Timing Analysis Techniques Review for sub-30 nm Circuit Designs

Juho Kim*, Sangwoo Han*, and Roy Jewell**

Abstract—With scaled technology, timing analysis of circuits becomes more and more difficult. In this paper, we review recently developed circuit simulation techniques created to deal with the cost issues of transistor-level simulations. Various techniques for fast SPICE simulations and Monte Carlo simulations are introduced. Moreover, process and aging variation issues are mentioned, along with promising methodologies.

Index Terms—Timing analysis, simulation, process variation, aging variation

I. INTRODUCTION

With continuous CMOS technology scaling, the number of devices per unit area have increased dramatically. This increased degree of integration enables us to design more powerful electronic devices. However circuit designs of less than 30 nm causes many problems that now have significant impacts on circuit performance.

One of these key design problems is process variation. Increasing fluctuations as a result of manufacturing processes have introduced unavoidable and significant uncertainty in circuit performance; hence, ensuring manufacturability has been identified as one of the top priorities of today's IC design process. Conventional methodologies for process variations, such as cornerbased approach are not accurate enough to estimate the impact of variations with process variation dependencies. In addition, they tend to overestimate the impact of process variations on circuit performance, which may lead to overly pessimistic designs. To improve the manufacturability and yield of designs, various analyses and simulation techniques are used at each design level. However, there are still many problems that need to be solved [1]. For example, transistor-level simulations using a SPICE-like engine can be used to analyze circuit blocks such as standard library cells, analog memory blocks, and interconnect wires. Even though a circuit block is small, the corresponding simulation-based analysis involves large computation costs for the following reasons:

- · Extremely complex device models
- · Characterization costs for process variations
- · Sophisticated interconnect models

Another problem is reliability. As reliability issues become more and more important for the semiconductor industry, modeling is increasingly requested to provide design tools to achieve better device performance and more robust reliability margins [2]. Previous reliabilityrelated works mostly focused on the physical mechanism of each reliability issue, such as bias temperature instability (BTI) and hot carrier injection (HCI) for modeling and simulation. However, aggressive CMOS technology scaling also leads to statistical variations of circuit degradation [3]. Therefore, SPICE and behavioral- level modeling of degradation with the presence of process variations is required to check an aging variability's impact on analog/digital circuit and memory.

In Section II, we review recently-developed circuit simulation techniques that address timing analysis and

Manuscript received Dec. 1, 2010; revised Dec. 13, 2010

^{*} Dep. CSE., Sogang University, Seoul, 121-742, Korea

^{**} Magma Design Automation, 1650 Technology Dr., San Jose, CA 95110, USA

E-mail : jhkim@sogang.ac.kr

verification problems in technologies less than 30 nm in size. Several process variation-aware design techniques and statistical reliability modeling approaches will be discussed in Sections III and IV.

II. CIRCUIT SIMULATION TECHNIQUES

Growing VLSI circuit size and increasing structure complexity make transistor-level simulations infeasible. In most transient analyses using SPICE tools, simulations of some moderate scale circuit designs takes days to accomplish, and low simulation efficiency becomes a critical bottleneck for modern CAD tools. As a result, there is significant motivation to speed up SPICE simulations without losing accuracy.

Recently developed fast SPICE methodologies are based on various techniques such as following:

- · Look-up table models
- · Event-driven or multi-time step algorithms
- · Hierarchical simulations
- · Parallel computations

The authors of [4] proposed speeding up circuit simulations using a graphics processing unit (GPU), namely the NVIDIA GeForce 8800 GTS GPU. They concentrated on the acceleration of SPICE by performing transistor model evaluations on the GPU. Since a large fraction of SPICE runtime is spent evaluating transistor model equations, a significant speedup could be achieved. When their accelerator was integrated in a commercial SPICE tool, they showed significant speedups (2.36X on average) through their experimental results.

One of the recent approaches in parallel SPICE simulation, proposed by X. Zhou et al., is the use of a circuit partition algorithm based on direct current connected blocks (DCCB) and strong connected components (SCC) [5]. They partitioned MOSFETs and related passive device networks in the circuit with a direct current passage to form DCCB and SCC. Since DCCBs are commonly seen as basic functional blocks in VLSI circuits, their method is robust for most application cases. Their method achieved similar or better results than a traditional k-way Fiduccia-Mattheyses partition

algorithm with a $20 \sim 50\%$ smaller CPU time.

Process variations increasingly impact the electrical behavior of a design. This is often tackled by performing Monte Carlo (MC) SPICE simulations, requiring significant computing and time resources. MC simulations attempt to estimate the probability distribution of the performance of circuits via three steps [1]:

- Generating a set of random samples for process parameters
- Running transistor-level simulations and evaluating performance values at all sampling points
- · Estimating the performance distribution

For transistor-level MC analyses, the simulation step is often the most time-consuming, since a huge number of sampling points are required to achieve sufficient accuracy. Therefore, efficient sampling methods are also essential for statistical transistor-level analyses of circuits.

Latin hypercube sampling (LHS) is one of these fast MC sampling techniques [6]. The key idea of LHS is to create a sampling point distribution close to the probability distribution function (PDF) of the random variable that researchers are trying to sample. In the onedimensional case where researchers attempt to generate N sampling points for a random variable, LHS consists of two steps. First, the cumulative distribution function (CDF) is evenly partitioned into N regions. Second, a single sampling point is randomly selected in each region. Therefore, LHS eliminates the possibility that the majority of samples come from the same small local region. Since LHS distributes sampling points all over a random space, it is more efficient than direct random sampling.

An alternative strategy for better controlling of MC samples is to use quasi monte carlo (QMC) technique which is deterministically chosen to more uniformly sample the statistical distribution [7]. This technique is widely used in many application domains, such as finance, and is, also applicable to the domain of scaled semiconductor problems; speedups of 10-50X have been demonstrated in [8], compared to a direct MC simulation.

III. VARIABILITY-AWARE DESIGN TECHNIQUES

The complexities in manufacturing transistors in below 30 nm technologies have caused significant variations in transistor parameters such as thickness of gate oxide (T_{ox}) and threshold voltage (V_{th}). These manufacturing uncertainties result in significant performance and power variations in circuit designs and variability of performance and power have increased with continuous technology scaling. In this section, we review several process variation-aware design techniques from variability-aware high level synthesis (HLS) to post-silicon tuning methods.

With increases in design complexity, the importance of system-level design methodologies increases. These methodologies will likely be enabled by HLS, which is the process of translating a behavioral description into a register transfer-level (RTL) description. However, process variation-aware HLS research is still in its immature stage compared to the existing research focused on process variations that occur at the lower (gate or transistor) levels. It is thus important to raise process variation awareness to a higher level, because the benefits gained from higher level optimization often far exceed those obtained through lower-level optimization. In addition, higher level statistical analyses enable early design decisions to consider lower-level process variations, reducing expensive design costs.

The HLS process usually consists of module selection, scheduling, resource binding, and clock selection. Traditionally, each step of HLS has performed with worst-case performance parameters for each resource, such as adders and multipliers. However, each functional unit's performance is no longer fixed value, but instead is represented by a probability density function (PDF).

To bring process variation awareness to the HLS flow, the resource library should be characterized as a form with PDFs. Fig. 1 shows a simple characterization flow of functional units [9]. At first, all the gates in a standard cell library are statistically characterized. After the gatelevel synthesis of functional units in the resource library, the delay PDF of functional units can be obtained using statistical timing analysis. We will discuss statistical timing analysis later.

With the characterized variation-aware resource



Fig. 1. A simple characterization flow for functional units.

library and statistical analysis methods, designers can use yield-driven HLS algorithms to perform design space explorations statistically and to search for solutions to improve performance yield.

One of the recent works on variation-aware HLS, by Feng Wang et al, introduced a new concept called statistical performance improvement, which is equivalent to performance yield improvement [10]. They proposed statistical performance improvement as the function of the statistical path delay improvement and the criticality to effectively represent performance improvements due to resource sharing or binding. The statistical path delay represents the magnitude of the path delay based on the statistical timing analysis and the statistical path delay improvement is the difference between the statistical path delay before and after resource sharing or binding. Criticality in HLS is defined as the probability of the operation being on the critical path. They used the cost function of statistical performance improvement as an effective metric to guide optimization during the resource sharing and binding steps. However, other steps they took (such as scheduling) remained conventional. Their work demonstrated that integrating performance yield into design space exploration can satisfy the yield requirement, and they achieved an area reduction of 30% under a 99% performance yield constraint, and runtime overhead of 10% compared to the traditional deterministic method.

Jung and Kim proposed a new yield computation technique that can handle the non-Gaussian timing variation of functional units and the correlation between resources [11]. The statistical timing methods used by the current variation-aware HLS techniques generally assume that the delay variations of resources follow the Gaussian distributions and are independent. However, the delay distributions of the circuit (gate) primitives in the modules are inherently correlated. Therefore, the capability of supporting non-Gaussian delay distributions and correlations between resources is required for the accurate computation of performance yields in HLS. They used the statistical analysis method in [12] to handle non-Gaussian variation sources. In addition, they formulated the correlation of delay variables between resources using Fourier series, integrating an incremental computation of non-Gaussian delay distributions into the performance yield computation in HLS.

Chen and Xie used latches as storage elements instead of flip-flops to improve the performance yield in HLS [13]. Latches have advantages in tolerating process variations compared with flip-flops, since latches are transparent during the active clock period and have the extra capability of passing time slacks between control steps. They also introduced the critical operation concept, which is used to solve hold-time violations and resource sharing problems of latch replacement. An average 28% yield improvement could be achieved using their latch replacement technique.

As stated previously, statistical timing analysis methods are essentially required to characterize the delay variations of functional units in HLS. Statistical timing analysis can be classified into two categories: path-based and block based. Path-based timing analysis predicts the maximal delay for a number of pre-selected logic paths and can be conducted with the consideration of random process variations [1]. The path-based technique is efficient and accurate if a few critical paths can be easily identified. On the other hand, block-based timing analysis propagates arrival times on a timing graph in a breadth-first order. It does not require pre-selecting any critical paths, and its computational complexity linearly scales with circuit size [14]. The block-based technique has been widely used for the full-chip timing analysis of digital circuits, where many equally critical paths may exist after timing optimization.

In the statistical timing analysis, delay variations are approximated as the linear models [15, 16]:

$$D = d_0 + \sum_{i=1}^n d_i X_i + d_{n+1} X_m$$
(1)

or the quadratic models [17]:

$$D = d_0 + \sum_{i=1}^n \sum_{j=1}^i A_{ij} X_i X_j + \sum_{i=1}^n B_i X_i + d_{n+1} X_m \quad (2)$$

where d_0 is the nominal delay of a component. The independent, normally distributed random variables X_i , X_j , and X_m model the variations in the process parameters; X_i and X_j are the correlated components of these variation parameters, such as channel length; and X_m is the purely random component. A_{ij} is the quadratic coefficient, and B_i and d_i are the linear coefficients.

Given delay models in Eqs. (1, 2), the basic operation in statistical timing analysis is to evaluate SUM or MAX and approximate the result as a new delay model of X_i . For the linear delay model in Eq. (1), the SUM operation can be easily handled by

$$x + y = (d_{0,x} + d_{0,y}) + (\mathbf{D}_{x} + \mathbf{D}_{y})^{\mathrm{T}} \mathbf{X} + (d_{n+1,x} + d_{n+1,y}) X_{m}$$
(3)

A similar formulation can be derived for the quadratic delay model in Eq. (2):

$$x + y = (d_{0,x} + d_{0,y}) + \mathbf{X}^{\mathsf{T}} (\mathbf{A}_{\mathsf{x}} + \mathbf{A}_{\mathsf{y}}) \mathbf{X}$$
$$+ (\mathbf{B}_{\mathsf{x}} + \mathbf{B}_{\mathsf{y}})^{\mathsf{T}} \mathbf{X} + (d_{n+1,x} + d_{n+1,y}) X_{m}$$
(4)

In Eqs. (3, 4), since the SUM operation is linear, adding two linear/quadratic models results in a new linear/quadratic model. However, the MAX operation is non-linear, and is thus much more difficult to approximate.

To perform the statistical MAX operation, the authors of [15, 16] proposed finding an approximated linear model by matching the first and second order moments. When x and y are approximated as a form with Eq. (1), they firstly calculated the first and second order moments for x and y. Using these moments, they obtained MAX(x, y) as an approximated linear model. Another approach for linear MAX approximation is based on the tightness probability proposed in [9]. The concept of tightness probability is related to the first-order Taylor expansion.

X. Li et al extended first-order statistical Taylor expansion to second order for quadratic MAX operations [17]. Quadratic approximation is more accurate, but also more expensive, than a simple linear approximation. In practice, quadratic MAX approximations should be selectively applied, depending on the accuracy and complexity requirements of a specific application.

On the other hand, post-silicon tuning techniques have been introduced that allow adjustment of device characteristics after a die has been manufactured to compensate for the specific deviations that occurred due to process variations [18]. Techniques such as adaptive body biasing (ABB) and adaptive supply voltage can be used to tune the manufactured chips, thus reducing variations in circuit performance.

For example, ABB techniques can effectively tighten performance distribution and minimize yield loss due to process variations. The body bias voltage for each individual die is different, and therefore, the body bias voltage is statistically distributed according to the probability distribution of performance. One of the postsilicon tuning methods using ABB, proposed by S. H. Kulkarni et al, clustered gates at design time into independent body bias groups, which were then individually tuned post-silicon for each die [18]. They generated the probability distribution of the post-silicon ideal body bias voltage using sampling methods for each gate. Then, they used these distributions and their correlations to derive a statistically-aware clustering technique. Their ABB approach produced designs with 2-9 times tighter delay distributions in comparison to the dual V_{th} assignment method and power reduction of 38 ~ 71%.

Post-silicon tuning methods can be also applied at the module level and can change modules' statistical characteristics. This tuning broadens the optimization space for variation-aware HLS algorithms. Meanwhile, when considering the granularity of post-silicon tuning techniques, module-level tuning is favorable because of its relatively low tuning cost [9].

IV. THE VARIABILITY ON CIRCUIT DEGRADATION

Continued miniaturization of the semiconductor process has caused new problems that were not taken into account in the past, including reliability issues such as Hot Carrier Injection (HCI), Bias-Temperature Instability (BTI) and Electromigration (EM), and the importance of such problems is gradually increasing for complementary metal oxide semiconductor (CMOS) devices of high reliability. In particular, considerations of these reliability issues can greatly affect product competitiveness from the early stages of a design since the time to market (TTM) of product development is diminishing.

For example, negative bias temperature instability (NBTI), which takes place when the V_{gs} of PMOS is lower than zero in other words, when the PMOS is turned on is one of these representative reliability issues. It increases the threshold voltage (V_{th}) of PMOS over time, causing alterations in circuit performance, including time delays. Various studies have been carried out to minimize the effect of NBTI the circuits, specifically regarding modeling, analysis, and design techniques of NBTI [19-22].

However, circuit aging by NBTI is also affected by the magnitude of T_{ox} and V_{th0} , and thus, process variation eventually affects NBTI. If the circuit time delay is expressed as (1), the NBTI, considering the process variation, increases not only the mean time delay $(d_{t=0})$ of the circuit, but also the standard deviation $(\sigma_{total,t=0} = \sqrt{\Sigma d^2_{t=0,i}})$ [18]. As shown in Table 1, change of the threshold voltage by Tox increased by about 2% when compared with the same change utilizing the conventional long-term model [19]. The effect of V_{th0} was 0.5%, smaller than that of T_{ox} . When both T_{ox} and V_{th0} were considered, the mean was increased by about 2.5% and the standard deviation was about 15% of the mean, larger than the other two cases. Therefore, the effects of process variation on aging phenomena such as NBTI, HCI, and EM should be considered to precisely analyze circuit aging and apply the optimum guard-band at the design-level.

One of the recent aging variation-related works shows the statistical variation of NBTI on random logic circuits and 6T SRAM cells [3]. They considered simultaneous random dopant fluctuation (RDF) and NBTI induced V_{th} variation (σ_{RDF} and σ_{NBTI}) as follows:

$$\sigma_{Vth} = \sqrt{\sigma_{RDF}^2 + \sigma_{NBTI}^2(t)}$$
(5)

where σ_{Vth} represents the total V_{th} variation after time t.

Their results showed that the standard deviation of

delay for logic circuits can change significantly (~ 100% at 32 nm) because of the impact of NBTI variations. In addition, their results on circuits designed at 22 nm technology nodes show a larger impact of 1.5 times the 32 nm node results. The NBTI variation also has an effect on the read static noise margin (SNM) of SRAM cells. If the NBTI variation is considered, the read SNM is reduced to a half. Hence, the NBTI variation should be considered for an accurate circuit analysis.

R. Kanj, et al also proposed a statistical optimization method for dual supply SRAM design under the NBTI variation [24]. They used a similar model with Eq. (4) for the NBTI variation. However, these works ignored the possible correlation between RDF and NBTI. For example, if V_{th} is skewed by the RDF variation, the oxide field and doping concentration also change, which could have a non-negligible effect on the NBTI variation. Therefore, this correlation should be considered to establish a more accurate NBTI model.

Table 1. The effect of process variation on NBTI (ΔV_{th})

Long Term[14]	Tox		Vth0		Tox+Vth0	
	μ	σ	μ	σ	μ	σ
0.158	0.161	0.018	0.159	0.015	0.162	0.024
	1.94%		0.29%		2.43%	

V. CONCLUSIONS

In this paper, we reviewed several recently developed works on timing analysis for below 30 nm technologies. Due to the increase of integration degree involved, traditional transistor-level simulations become too expensive to use in practical situations. In addition, the impact of process variations on circuit performance keeps growing, and aging variations have also become unignorable. Many researchers are attempting to consider these problems and promising results continue to develop for use with future technologies. In the end, it can be concluded that there is a strong need for a fast/accurate analysis method and an analytical model that accurately describes the impact of process/aging variations on circuits.

ACKNOWLEDGMENTS

This paper has been supported by Nano IP/SoC Promotion Group of Seoul R&BD Program (10560) in 2010.

REFERENCES

- [1] Xin Li, Jiayong Le, and Lawrence T. Pileggi, *Statistical Performance Modeling and Optimization*, Lightining Source Inc., 2007.
- [2] ITRS 2009 Design Roadmap, http://www.itrs.net/ Links/2009ITRS
- [3] Sang Phill Park, K. Roy, and K. Kang, "Reliabiliability Implications of Bias-Temperature Instability in Digital ICs," *IEEE Design and Test of Computers*, Vol.26, No.6, pp.8-17, 2009.
- [4] K. Gulati, J. F. Croix, S. P. Khatri, and R. Shastry, "Fast Circuit Simuation on Graphics Processing Units," *Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp.403-408, Jan. 2009.
- [5] X. Zhou, Y. Wang, and H. Yang, "DCCB and SCC based Fast Circuit Partition Algorithm for Parallel SPICE simulation," *International Conference on* ASIC (ASICON), pp.1247-1250, Oct. 2009.
- [6] E. Pebesma and G. Heuvelink, "Latin Hypercube Sampling of Gaussian random fields," *Technometrics*, Vol.41, No.4, pp.303-312, 1999.
- [7] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2005.
- [8] A. Singhee and R. Rutenbar, "From Finance to Filp flops: A Study of fast quasi-Monte Carlo methods from Computational finance applied to Statistical Circuit Analysis," *International Symposium on Quality Electronic Design (ISQED)*, pp.685-692, March 2007.
- [9] Y. Xie and Y. Chen, "Statistical High-Level Synthesis under Process Variability," *IEEE Design* and Test of Computers, Vol.26, No.4, pp.78-87, 2009.
- [10] F. Wang, Y. Xie, and A. Takach, "Variation-aware Resource Sharing and Binding in Behavioral Synthesis," Asia and South Pacific Design Automation Conference (ASP-DAC), pp.79-84, Jan. 2009.

- [11] J. Jung and T. Kim, "Timing Variation-aware High-Level Synthesis Considering Accurate Yield Computation," *International Conference on Computer Design (ICCD)*, pp.207-212, Oct. 2009.
- [12] L. Cheng, J. Xiong, and L. He, "Non-Linear Statistical Static Timing Analysis for Non-Gaussian Variation Sources," *Design Automation Conference* (*DAC*), pp.250-255, June 2007.
- [13] Y. Chen and Y. Xie, "Tolerating Process Variation in High-Level Synthesis Using Transparent Latches," Asia and South Pacific Design Automation Conference (ASP-DAC), pp.73-78, Jan. 2009.
- [14] C. Visweswariah, K. Ravindran, K. Kalafala, S. Walker, S. Narayan, D. Beece, J. Piaget, N. Venkateswaran, and J. Hemmett, "First-Order Incremental Block-Based Statistical Timing Analysis," *Computer-Aided Design of Integrated Circuits and Systems*, Vol.25, Issue 10, pp.2170-2180, Oct. 2006.
- [15] C. Hongliang and S. Sapatnekar, "Statistical timing analysis under spatial correlations," *Computer-Aided Design of Integrated Circuits and Systems*, Vol.24, Issue 9, pp. 1467-1482, Sept. 2005.
- [16] S. Tsukiyama, M. Tanaka, and M. Fukui, "A statistical static timing analysis considering correlations between delays," *Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp.353-358, Jan 2001- Feb 2001.
- [17] X. Li, J. Le, P. Gopalakrishnan, and L. Pileggi, "Asymptotic Probability Extraction for Nonnormal Performance Distributions," *Computer-Aided Design* of Integrated Circuits and Systems, Vol.26, Issue 1, pp.16-37, Jan. 2007.
- [18] S. H. Kulkarni, D. Sylvester, and D. Blaauw, "A Statistical Framework for Post-Silicon Tuning through Body Bias Clustering," *International Conference on Computer-Aided Design (ICCAD)*, pp.39-46, Nov. 2006.
- [19] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Scalable model for predicting the effect of negative bias temperature instability for reliable design," *Circuits, Devices & Systems, IET*, Vol.2, No.4, pp.361-371, 2008.
- [20] B. C. Paul, K. Kang, H. Kufluoglu, M. A. Alam, and K. Roy, "Temporal performance degradation under NBTI: Estimation and design for improved

reliability of nanoscale circuits," *Design, Automation and Test in Europe (DATE)*, pp.780-785, Mar. 2006.

- [21] D. R. Bild, G. E. Bok, and R. P. Dick, "Minimization of NBTI Performance Degradation Using Internal Node Control," *Design, Automation* and Test in Europe (DATE), pp.148-153, Apr. 2009.
- [22] M. B. da Silva, V. V. A. Camargo, L. Brusamarello, G. I. Wirth, and R. da Silva, "NBTI-aware technique for transistor sizing of high-performance CMOS gates," *Latin American Test Workshop* (*LATW*), pp.1-5, Mar. 2009.
- [23] B. Vaidyanathan, A.S Oates, Y. Xie, and Y. Wang, "NBTI-aware statistical circuit delay assessment," *International Symposium on Quality Electronic Design (ISQED)*, pp.13-18, Mar. 2009.
- [24] R. Kanj, R. Joshi, C. Adams, J. Warnock, and S. Nassif, "An Elegant Hardware-corroborated Statistical Repair and Test Methodology for Conquering Aging Effects," *International Conference on Computer-Aided Design (ICCAD)*, pp.497-504, Nov. 2009.



Juho Kim received B.S degree and Ph.D degree in Computer and Information Science from University of Minnesota in 1987 and 1995, respectively. After getting Ph.D degree, he worked as a senior member of technical staff at Cadence

Design System until 1997. Professor Kim joined the department of computer science and engineering in Sogang University, Seoul, Korea in 1997, and he was a department chair from 2005 to 2008.



Sangwoo Han received the BS degree in Physics and Computer Science from Sogang University in 2005. He also received the MS degree in Computer Science and Engineering from Sogang University in 2008. He is currently pursuing a

Ph.D. degree in computer engineering at Sogang University. His research interests are SSTA, robust design for process variation, and design for reliability enhancement.



Roy Jewell received B.S. and M.A. degrees in physics from University of South Florida, and M.S. degree in management and administration science from University of Texas, Dallas. Jewell joined Magma in 2001 as President and Chief Operating

Officer. From 1999 to 2001 Jewell served initially as the CEO of Clarisay, Inc. a company he co-founded. From 1998 to 1999 Jewell was a member of the executive staff at Avanti Corporation. From 1992 to 1998 Jewell was president and CEO of Technology Modeling Associates, Inc. Jewell previously served in various marketing positions at TMA, and prior to joining TMA in 1988 he served in various technical and management positions at Texas Instruments.