

Design Sensitivities to Variability: Extrapolations and Assessments in Nanometer VLSI *

Y. Cao[†], P. Gupta[‡], A.B. Kahng[‡], D. Sylvester* and J. Yang*

[†]EECS Department, University of California at Berkeley

[‡]ECE Department, University of California at San Diego

*EECS Department, University of Michigan at Ann Arbor

ycao@eecs.berkeley.edu, (puneet,abk@ucsd.edu), (dennis,jiey@eecs.umich.edu)

Abstract

We propose a new framework for assessing (1) the impact of process variation on circuit performance and product value, and (2) the respective returns on investment for alternative process improvements. Elements of our framework include accurate device models and circuit simulation, along with Monte-Carlo analyses, to estimate parametric yields. We evaluate the merits of taking into account such previously unconsidered phenomena as correlations among process parameters. We also evaluate the impact of process variation with respect to such relevant metrics as *parametric yield at selling point*, and amount of *required design guardbanding*. Our experimental results yield insights into the scaling of process variation impacts through the next two ITRS technology nodes.

1 Introduction

Aggressive technology scaling has introduced new variation sources and made process control more difficult. As a result, future technology nodes are expected to see increased process variation and decreased predictability of nanometer-scale circuit performance [1]. Despite the relaxation of some 3σ tolerances, there are no known solutions for a number of near-term variability control requirements (according to the ITRS [1]). Moreover, observation of key markets that drive the semiconductor industry reveals the potentially large impact of variability on the value of semiconductor products. Semiconductor enterprises must be cognizant of the different risks and ROI opportunities from, e.g., an extra increment of T_{ox} or L_{eff} CD control, versus new design for value design technologies, versus revised performance targets for products, etc. In this work, we describe key elements of a framework that will allow the semiconductor industry to assess impact of process variation on circuit performance, manufacturing cost, and product value in nanometer technologies (180nm through 70nm). Our framework is built on accurate circuit design models, statistical models of process variation, a combination of circuit simulators and analytical performance models, and application of Monte Carlo analyses to estimate parametric yields. We evaluate the merits of taking into account previously unconsidered phenomena such as correlations among process parameters. We also evaluate the impact of process variation with respect to such metrics as *parametric yield at selling point*, and

*This work was in part supported by Semiconductor Research Corporation.

amount of *required design guardbanding*. Key contributions of our work include:

- a self-consistent taxonomy of variations, focusing on both within-die and die-to-die process variation sources;
- accurate models of *correlations* of variation;
- realistic and quantified projection to future process nodes of the impact of variability on critical-path delays; and
- analysis of the sensitivity of performance variation to improved control of individual device parameters, measured by change in number of “sellable” chips produced and extent of guardbanding required to meet a given parametric yield target.

Our experimental results yield surprising insights into the scaling of process variation impacts through the next two ITRS technology nodes, as well as the prioritization of various areas for future technology investment. The latter type of contribution, even if not fully achieved by this initial work, is required for principled allocation of R&D resources among multiple semiconductor supplier industries to solve the variability problem (cf. “shared red bricks” [15])¹.

The remainder of our paper is organized as follows. Section 2 presents a taxonomy of random process variations. Section 3 defines our simulation setup, including circuit models and variability distributions (as well as correlations) for device and interconnect parameters. Section 4 gives results for projection of variability into future technology nodes, as well as interesting views of the sensitivities of overall circuit performance variability to improved process control. Section 5 concludes with ongoing research directions.

2 Taxonomy of Variation

Circuit variability refers to deviations of circuit parameters (e.g., L_{eff} , V_{dd} , interconnect size, etc.) from nominal values. It is introduced either during chip fabrication or due to circuit operation. According to the inherent length scale of such variability, it can be characterized as within-die variations (e.g., interference in optical lithography) and

¹Sources at a major semiconductor vendor indicate that substantial effort and capital has been invested for V_{th} control because of its huge impact on design performance. This anecdotal evidence supports our claim that significant engineering effort and capital investment can be expended to reduce many sources of variability, but that we must focus investment on key sources.

die-to-die variations. In this study, we consider the impact of both types of variations. Our taxonomy of variability is as follows.

- *Intrinsic Variation.* Intrinsic variation is caused by the fabrication of the integrated circuit, in contrast to dynamic variation (sometimes referred to as extrinsic variation), which arises during circuit operation. Intrinsic variability can be further classified as either *systematic* or *random*.
 - *Systematic variation* implies that changes in parameter values, such as L_{eff} , are due to known and predictable phenomena. This is the most significant limiter to performance in future processes [2] [5]. Successful scaling of MOSFET technology to sub-100nm process geometries relies on compensation of systematic variation components at the design and reticle stages. We assume that such corrections to systematic process variation are applied.
 - *Random variation* is due to the inherent unpredictability of the semiconductor fabrication process. Fluctuations in channel doping, gate oxide thickness, and ILD permittivity are primarily due to random variation. As random phenomena cannot be compensated for and are difficult to minimize, this type of variability may eventually pose the most significant challenge to design of adequately yielding nanometer-scale MOSFET circuits.
- *Dynamic Variation.* In contrast to intrinsic variation, dynamic variation is due to factors such as temperature and supply voltage, which vary with the operation of the circuit. These phenomena, while possible to model during the design phase, are difficult to compensate for in that they are transient and not always present. Therefore, designers focus on minimizing the variation itself (e.g., re-designing the power distribution network to ensure no voltage drop greater than 5-10% of Vdd) rather than changing the circuit design itself. Our studies acknowledge dynamic variability in a similar fashion as other variations. We will see below that dynamic variation is a potentially very significant source of performance loss in future technology nodes.

3 Experimental Testbed and Methodology

In this section, we describe elements of our experimental testbed. The key components are

1. a parameterized scalable critical path circuit model;
2. introduction of correlations among the parameter variations; and
3. use of detailed device modeling and circuit simulation within Monte-Carlo methodology.

We study a parameterized critical path, shown in Figure 1, composed of n identical local stages and one long top-level buffered global interconnect. The parameter n is set to 11 at the 180nm technology node and is reduced by one in each subsequent generation to reflect aggressive pipelining techniques and other micro-architectural advancements. In each local stage, a 2-input NAND gate drives a short local

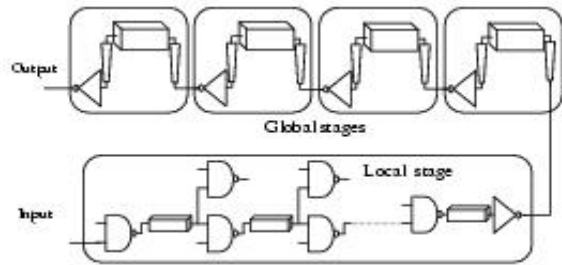


Figure 1: Critical path structure for performance study.

line with length estimated by [1]. The NAND is sized to optimize the speed-power tradeoff (fanout=2), i.e., the knee of the delay vs. sizing curve as in the Berkeley Advanced Chip Performance Calculator (BACPAC) [12]. The global line length remains the same in all technology nodes at 10mm, consistent with the 2001 ITRS projections of fixed die size for future microprocessors [1]. Optimal inverting repeaters are inserted at even intervals into the global line to minimize delay. Parasitic via resistance is considered. Overall, we closely follow the 2001 ITRS (high-performance MPU) critical path model [1]. For each local stage and the global line, we add two parallel neighboring lines for delay and noise analysis. Input transition times to initial stages are set at 20% of the clock period. Further details of the critical path models, including line lengths, gate sizes, and signal transition times, are listed in Table 1. The nominal dimensions of interconnect are taken from [14]. The simulation setup includes an active line on the critical path coupled with two quiet identical neighboring lines. Each line is modelled as a sufficiently long chain of L segments to capture the distributed RLC characteristics.

Previous works assume that variation sources are independent of each other [3]. By contrast, our work recognizes several strong correlations. Specific examples include:

- V_{th} is a function of T_{ox} , N_{ch} , L_{eff} and X_t , calculated from a delta-doping approximation and BSIM3v3 models.
- Corresponding parameters of NMOS and PMOS have a correlation coefficient of one, i.e., NMOS and PMOS in the same gate exhibit the same deviations from respective means.
- Assuming a fixed wire pitch, wire spacing variation is the negative of wire width variation. Metal thickness (T) and underlying interlevel dielectric (ILD) thickness (H) are negatively correlated with a correlation coefficient of -0.5 (this value stems from the relationship of trench etch depth in damascene processes as well as chemical-mechanical polishing effects which act to reduce the correlation between T and H).
- Spatially proximate devices and interconnections (e.g., in local stages) have similar variations.
- We model the spatial correlation among devices by incorporating a distance-dependent correlation parameter. This correlation decays linearly with distance to a value of zero (implying complete independence) over a length scale of 1cm [6].

Technology	180nm		130nm		100nm		70nm	
Device	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS	NMOS	PMOS
L_{eff} (nm)	100 \pm 16.7%	100 \pm 16.7%	65 \pm 16.7%	65 \pm 16.7%	45 \pm 16.7%	45 \pm 16.7%	28 \pm 16.7%	28 \pm 16.7%
T_{ox} (Å)	22 \pm 4%	22 \pm 4%	16 \pm 4%	16 \pm 4%	13 \pm 4%	13 \pm 4%	10 \pm 4%	10 \pm 4%
R_{dsW} ($\Omega - \mu m$)	250 \pm 10%	450 \pm 10%	200 \pm 10%	400 \pm 10%	180 \pm 10%	300 \pm 10%	150 \pm 10%	280 \pm 10%
X_i (nm)	30		24		20		13	
V_{th} (V)	0.214 \pm 30%	-0.327 \pm 30%	0.232 \pm 30%	-0.273 \pm 30%	0.217 \pm 30%	-0.253 \pm 30%	0.169 \pm 30%	-0.218 \pm 30%
Interconnect	Local		Global		Local		Global	
E	3.5 \pm 3%		3.2 \pm 5%		2.8 \pm 5%		2.2 \pm 5%	
w(nm)	250 \pm 20%	525 \pm 20%	175 \pm 20%	335 \pm 20%	123 \pm 20%	237 \pm 20%	85 \pm 20%	160 \pm 20%
s(nm)	250 \pm 20%	525 \pm 20%	175 \pm 20%	335 \pm 20%	123 \pm 20%	237 \pm 20%	85 \pm 20%	160 \pm 20%
t(nm)	500 \pm 10%	1050 \pm 10%	280 \pm 10%	670 \pm 10%	197 \pm 10%	498 \pm 10%	145 \pm 10%	325 \pm 15%
h(nm)	500 \pm 15%	1050 \pm 15%	280 \pm 15%	670 \pm 15%	197 \pm 15%	498 \pm 15%	145 \pm 15%	325 \pm 15%
ρ (Ω -m)	2.2e-8 \pm 30%		2.2e-8 \pm 30%		2.2e-8 \pm 30%		2.2e-8 \pm 30%	
Rvia(Ω)	23 \pm 20%		25 \pm 20%		27 \pm 20%		29 \pm 30%	
Length(μm)	62.5	5000	55.56	3333	50	2500	45.45	2000
Wn(μm)	1.26	15	1.01	9.75	0.76	6.75	0.48	4.2
Dynamic								
V_{dd} (V)	1.8 \pm 10%		1.2 \pm 10%		1.0 \pm 10%		0.9 \pm 10%	
Tr(ps)	160		125		110		87	
Temp ($^{\circ}C$)	25		25		25		25	

Table 1: Parameter values and 3σ variations

- Our interconnect spatial correlation modeling is more involved. We divide the global line into $100\mu m$ segments. Interconnect parameters within each segment are perfectly correlated. We assume that correlation between segments decays linearly with separation. At a certain distance, this correlation equals zero. For interconnect width and space, we take this separation distance to be $5mm$ for all the technology nodes while it is $2mm$ for metal thickness and ILD thickness. Numerous prior studies have investigated the concept of CMP planarization length; this relates to the distances over which features can be considered to be correlated due to pad deformation and other physical phenomena. This planarization length is typically found to be on the order of $2mm$, motivating our choice of separation distance.

We assume parameter variations to be normally distributed with mean and sigma values derived from [1], [14] and industry sources. In [1], the allowed variability in physical gate length is fixed at 10%. The magnitude of the physical gate length is approximately half of the technology node, or the DRAM half-pitch. Translating this uncertainty to effective channel length, which is also a fraction of physical gate length due to source-drain extension (SDE) underdiffusion, we expect a 3σ for L_{eff} of greater than 10%. In this work, we approximate $L_{eff} = 0.6 \times L_{physical}$, leading to a 3σ process tolerance throughout the roadmap of 16.7%. Different approaches may be taken including an assumption that the SDE underdiffusion has a fundamental lower limit that pushes L_{eff} to be a smaller fraction of $L_{physical}$. This will result in either (1) larger uncertainty in L_{eff} or (2) less aggressive scaling of $L_{physical}$ to compensate. Either of these alternatives can be readily investigated in our framework.

We perform circuit simulation with a distributed-lumped RLC interconnect model and all correlations included. Figure 2 compares the delay distributions obtained using our Monte Carlo simulation methodology for (1) RC interconnect model without correlations, (2) RLC interconnect model without correlations and (3) RLC interconnect model with correlations. This demonstrates the effect of more accurate and detailed circuit modeling for purposes of eventual accurate assessment of variability impacts.

In contrast with the linear regression analysis used in [3], our studies use a Monte Carlo (MC) approach with 5000 trials where the variation sources all vary simultaneously. Each model of process variability, at each technology node, gives rise to 5000 sets of random parameter values within the critical path model which we simulate using HSPICE.

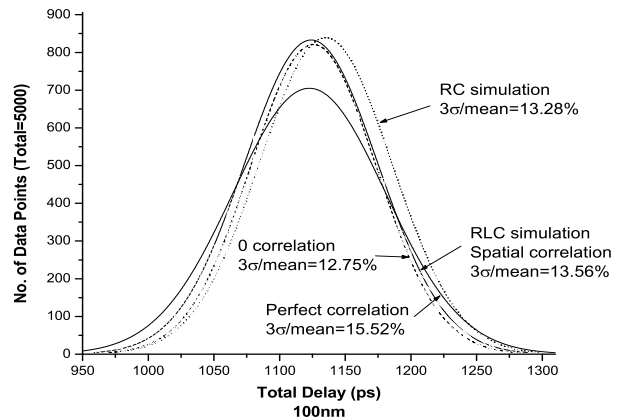


Figure 2: Comparison of RC vs. RLC and RLC with correlations vs. RLC without correlations studies

We then investigate the resulting delay distributions in the critical path model, in an attempt to gauge (1) the true impact of variability on circuit performance, and (2) the true value of developing, e.g., 1nm better control on interconnect thicknesses.

4 Impact on Future Circuit Performance

To assess the impact of process variation on critical path delay we adopt two different metrics as outlined below.

1. *Selling point parametric yield.* We assume target parametric yield to be 99.7%. This corresponds to $mean + 3\sigma$ point on the delay distribution and is taken to be the *selling point* of the chip. We take the delay distributions for values drawn from Table 1 as the “baseline” results for all technologies. The selling point is calculated from the baseline distribution. The change in parametric yield at the selling point is then taken as a measure of impact of process variation.
2. *Guardbanding Analysis.* Guardbanding is the typical approach followed in industry to account for variability. A larger guardband implies a more conservative design and hence is not preferred. The expected (“designed-for”) value of performance is given

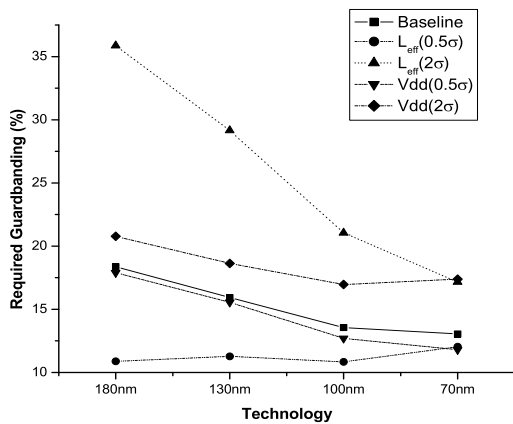


Figure 3: Effect of process control on required guardbanding to achieve 99.7% parametric yield.

by the mean of the delay distribution. Thus, the difference between the selling point and the mean gives the amount of guardbanding required. I.e. $\frac{3\sigma}{mean}$ expressed as a percentage gives the required guardbanding.

We have conducted experiments that vary all parameters listed in Table 1, but due to space constraints we report results only for V_{dd} and L_{eff} ; these are two of the most significant contributors to process variation impact.

4.1 Cumulative effect of all parameter variations

We simulate the critical path and measure delay with *all* the parameters varying with 3σ and mean values as specified in Table 1. This simulation result is taken as the “baseline” result for all the comparisons and analysis explained in the subsequent subsections. Figure 3 shows the baseline delay variation for the four technology nodes. The $\frac{3\sigma}{mean}$ value drops from 18.5% for the 180nm node to 13% for the 70nm node. As the MC-predicted performance-variation is not as severe as previously reported [8], there may be some flexibility to tradeoff between process control and performance control, i.e., we can relax process control to reduce costs without substantial performance penalties.

4.2 Sensitivity Analysis

To determine how sensitive performance is to individual parameter tolerances, we changed the σ values from those in Table 1 to 0.5 and 2 times their original values. This was done for L_{eff} and V_{dd} individually while maintaining the normal σ for other parameters for each technology node. Figure 3 shows the impact on guardbanding of varying L_{eff} and V_{dd} control. Figure 4 and 5 show the impact on selling point yield. Loose $L_{eff}(V_{dd})$ control can cause loss of up to 5%(2%) in yield. It is clear that impact of L_{eff} control is greater than that of V_{dd} but the difference continues to reduce such that at 70nm V_{dd} control is as valuable as L_{eff} control.

With increasing power budgets and lower supply voltages, supplying a stable V_{dd} requires a low-impedance

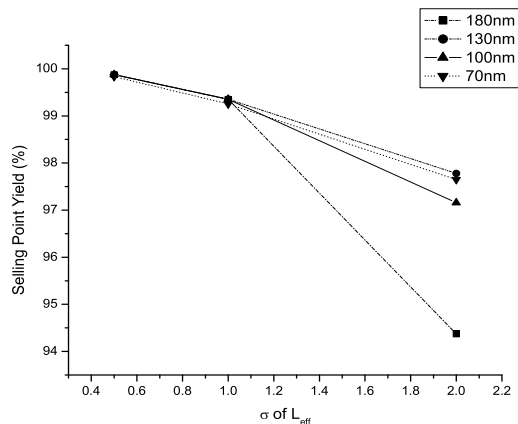


Figure 4: Effect of L_{eff} control on selling point parametric yield.

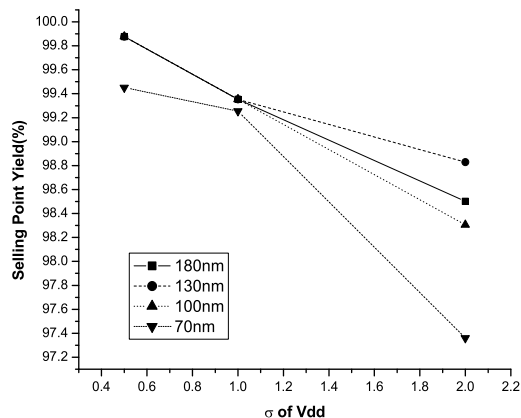


Figure 5: Effect of V_{dd} control on selling point parametric yield.

power distribution network. To achieve this, a larger fraction of routing resources need to be allocated to power distribution, thus increasing the cost associated with reduced IR drop. In contrast, a fixed amount of routing resources for power supply could be used if performance sensitivity to V_{dd} is low. Our results indicate that the latter is not feasible due to intense sensitivity of delay to V_{dd} variation. Distributing a reliable voltage supply can be addressed at both the design and process levels. For instance, for metal layers with a large fraction of routing allocated to the power grid, higher aspect ratio lines should be used than for signal routing layers since coupling capacitance is not an issue in power grid routing.

4.3 Impact of Technology Roadmap Deceleration

In this subsection, we consider what happens when no further (beyond 180nm technology) investments are made in control of a given process parameter, while control of other parameters scales according to Table 1. In other words,

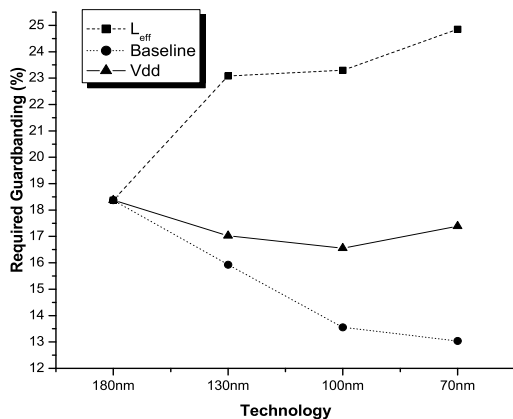


Figure 6: Impact of technology roadmap deceleration in V_{dd} and L_{eff} control on required guardbanding.

the absolute σ value for the given parameter (here V_{dd} and L_{eff} as examples) is kept constant at its 180nm technology node value. Figure 6 gives the “worst-case” impact of no further investments for control of a given parameter.

Our results confirm the prevailing wisdom that sensitivity of performance to L_{eff} variation is high. On the other hand L_{eff} control is very expensive and may not offer the best ROI, in terms of metrics highlighted above, as process technology scales. It may be more cost-effective to tackle L_{eff} variation from a design perspective rather than a process perspective; this is a major focus of our ongoing work.

5 Conclusions and Ongoing Research

In this paper, we have presented a new framework for assessing the impact of process variation on circuit performance, product value and return on investment on alternative process improvements. We apply new metrics such as guardbanding and parametric yield at selling point. We have presented a self-consistent taxonomy of variations. We use accurate models of correlations and Monte Carlo techniques based on circuit simulation. Our main conclusions are as follows.

- The impact of variability is decreasing whether measured as the amount of guardbanding required to circumvent it or the decrease in parametric yield that may need to be tolerated.
- There are both process and design implications of variability. Variability impact can be restricted by innovative design and this should be preferred due to very costly nature of process improvement techniques.
- Performance is very sensitive to L_{eff} variation but the huge cost of L_{eff} control motivates the need for design methods to contain the effect of its variation.
- V_{dd} is an important source of variation and its control may give a better ROI than L_{eff} control for achieving the same amount of variability-tolerance as technology scales.

Our results suggest the potential utility of *Design for Value* methodologies which take variability into account for design optimizations. For instance, one may argue in favour of *selling point optimization* rather than traditional nominal performance optimization. Also, we may have multiple selling points with some pre-specified *value* associated with each selling point. The total design value is then given by $\Sigma v(f) * yield(f)$, for a given value function v of performance measure f , and given parametric yield distribution $yield(f)$. Design for value seeks to find values of design parameters to maximize value, assuming normally distributed process parameters. Our ongoing work looks into possibilities of such probabilistic optimizations, and seeks to quantify value associated with the process and the cost associated with process control. Finally, we are running more comprehensive experiments to assess sensitivity of our variability impact projections not only to additional process parameters but also to factors such as values of spatial correlations.

References

- [1] *2001 International Technology Roadmap for Semiconductors (ITRS)*, <http://public.itrs.net>.
- [2] M. Orshansky, *Personal communication*.
- [3] S. R. Nassif, “Delay Variability: Sources, Impacts and Trends”, *ISSCC*, 2000, pp. 368-369.
- [4] S. R. Nassif, “Design for Variability in DSM Technologies”, *ISQED*, 2000, pp. 451-454.
- [5] M. Orshansky, L. Milor, P. Chen, K. Keutzer and C. Hu, “Impact of Systematic Spatial Intra-Chip Gate Length Variability on Performance of High-Speed Digital Circuits”, *ICCAD*, 2000, pp. 62-67.
- [6] D. Boning and S. Nassif, *Design of High-Performance μP Circuits, Chapter 6: Models of Process Variation in Device and Interconnect*, 2001, pp. 98-116.
- [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”, *ISSCC*, 2001, pp. 278-279.
- [8] K. A. Bowman and J. D. Meindl, “Impact of Within-Die Parameter Fluctuations on Future Maximum Clock Frequency Distributions”, *CICC*, 2001, pp. 229-232.
- [9] X. Qi, G. Wang, Z. Yu, R. Dutton, T. Young and N. Chang, “On-Chip Inductance Modeling and RLC Extraction of VLSI Interconnects for Circuit Simulation”, *CICC*, 2000, pp. 487-490.
- [10] T. Sakurai and R. Newton, “Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas”, *IEEE Journal of Solid State Circuits*, 25(2), 1990, pp. 584-594.
- [11] T. Sakurai, “Closed-Form Expressions for Interconnect Delay, Coupling and Crosstalk in VLSI’s”, *IEEE Transactions on Electron Devices*, 40(1), 1993, pp. 118-124.
- [12] D. Sylvester and K. Keutzer, “System-Level Modeling Using BACPAC”, *International Workshop on System-level Interconnect Prediction*, 1999, pp. 109-114.
- [13] S.-C. Wong, G.-Y. Lee and D.-J. Ma, “Modeling of Interconnect Capacitance, Delay and Crosstalk in VLSI”, *IEEE Transactions on Semiconductor Manufacturing*, 40(1), 2000, pp. 108-111.
- [14] “Berkeley Predictive Technology Model”, <http://www-device.eecs.berkeley.edu/ptm>
- [15] A. B. Kahng, “The Road Ahead: Shared Red Bricks”, *IEEE Design and Test*, March 2002, pp. 70-71.