process variations on the three most important parameters that have the greatest impact on CMOS transistor power and performance [87].

(1) *Gate Length Variations.* The transistor gate length[1] (L_g) is typically the smallest feature that can be printed in a given technology node and is therefore the most susceptible to process variations. The gate length of a transistor is impacted by both systematic variations from the lithographic process and random variations that include line-edge roughness (LER) [33]. In addition, systematic variations at the wafer scale, for example from lens defocus, can be modeled as random variations on a per-die basis. This will be made clearer shortly.
(2) *Oxide Thickness Variations.* The height of the gate oxide plays a critical role in determining the gate capacitance and therefore the power and performance of a CMOS transistor. Gate oxide thickness is relatively well controlled across a wafer, but exhibits wafer to wafer variations. Again, at the level of a single die, these can be viewed as random variations.
(3) *Threshold Voltage Variations.* The threshold voltage (V_{th}) of a transistor is a function of a number of physical process parameters including the gate length, L_g, the oxide thickness, T_{ox}, and the channel doping, n_{ch}. Gate length and oxide thickness variations therefore also result in threshold voltage variations. In addition, the channel doping, which is essentially related to the number and distribution of dopant atoms

[1] Technically, the process parameter of interest is the *effective* gate length which is typically smaller than the *drawn* gate length because of source and drain overlaps. The term gate length in this survey refers to effective gate length.

in the channel of a transistor, is subject to random variations because of the inherent randomness of the dopant diffusion process.

It is to be noted that the process parameters of a chip also change with time, i.e., in the field, through various aging mechanisms. Although these time-dependent changes in process parameters can also be interpreted as process variations, in this survey we only address variations that arise from the manufacturing process.

1.1.2 Process Variation Models

The first step in enabling statistical analysis and optimization methodologies is to mathematically model process variations as random variables — in other words, each process parameter for every transistor on the chip is written as a random variable instead of as a deterministic value. The impact of process variations on each transistor can then be thought of as the sum of two constituent components:

(1) *Die-to-Die (D2D) Variations.* D2D variations impact each transistor on a die in the same way but affect different dice differently.
(2) *Within-die (WID) Variations.* WID variations impact each transistor on a die differently.

For the sake of clarity, we note that different terminology is used for the two variation components in the analog circuit design community. D2D variations are referred to as simply process variations, and within-die variations are referred to as mismatch. Nonetheless, in this survey, we will consistently use the terms D2D and WID variations.

Figure 1.1 shows a color coded map of process variations, in this case the critical dimension, at discrete points on a semiconductor wafer [32]. The dark black lines represent the boundaries of the dice that the wafer is split up into. It is clear from the picture that no two dice are identical and that there are significant variations within a given die. Figure 1.1 also illustrates how process variations for a single die can be modeled as the sum of D2D and WID variations.

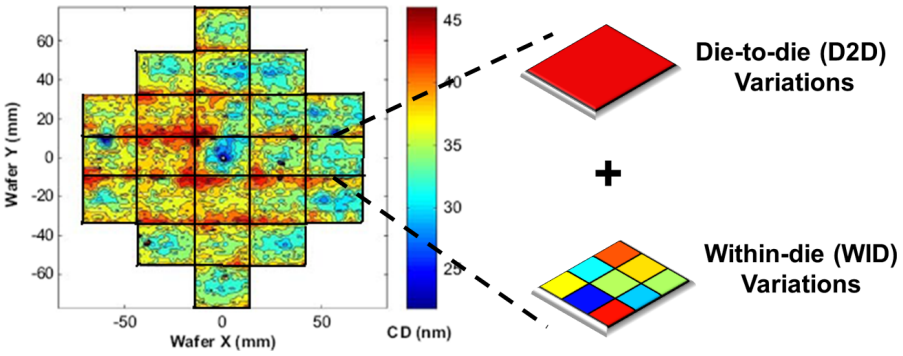

Fig. 1.1 Wafer map of critical dimension variations [32] and decomposition of the variations on a single die into its die-to-die and within-die components.

Mathematically, let L_g^i be a random variable that represents the gate length of the ith transistor on the die. We can write this random variable as:

$$L_g^i = L_{g,\text{nom}}^i + \Delta L_{g,\text{D2D}} + \Delta L_{g,\text{WID}}^i \tag{1.1}$$

In this equation, $L_{g,\text{nom}}^i$ is a deterministic number that represents the nominal value of gate length, $\Delta L_{g,\text{D2D}}$ is a zero mean random variable that represents the impact of D2D variations on gate length, and $\Delta L_{g,\text{WID}}^i$ is a zero mean random variable that represents the impact of WID variations on gate length. Note that while $\Delta L_{g,\text{WID}}^i$ is unique to each transistor, $\Delta L_{g,\text{D2D}}$ is shared across transistors.

The D2D and WID random variables are typically assumed to have a Gaussian distribution. In particular

$$\Delta L_{g,\text{D2D}}^i \sim N(0, \sigma_{\text{D2D}})$$

and

$$\Delta L_{g,\text{WID}}^i \sim N(0, \sigma_{\text{WID}})$$

In these equations σ_{D2D} and σ_{WID} are the standard deviations of process variation for D2D and WID variations, respectively, and are typically specified by the foundry in its process models.

As a final note, the WID process variations might be spatially correlated. In other words, the random variables $\Delta L_{g,\text{WID}}^i$ and $\Delta L_{g,\text{WID}}^j$

for two transistors i and j on the same die can be correlated with each other. The stregth of the correlation is quantified by the correlation co-efficient, ρ_{ij}, which is defined as:

$$\rho_{ij} = \frac{\sqrt{E(\Delta L^i_{g,\text{WID}} \times \Delta L^j_{g,\text{WID}})}}{\sigma_{\text{WID}}}$$

In this equation, $E(.)$ refers to the expectation of a random variable.

The value of ρ_{ij} is dictated by the distance $d(i,j)$ between the two transistors — transistors that are closer together tend to be more correlated than transistors that are further apart. A number of spatial correlation models have been proposed in literature based on fitting empirically observed data [32, 91, 108]. Xiong et al. [108] have shown that the following model is mathematically consistent:

$$\rho_{ij} = e^{-b \times d(i,j)},$$

where b is an empirically determined scalar constant. This model posits that correlation between the WID process parameters of any two transistors on a die decreases exponentially with increasing distance between transistors. The spatial correlation model can be used to generate representative die maps of process variation, similar to the empirically measured map shown in Figure 1.1, which can then be used as an input to statistical analysis and optimization procedures. As an example, Figure 1.2 shows process variation maps for dice with varying strength of spatial correlations, i.e., with $b = 100$ (low correlations), $b = 10$ (medium correlations), and $b = 1$ (high correlations).

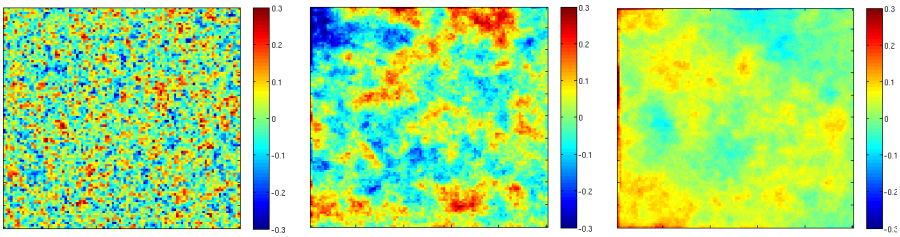

Fig. 1.2 Process variation maps with different values of parameter b.

1.1.3 Process Variation Impact

Having explained how process variations are modeled at the transistor level, we are now in a position to understand the impact of process variations at the gate and circuit levels. At the gate level, process variations impact both the gate delay and the leakage power dissipation of the gate. For the rest of this discussion, we will assume that all transistors in a gate are equally affected by process variations, and that the impact is captured by a single processor parameter p^i for the ith gate. The process parameter p^i can represent either gate length, threshold voltage, or oxide thickness.

The delay of gate i, d^i, is a nonlinear function of process parameters but can be linearized with respect to the process parameter as follows:

$$d^i \approx d^i_{\text{nom}} + \left.\frac{\partial d^i}{\partial p^i}\right|_{d^i_{\text{nom}}} p^i$$

This linear relationship implies that if, for example, the 3σ value of the process parameter is 30% of its nominal value, the 3σ of gate delay will also be 30% of its nominal value. The linear approximation is reasonably accurate because, unless the supply voltage is set to near or sub-threshold voltage values, the gate delay is only a weakly nonlinear function of process parameters.

The leakage power dissipation of a gate is, unlike gate delay, a *strongly* nonlinear function of process parameters. In fact, the leakage power dissipation can be modeled as an exponential function of the gate length, threshold voltage and oxide thickness. Therefore, a linear approximation is not appropriate for leakage power dissipation. The leakage power of gate i can be expressed as

$$I^i_{\text{Leak}} \propto e^{\alpha^i p^i},$$

where α^i is a constant fitting parameter. Since random variable p^i is normally distributed, the distribution of I^i_{Leak} is log-normal [66].

From a full chip perspective, variations in the delay of each gate result in variations in the *maximum critical path delay* of the circuit, i.e., the worst case delay from any flip–flop output to any flip–flop input, including set-up and hold times. The maximum critical path

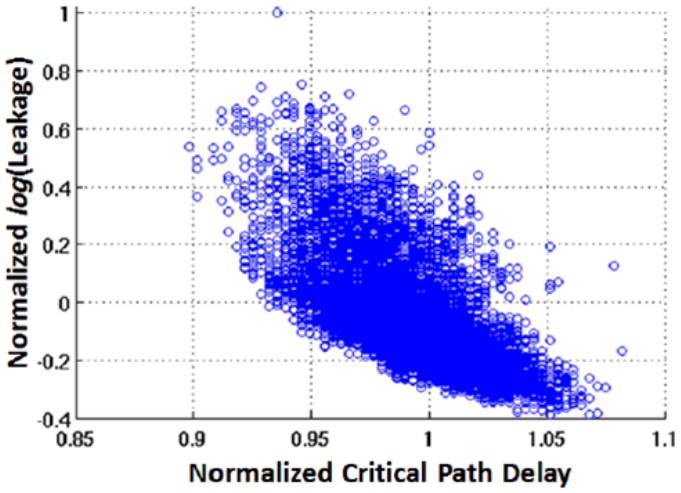

Fig. 1.3 Scatter plot of leakage power dissipation versus maximum critical path delay.

delay determines the maximum operating frequency of the circuit. This results in variations in the clock frequency of a circuit under the impact of process variations. At the same time, the peak power dissipation of the full chip can be computed by summing up the power dissipation of each gate. Since the leakage component of gate level power dissipation varies with process variations, so does the full chip power dissipation. We note that the impact of process variations on the full-chip dynamic power dissipation is, comparatively, negligible [96].

Figure 1.3 shows a scatter plot for the measured leakage power dissipation and maximum critical path delay for an industrial test chip manufactured in a relatively mature 0.13 μm process.[2] As expected, the variations in leakage power dissipation have a much wider spread than the variations in critical path delay because of the exponential dependence of leakage power on process parameters.

1.2 System Level Design Space

System level design techniques encompass a wide range of computing platforms across application domains — from general purpose chip

[2] Due to contractual reasons, more details about the test chip are not provided here.

multi-processors (CMP) to multi-processor systems-on-chip (MPSoC) and application specific integrated circuits (ASIC). To help understand how process variation aware statistical analysis and design techniques fit within a system level design framework, we now briefly discuss each of the three application domains.

1.2.1 General Purpose Chip Multi-Processors

General purpose CMPs represent the work horses of modern day computing infrastructure. A CMP consists of multiple, typically identical, processors that support a specific instruction set architecture (ISA). Processors consist of a pipeline of components that fetch, decode, execute, and commit instructions in program order. Instructions and data are fetched from an off-chip main memory and cached in on-chip static random access memory (SRAM) arrays called caches. The processing cores communicate using an on-chip interconnect, for example, a bus or a network-on-chip (NoC).

System-level design of a CMP typically encompasses both design-time and run-time optimizations. These include:

(1) *Design-time decisions* consist of determining the number of cores, the cache capacity and the micro-architectural parameters of each core. These include issue width, pipeline depth, dispatch policy (in-order versus out-of-order), issue queue size, branch predictor size, etc. The most important optimization metrics are the execution latency of benchmark programs, peak power consumption and peak temperature. These metrics are typically computed via cycle-accurate simulations.
(2) *Run-time policies* try to maximize performance within a power budget or meet performance requirement while minimizing power and energy consumption. The most effective knobs to control power and performance are, in fact, system level knobs like the voltage and frequency at which cores operate and additional power states like sleep and idle modes. For CMPs, these policies are typically best-effort, i.e., for

complexity reasons, they do not provide any guarantees other than attempting to reach the specified power/performance goals.

1.2.2 Multi-Processor Systems-on-Chip

MPSoCs are typically more specialized than CMPs and are targeted toward a specific application or set of applications that are known in advance. Important application domains include entertainment, for example, set-top boxes and automotive electronics. Since the applications are known and characterized at design time, the system hardware can be tailored and optimized for this set of applications. MPSoCs typically feature both greater parallelism and greater heterogeneity in terms of available on-chip hardware resources than CMPs. The steps in the system-level design of an MPSoC include:

(1) Design-time decisions include platform synthesis, i.e., determining the type and number or cores, memories and communication resources on the chip, application mapping and scheduling on the available hardware resources and, in some cases, direct synthesis of critical applications to hardware.
(2) Run-time knobs are similar to those available for CMPs, but the policies are geared toward providing run-time *guarantees* as opposed to best effort. This is particularly true for real-time applications with hard or soft deadline constraints.

1.2.3 Application Specific Integrated Circuits

ASICs refer to dedicated hardware logic that is synthesized for a specific task or applications. Compared to CMPs and MPSoCs, ASICs are the most power efficient but also the least re-configurable. ASICs have conventionally designed with the aid of low-level hardware description languages such as Verilog or VHDL. However, to address increasing design complexity, there has been a move toward direct synthesis from higher levels language, for example C/C++. This is referred to as high-level synthesis (HLS). HLS begins with extracting a task graph from

the high-level code and then mapping each task in the graph to an available logic module from a library of modules, for example, comparators, adders, multipliers, and so on. Optimization metrics include latency, throughput, power dissipation and energy consumption and optimization is primarily focused on design time decisions. The HLS flow can be used to either synthesize stand-alone ASICs or to synthesize custom logic for use as a hardware accelerator in an MPSoC or even a CMP.

1.3 Survey Overview

In this survey, we have chosen highlight how the impact of process variations can be modeled and mitigated at the system level by making use of specific and detailed examples in each section, followed by a survey of related techniques and the existing state-of-the-art. Given the focus on emerging techniques in design automation, we address both planar ICs *and* emerging 3D IC technology for which there is an increasing body of work on system level variability modeling and mitigation. We begin by discussing process variation macro-models for single logic blocks or modules in Section 2. Section 3 discusses compositional variability analysis for multi-core systems, i.e., when the modules characterized in Section 2 are composed into full systems. Section 4 discusses system level variability mitigation techniques that are deployed post-fabrication or at run-time. We conclude with pointers to future research directions in Section 5.

References

[1] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Proceedings of the 2003 IEEE/ACM International Conference on Computer-aided Design*, p. 900, 2003.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *Communications Magazine (IEEE)*, vol. 40, no. 8, pp. 102–114, 2002.

[3] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 602–633, 2001.

[4] L. A. D. Bathen, N. D. Dutt, A. Nicolau, and P. Gupta, "Vamv: variability-aware memory virtualization," in *Proceedings of the Design, Automation and Test in Europe Conference & Exhibition (DATE), 2012*, pp. 284–287, 2012.

[5] L. A. D. Bathen, M. Gottscho, N. Dutt, A. Nicolau, and P. Gupta, "ViP-ZonE: OS-level memory variability-driven physical address zoning for energy savings," in *Proceedings of the IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pp. 33–42, 2012.

[6] K. A. Bowman, A. R. Alameldeen, S. T. Srinivasan, and C. B. Wilkerson, "Impact of die-to-die and within-die parameter variations on the throughput distribution of multi-core processors," in *Proceedings of the 2007 International Symposium on Low Power Electronics and Design*, pp. 50–55, 2007.

[7] K. A. Bowman, S. G. Duvall, and J. D. 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 Circuits*, vol. 37, no. 2, pp. 183–190, 2002.

[8] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2004.

[9] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations," in *ACM SIGARCH Computer Architecture News,* vol. 28, pp. 83–94, 2000.

[10] B. H. Calhoun, Y. Cao, X. Li, K. Mai, L. T. Pileggi, R. A. Rutenbar, and K. L. Shepard, "Digital circuit design challenges and opportunities in the era of nanoscale CMOS," *Proceedings of the IEEE*, vol. 96, no. 2, pp. 343–365, 2008.

[11] K. Chae and S. Mukhopadhyay, "Tier-adaptive-voltage-scaling (TAVS): A methodology for post-silicon tuning of 3D ICs," in *Proceedings of the Asia and South Pacific Design Automation Conference*, pp. 277–282, 2012.

[12] K. Chakraborty and S. Roy, "A Novel threshold voltage assignment for 3D multicore designs," *Journal of Low Power Electronics*, vol. 6, no. 3, pp. 436–446, 2010.

[13] S. Chandra, K. Lahiri, A. Raghunathan, and S. Dey, "Considering process variations during system-level power analysis," in *Proceedings of the 2006 International Symposium on Low Power Electronics and Design*, pp. 342–345, New York, NY, USA, 2006.

[14] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proceedings of the 2003 IEEE/ACM International Conference on Computer-aided Design*, p. 621, 2003.

[15] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, "Parameterized block-based statistical timing analysis with non-gaussian parameters, nonlinear delay functions," in *Proceedings of the Annual Conference on Design Automation*, pp. 71–76, New York, NY, USA, 2005.

[16] D. M. Chapiro, "Globally-asynchronous locally-synchronous systems," PhD Thesis, vol. 1, p. 50, 1984.

[17] T. Chelcea and S. M. Nowick, "Robust interfaces for mixed-timing systems with application to latency-insensitive protocols," in *Proceedings of the Conference on Design Automation*, pp. 21–26, New York, NY, USA, 2001.

[18] Y. Chen, Y. Xie, Y. Wang, and A. Takach, "Parametric yield driven resource binding in behavioral synthesis with multi-Vth/Vdd library," in *Proceedings of the 2010 Asia and South Pacific Design Automation Conference*, pp. 781–786, 2010.

[19] Y. Chen and Q. Zhao, "On the lifetime of wireless sensor networks," *Communications Letters (IEEE)*, vol. 9, no. 11, pp. 976–978, 2005.

[20] C. Y. Cher and E. Kursun, "Exploring the effects of on-chip thermal variation on high-performance multicore architectures," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 8, no. 1, p. 2, 2011.

[21] E. Chun, Z. Chishti, and T. N. Vijaykumar, "Shapeshifter: Dynamically changing pipeline width and speed to address process variations," in *MICRO '08: Proceedings of the 2008 IEEE/ACM International Symposium on Microarchitecture*, 2008.

[22] A. Das, S. Ozdemir, G. Memik, and A. Choudhary, "Evaluating voltage islands in CMPs under process variations," in *Proceedings of the International Conference on IEEE*, pp. 129–136, 2007.

[23] A. Dasdan and R. K. Gupta, "Faster maximum and minimum mean cycle algorithms for system-performance analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 10, pp. 889–899, 1998.

[24] A. Dasdan and R. K. Gupta, "Faster maximum and minimum mean cycle algorithms for system-performance analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 1998.

[25] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *Design & Test of Computers (IEEE)*, vol. 22, no. 6, pp. 498–510, 2005.

[26] R. P. Dick, D. L. Rhodes, and W. Wolf, "TGFF: Task graphs for free," in *Proceedings of the 6th International Workshop on Hardware/Software Codesign*, pp. 97–101, IEEE Computer Society, 1998.

[27] S. Dighe, S. Vangal, P. Aseron, S. Kumar, T. Jacob, K. Bowman, J. Howard, J. Tschanz, V. Erraguntla, and *et al.*. N. Borkar, "Within-die variation-aware dynamic-voltage-frequency scaling core mapping and thread hopping for an 80-core processor," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 174–175, 2010.

[28] J. Donald and M. Martonosi, "Power efficiency for variation-tolerant multicore processors," in *Proceedings of the 2006 International Symposium on Low Power Electronics and Design*, 2006.

[29] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, and K. Flautner, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture*, 7–18.

[30] C. Ferri, S. Reda, and R. I. Bahar, "Strategies for improving the parametric yield and profits of 3D ICs," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 220–226, 2007.

[31] T. Fischer, J. Desai, B. Doyle, S. Naffziger, and B. Patella, "A 90-nm variable frequency clock system for a power-managed Itanium architecture processor," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 218–228, 2006.

[32] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization," in *Proceedings of the Sixth International Symposium on Quality of Electronic Design*, pp. 516–521, IEEE, 2005.

[33] G. M. Gallatin, "Resist blur and line edge roughness (Invited Paper)," in *Proceedings of the SPIE,* vol. 5754, p. 38, 2005.

[34] S. Garg and D. Marculescu, "On the impact of manufacturing process variations on the lifetime of sensor networks," in *Proceedings of the IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ ISSS)*, pp. 203–208, 2007.

[35] S. Garg and D. Marculescu, "System-level process variation driven throughput analysis for single and multiple voltage-frequency island designs," in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pp. 1–6, 2007.

[36] S. Garg and D. Marculescu, "System-level throughput analysis for process variation aware multiple voltage-frequency island designs," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 13, no. 4, p. 59, 2008.

[37] S. Garg and D. Marculescu, "3D-GCP: An analytical model for the impact of process variations on the critical path delay distribution of 3D ICs," in *Proceedings of the International Symposium on Quality of Electronic Design*, pp. 147–155, IEEE, 2009.

[38] S. Garg and D. Marculescu, "System-level process variability analysis and mitigation for 3D MPSoCs," in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pp. 604–609, 2009.

[39] A. H. Ghamarian, M. Geilen, T. Basten, and S. Stuijk, "Parametric throughput analysis of synchronous data flow graphs," in *Proceedings of the Design, Automation and Test in Europe, 2008. DATE'08*, pp. 116–121, IEEE, 2008.

[40] R. Ginosar, "Metastability and synchronizers: A tutorial," *Design & Test of Computers, IEEE*, vol. 28, no. 5, pp. 23–35, 2011.

[41] S. Golshan, L. Singhal, and E. Bozorgzadeh, "Process variation aware system-level load assignment for total energy minimization using stochastic ordering," in *Proceedings of the 12th International Symposium on Quality Electronic Design*, pp. 1–6, IEEE, 2011.

[42] J. Guo, S. Datta, M. Lundstrom, M. Brink, P. McEuen, A. Javey, H. Dai, H. Kim, and P. McIntyre, "Assessment of silicon MOS and carbon nanotube FET performance limits using a general theory of ballistic transistors," in *Proceedings of the International Electron Devices Meeting*, pp. 711–714, IEEE, 2002.

[43] M. S. Gupta, J. A. Rivers, P. Bose, G. Y. Wei, and D. Brooks, "Tribeca: design for PVT variations with local recovery and fine-grained adaptation," in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 435–446, IEEE, 2009.

[44] A. Hemani, T. Meincke, S. Kumar, A. Postula, T. Olsson, P. Nilsson, J. Oberg, P. Ellervee, and D. Lundqvist, "Lowering power consumption in clock by using globally asynchronous locally synchronous design style," in *Proceedings of the Design Automation Conference*, pp. 873–878, 1999.

[45] S. Herbert, S. Garg, and D. Marculescu, "Reclaiming Performance and Energy Efficiency from Variability," *Proceedings of IBM Pac2*, 2006.

[46] S. Herbert, S. Garg, and D. Marculescu, "Exploiting process variability in voltage/frequency control," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 8, pp. 1392–1404, 2012.

[47] S. Herbert and D. Marculescu, "Analysis of dynamic voltage/frequency scaling in chip-multiprocessors," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 38–43, IEEE, 2007.

[48] S. Herbert and D. Marculescu, "Characterizing chip-multiprocessor variability-tolerance," in *Proceedings of the Annual Conference on Design Automation*, pp. 313–318, 2008.

[49] S. Herbert and D. Marculescu, "Variation-aware dynamic voltage/frequency scaling," in *Proceedings of the International Symposium on High Performance Computer Architecture*, pp. 301–312, 2009.

[50] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 5, pp. 501–513, 2006.

[51] E. Humenay, D. Tarjan, and K. Skadron, "Impact of process variations on multicore performance symmetry," in *Proceedings of the conference on Design, automation and test in Europe*, pp. 1653–1658, 2007.

[52] C. Isci, A. Buyuktosunoglu, C. Y. Cher, P. Bose, and M. Martonosi, "An analysis of efficient multi-core global power management policies: Maximizing performance for a given power budget," in *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 347–358, IEEE, 2006.

[53] A. Iyer and D. Marculescu, "Power and performance evaluation of globally asynchronous locally synchronous processors," in *Proceedings of the 29th Annual International Symposium on Computer Architecture*, pp. 158–168, IEEE, 2002.

[54] A. Iyer and D. Marculescu, "Power efficiency of voltage scaling in multiple clock, multiple voltage cores," in *Proceedings of the 2002 IEEE/ACM International Conference on Computer-aided Design*, pp. 379–386, 2002.

[55] D. C. Juan, S. Garg, and D. Marculescu, "Statistical thermal evaluation and mitigation techniques for 3D chip-multiprocessors in the presence of process variations," in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, pp. 1–6, 2011.

[56] J. Jung and T. Kim, "Timing variation-aware high-level synthesis considering accurate yield computation," in *Proceedings of the IEEE International Conference on Computer Design*, pp. 207–212, 2009.

[57] R. M. Karp, "A characterization of the minimum cycle mean in a digraph," *Discrete Math*, vol. 23, 1978.

[58] U. R. Karpuzcu, K. B. Kolluru, N. S. Kim, and J. Torrellas, "VARIUS-NTV: A microarchitectural model to capture the increased sensitivity of manycores to process variations at near-threshold voltages," in *Proceedings of the 42nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pp. 1–11, 2012.

[59] K. Keutzer, A. R. Newton, J. M. Rabaey, and A. Sangiovanni-Vincentelli, "System-level design: Orthogonalization of concerns and platform-based design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 12, pp. 1523–1543, 2000.

[60] N. S. Kim, T. Kgil, K. Bowman, V. De, and T. Mudge, "Total power-optimal pipelining and parallel processing under process variations in nanometer technology," in *Proceedings of the IEEE/ACM International Conference on Computer-aided Design*, pp. 535–540, 2005.

[61] V. J. Kozhikkottu, R. Venkatesan, A. Raghunathan, and S. Dey, "VESPA: Variability emulation for System-on-Chip performance analysis," in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011*, pp. 1–6, 2011.

[62] C. Y. Lee and N. K. Jha, "CACTI-FinFET: an integrated delay and power modeling framework for FinFET-based caches under process variations," in *Proceedings of the ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 866–871, 2011.

[63] J. Lee, P. P. Ajgaonkar, and N. S. Kim, "Analyzing throughput of GPGPUs exploiting within-die core-to-core frequency variation," in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 237–246, IEEE, 2011.

[64] J. Lee and N. S. Kim, "Optimizing total power of many-core processors considering voltage scaling limit and process variations," in *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*, pp. 201–206, ACM, 2009.

[65] B. Li, L. S. Peh, and P. Patra, "Impact of process and temperature variations on network-on-chip design exploration," in *Proceedings of the Second ACM/IEEE International Symposium on Networks-on-Chip*, pp. 117–126, IEEE Computer Society Washington, DC, USA, 2008.

[66] X. Li, J. Le, and L. T. Pileggi, "Projection-based statistical analysis of full-chip leakage power with non-log-normal distributions," in *Proceedings of the Annual Design Automation Conference*, pp. 103–108, 2006.

[67] X. Li, P. Li, and L. T. Pileggi, "Parameterized interconnect order reduction with explicit-and-implicit multi-parameter moment matching for inter/intra-die variations," in *Proceedings of the IEEE/ACM International Conference on Computer-aided Design*, pp. 806–812, 2005.

[68] X. Liang and D. Brooks, "Microarchitecture parameter selection to optimize system performance under process variation," in *Proceedings of the 2006 IEEE/ACM International Conference on Computer-aided Design*, pp. 429–436, New York, NY, USA, 2006.

[69] X. Liang, G.-Y. Wei, and D. Brooks, "Revival: A Variation-Tolerant Architecture Using Voltage Interpolation and Variable Latency," *IEEE Micro*, vol. 29, no. 1, 2009.

[70] W. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 24, no. 7, pp. 1042–1053, 2005.

[71] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *Journal of the ACM (JACM)*, vol. 20, no. 1, pp. 46–61, 1973.

[72] S. Liu, Q. Qiu, and Q. Wu, "Full-chip leakage current estimation based on statistical sampling techniques," in *Proceedings of the Great Lakes Symposium on VLSI*, 2008.

[73] G. Lucas, S. Cromar, and D. Chen, "FastYield: Variation-aware, layout-driven simultaneous binding and module selection for performance yield optimization," in *Proceedings of the 2009 Asia and South Pacific Design Automation Conference*, pp. 61–66, 2009.

[74] G. Magklis, P. Chaparro, J. González, and A. González, "Independent front-end and back-end dynamic voltage scaling for a GALS microarchitecture," in *Proceedings of the 2006 International Symposium on Low Power Electronics and Design*, pp. 49–54, ACM, 2006.

[75] D. Marculescu, "On the use of microarchitecture-driven dynamic voltage scaling," in *Workshop on Complexity-Effective Design*, 2000.

[76] D. Marculescu and S. Garg, "System-level process-driven variability analysis for single and multiple voltage-frequency island systems," in *Proceedings of the 2006 IEEE/ACM International Conference on Computer-aided Design*, pp. 541–546, New York, NY, USA, 2006.

[77] D. Marculescu and S. Garg, "Process-Driven Variability Analysis of Single and Multiple Voltage–Frequency Island Latency-Constrained Systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 5, pp. 893–905, 2008.

[78] D. Marculescu and E. Talpes, "Energy awareness and uncertainty in microarchitecture-level design," *Micro, IEEE*, vol. 25, no. 5, pp. 64–76, 2005.

[79] D. Marculescu and E. Talpes, "Variability and energy awareness: a microarchitecture-level perspective," in *Proceedings of the Annual Conference on Design Automation*, pp. 11–16, New York, NY, USA, 2005.

[80] T. N. Miller, X. Pan, R. Thomas, N. Sedaghati, and R. Teodorescu, "Booster: reactive core acceleration for mitigating the effects of process variation and application imbalance in low-voltage chips," in *Proceedings of the 2012 IEEE 18th International Symposium on High Performance Computer Architecture (HPCA)*, pp. 1–12, IEEE, 2012.

[81] M. Miranda, B. Dierickx, P. Zuber, P. Dobrovoln, F. Kutscherauer, P. Roussel, and P. Poliakov, "Variability aware modeling of SoCs: From device variations to manufactured system yield," in *Proceedings of the International Symposium on Quality of Electronic Design*, pp. 547–553, IEEE, 2009.

[82] J. Muttersbach, T. Villiger, H. Kaeslin, N. Felber, and W. Fichtner, "Globally-asynchronous locally-synchronous architectures to simplify the design of on-chip systems," in *Proceedings of the Annual IEEE International ASIC/SOC Conference*, pp. 317–321, 1999.

[83] S. R. Nassif, "Modeling and forecasting of manufacturing variations," in *Proceedings of the International Workshop on Statistical Metrology*, pp. 2–10, 2000.

[84] S. Ozdemir, Y. Pan, A. Das, G. Memik, G. Loh, and A. Choudhary, "Quantifying and coping with parametric variations in 3D-stacked microarchitectures," in *Proceedings of the Design Automation Conference*, pp. 144–149, 2010.

[85] S. Pasricha, Y. H. Park, N. Dutt, and F. J. Kurdahi, "System-level PVT variation-aware power exploration of on-chip communication architectures," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 14, no. 2, p. 20, 2009.

[86] A. Rahimi, L. Benini, and R. Gupta, "Procedure hopping: a low overhead solution to mitigate variability in shared-L1 processor clusters," in *Proceedings of the 2012 ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 415–420, ACM, 2012.

[87] R. Rao, A. Srivastava, D. Blaauw, and D. Sylvester, "Statistical estimation of leakage current considering inter-and intra-die process variation," in *Proceedings of the 2003 International Symposium on Low Power Electronics and Design*, pp. 84–89, ACM, 2003.

[88] S. Reda, A. Si, and R. Bahar, "Reducing the leakage and timing variability of 2D ICs using 3D ICs," in *Proceedings of the 14th ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 283–286, ACM, 2009.

[89] J. Samandari-Rad, M. Guthaus, and R. Hughey, "VAR-TX: A variability-aware SRAM model for predicting the optimum architecture to achieve minimum access-time for yield enhancement in nano-scaled CMOS," in *Proceedings of the 13th International Symposium on Quality Electronic Design*, pp. 506–515, IEEE, 2012.

[90] S. Sarangi, B. Greskamp, A. Tiwari, and J. Torrellas, "EVAL: Utilizing processors with variation-induced timing errors," in *Proceedings of the 41st IEEE/ACM International Symposium on Microarchitecture*, 2008.

[91] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "VARIUS: A model of process variation and resulting timing errors for microarchitects," *IEEE Transactions on Semiconductor Manufacturing*, vol. 21, no. 1, pp. 3–13, 2008.

[92] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott, "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling," in *Proceedings of the Eighth International Symposium on High-Performance Computer Architecture*, pp. 29–40, IEEE, 2002.

[93] S. Seo, R. G. Dreslinski, M. Woh, Y. Park, C. Charkrabari, S. Mahlke, D. Blaauw, and T. Mudge, "Process variation in near-threshold wide SIMD architectures," in *Proceedings of the Annual Design Automation Conference*, pp. 980–987, 2012.

[94] M. Seok, D. Sylvester, and D. Blaauw, "Optimal technology selection for minimizing energy and variability in low voltage applications," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 9–14, IEEE, 2008.

[95] L. Singhal and E. Bozorgzadeh, "Process variation aware system-level task allocation using stochastic ordering of delay distributions," in *Proceedings of the IEEE/ACM International Conference on Computer-aided Design*, pp. 570–574, 2008.

[96] D. Sylvester, K. Agarwal, and S. Shah, "Variability in nanometer CMOS: Impact, analysis, and minimization," *Integration, the VLSI Journal*, vol. 41, no. 3, pp. 319–339, 2008.

[97] E. Talpes and D. Marculescu, "Toward a multiple clock/voltage island design style for power-aware processors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, no. 5, pp. 591–603, 2005.

[98] R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas, "Mitigating Parameter Variation with Dynamic Fine-Grain Body Biasing," in *MICRO 40: Proceedings of the 40th annual ACM/IEEE International Symposium on Microarchitecture*, 2007.

[99] R. Teodorescu and J. Torrellas, "Variation-aware application scheduling and power management for chip multiprocessors," in *Proceedings of the 35th annual International Symposium on Computer Architecture*, 2008.

[100] A. Tiwari, S. R. Sarangi, and J. Torrellas, "ReCycle: pipeline adaptation to tolerate process variation," in *ACM SIGARCH Computer Architecture News*, vol. 35, pp. 323–334, ACM, 2007.

[101] J. W. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, and V. De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-die parameter variations on Microprocessor Frequency and Leakage," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 11, 2002.

[102] N. Verma, J. Kwong, and A. P. Chandrakasan, "Nanometer MOSFET variation in minimum energy subthreshold circuits," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 163–174, 2008.

[103] F. Wang, Y. Chen, C. Nicopoulos, X. Wu, Y. Xie, and N. Vijaykrishnan, "Variation-aware Task and Communication Mapping for MPSoC Architecture," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 30, no. 2, pp. 295–307, 2011.

[104] F. Wang, G. Sun, and Y. Xie, "A variation aware high level synthesis framework," in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 1063–1068, New York, NY, USA, 2008.

[105] F. Wang, X. Wu, and Y. Xie, "Variability-driven module selection with joint design time optimization and post-silicon tuning," in *Proceedings of the 2008 Conference on Asia and South Pacific Design Automation*, pp. 2–9, Los Alamitos, CA, USA, 2008.

[106] L. Wanner, R. Balani, S. Zahedi, C. Apte, P. Gupta, and M. Srivastava, "Variability-aware duty cycle scheduling in long running embedded sensing systems," in *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6, 2011.

[107] J. Wawrzynek, D. Patterson, M. Oskin, S. L. Lu, C. Kozyrakis, J. C. Hoe, D. Chiou, and K. Asanovic, "RAMP: Research accelerator for multiple processors," *Micro, IEEE*, vol. 27, no. 2, pp. 46–57, 2007.

[108] J. Xiong, V. Zolotov, and L. He, "Robust extraction of spatial correlation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 4, pp. 619–631, 2007.

[109] H. Xu, V. F. Pavlidis, and G. D. Micheli, "Process-induced skew variation for scaled 2-D and 3-D ICs," in *Proceedings of the 12th ACM/IEEE International Workshop on System Level Interconnect Prediction*, pp. 17–24, ACM, 2010.

[110] J. Zhang, N. Patil, A. Hazeghi, and S. Mitra, "Carbon nanotube circuits in the presence of carbon nanotube density variations," in *Proceedings of the 46th ACM/IEEE Design Automation Conference*, pp. 71–76, 2009.