Modeling and Forecasting of Manufacturing Variations

Sani R. Nassif
IBM Austin Research Laboratory, 11501 Burnet Rd., Austin, TX 78758, USA

1 Abstract

Process-induced variations are an important consideration in the design of integrated circuits. Until recently, it was sufficient to model die-to-die shifts in device performance, leading to the well known worst-case modeling and design methodology [1, 2]. However, current and near-future integrated circuits are large enough that device and interconnect parameter variations within the chip are as important as those same variations from chip to chip. This presents a new set of challenges for process modeling and characterization and for the associated design tools and methodologies.

This paper examines the sources and trends of process variability, the new challenges associated with the increase in within-die variability analysis, and proposes a modeling and simulation methodology to deal with this variability.

2 Sources of Variability

The electrical performance of an integrated circuit is impacted by two distinct sources of variation:

- *Environmental factors* which include variations in power supply voltage and temperature. These factors are highly design dependent and exhibit time constants similar in scale to the clock frequency.

- *Physical factors* which include variations in the electrical and physical parameters characterizing the behavior of active and passive devices. These variations are caused by processing and mask imperfections and various wearout mechanisms (e.g. electromigration). These factors exhibit long time constants, typically measured in years, and can be further divided into two categories:

  – *Die-to-Die Physical Variations* which are largely independent of the design implementation and are usually modeled using worst-case corners.

  – *Within-Die Physical Variations*, the most pronounced of which is the variation in polysilicon gate dimension, and which depend on the design implementation (layout) and for which no general effective modeling and analysis methodologies yet exist.

We denote the physical sources of variability by $\mathcal{P}$, and divide it into the *device* variations which we will denote by $\mathcal{D}$ and the *wire* or interconnect variations which we will denote by $\mathcal{W}$. $\mathcal{D}$ contains parameters such as V_{th}, T_{ox}, while $\mathcal{W}$ contains parameters such as wire geometry and sheet resistivity. If $\mathcal{P}$ is *constant* within a die, but varies within a wafer or lot, then $\mathcal{P}$ is independent of local differences within the chip, thus we can treat variations in $\mathcal{P}$ as noise *imposed* upon the circuit, and analyze it using worst-case or Monte-Carlo analysis. In this case, we need only the distribution describing $\mathcal{P}$ to estimate the performance variation in the design.

If $\mathcal{P}$ varies within a die because (a) the die is large relative to the wafer, or (b) $\mathcal{P}$ has a strong layout dependence (e.g. nested vs. isolated effect on polysilicon line dimensions [3]), then the task of determining the design performance variation becomes more difficult because the number of entities varying is larger [4, 5] and simple worst-case analysis is not possible.

We denote the environmental sources of variability by $\mathcal{E}$ and include in them variations in the power supply and operating temperature. We further note that delay variations induced by cross-talk and other on-chip noise sources can potentially be treated as a source of environmental variability, but that we will not do so in the remainder of this paper.

3 Trends in Variability

In order to assess trends in variability, we use the circuit in figure 1 which is composed of a buffer driving an identical buffer through a length of minimum-width wire. We perform a simulation study of the circuit for a variety of technologies defined in the 1997 SIA technology roadmap[6].

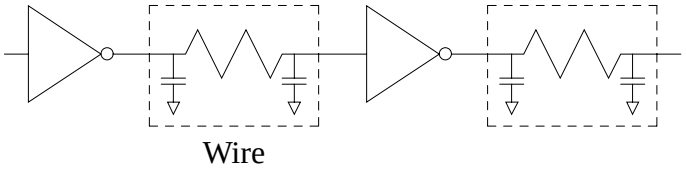

Figure 1: Canonical circuit.

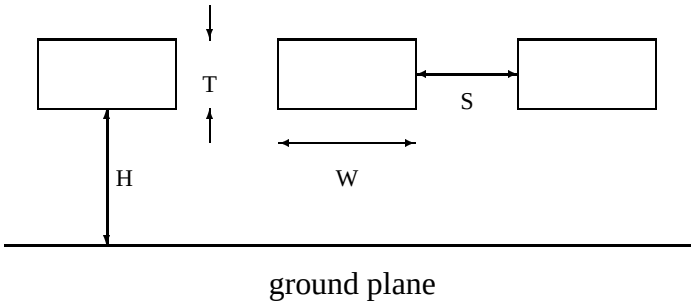

Figure 2: Cross-section showing wire geometry.

In order to assure consistency across the various technologies, we maintained the width-to-length ratio in the buffers, i.e. we scaled the area at the same rate as L_{eff}. We determined the length of minimum-width wire for each technology by observing that the optimal delay is achieved by the application of a buffer insertion strategy[7]. Such a strategy sets the maximum line length at: $L_{max} = \sqrt{2(\tau_B + R_B C_B)/R_w C_w}$, where τ_B, R_B and C_B are the delay, output resistance and input capacitance of the buffer, and R_w and C_w are per unit length of the wire.

We considered five technologies in the 250 to 70 nm gate length range conforming to the 1997 SIA technology roadmap[6], and computed L_{max}. The results are shown in table I and in figure 2 which shows the wire geometry parameters. The salient feature of the table is plotted in figure 2 and shows that L_{max} is not scaling as fast as L_{eff} which recalls the increasing influence of interconnect in advanced technologies.

Year	L_{eff} nm	T_{ox} nm	V_{dd} V	V_T V	W μ	H μ	ρ $\frac{m\Omega}{\square}$	L_{max}
1997	250	5	2.5	0.5	0.8	1.2	45	*2123*
1999	180	4.5	1.8	0.45	0.65	1.0	50	*1920*
2002	130	4	1.5	0.4	0.5	0.9	55	*1670*
2005	100	3.5	1.2	0.35	0.4	0.8	60	*1526*
2006	70	3	0.9	0.3	0.3	0.7	75	*1303*

Table 1: Technology parameters.

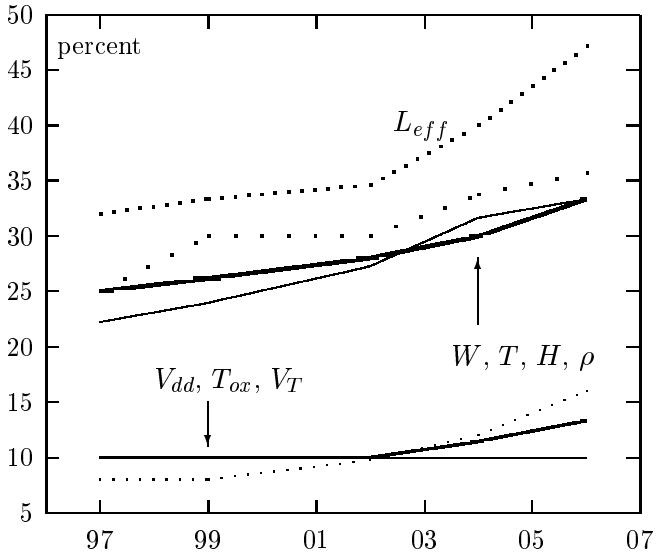

Figure 3: Technology parameter variations.

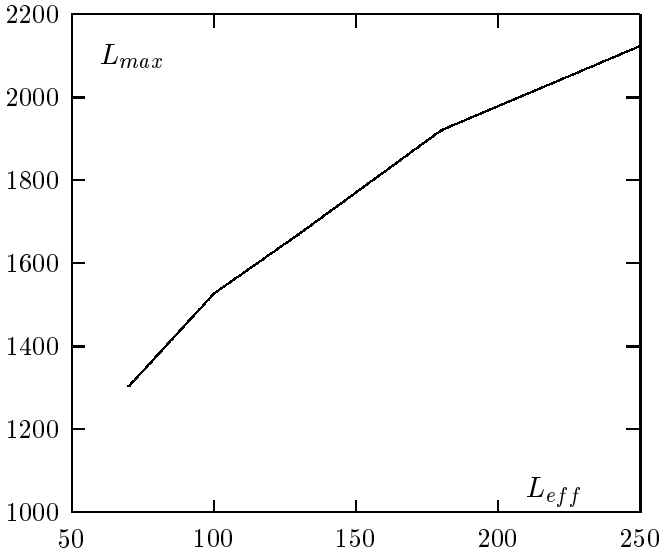

Figure 4: L_{max} vs. L_{eff}.

Year	L_{eff} nm	T_{ox} nm	V_T mV	W μ	H μ	ρ $\frac{m\Omega}{\square}$
1997	80	0.4	50	0.2	0.3	10
1999	60	0.36	45	0.17	0.3	12
2002	45	0.39	40	0.14	0.27	15
2005	40	0.42	40	0.12	0.27	19
2006	33	0.48	40	0.1	0.25	25

Table 2: Technology parameter 3σ variations.

We now turn our attention to the impact of physical and environmental variability on the performance of the simple circuit in figure 1. Table II lists the three-sigma ranges assumed for the physical variables and broadly conforms to the variability assumptions made in the SIA roadmap. In order to assess the environmental variability, we assumed that V_{dd} has a tolerance of +/-10% and that the temperature varies from 25 to 125 degrees Celsius.

We performed the variational assessment by performing 100 simulations for each technology node while varying all the environmental and physical parameters using Latin-Hypercube sampling[8]. We then performed linear regression to build a model for the delay as a function of the sources of variability:

$$T = T_0 + \sum_{i=1}^{N_{\mathcal{E}}} a_i \mathcal{E}_i + \sum_{i=1}^{N_{\mathcal{D}}} b_i \mathcal{D}_i + \sum_{i=1}^{N_{\mathcal{W}}} c_i \mathcal{W}_i \qquad (1)$$

If we center and normalize all distributions to have a mean of zero and a unity standard deviation, and making the natural assumption that the source of variations are independent, we find that the sums of the coefficients a_i, b_i and c_i can be interpreted as the relative (percentage) impact on delay of the device, wire and environmental variations. Figure 4 shows the results of this analysis for the five technologies considered. The radical increase in sensitivity to wire variations can be explained by observing from table III the relatively large increases in variability in the wire parameters. This increase in variability is an important technology limitation and needs to be addressed, especially since -as is clear from table III, the tolerances in the BEOL (Back End Of Line, i.e. wiring levels) is -in fact- assumed to scale slightly better than the FEOL (Front End Of Line, i.e. polysilicon dimensions).

4 Analysis of Variability

A single scalar performance z of a network may be expressed as a function of a vector of designable parameters, $\mathcal{X}$, and a vector of model parameters p:

$$z = f(\mathcal{X}, p) \qquad (2)$$

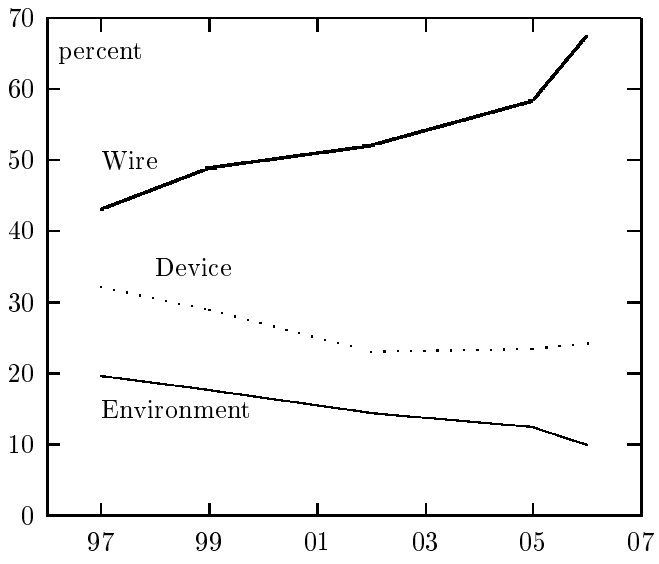

Figure 5: Relative importance of variations in $\mathcal{E}$, $\mathcal{D}$ and $\mathcal{W}$.

For digital integrated circuits, the designable parameter vector $\mathcal{X}$ typically contains device sizes. The parameter vector p contains elements which are physical (e.g. the gate oxide thickness T_{ox}), elements which are empirical (e.g. mobility reduction), and elements which are environmental (e.g. the power supply voltage V_{dd} or temperature). The function f is typically evaluated by simulation. In this paper, we will assume that a circuit simulator such as Spice [9] is used to evaluate f.

Consider the case where there is no within-die variation. The vector p will include the unique set of physical, empirical and environmental parameters needed to predict the behavior of the circuit. For example, for a CMOS circuit, p would include the voltage supply V_{dd}, the temperature T and the parameters of the simulation models for the N and P-channel transistors. The dimension of p is of order 10. New complex transistor models (e.g. BSIM [10]) can appear to increase the dimension of p, but many model parameters will be constant for a given technology.

To model intra-chip variation, a statistical distribution is associated with p. This is done according to the type of parameter:

- The environmental portion of p is usually dealt with based on *specifications*. For example, if the chip is to operate for a range of supply voltage V_{dd} from 1.6 to 2.0, then one might use a uniform distribution over that range.
- The device model portion of p has been studied ex-

tensively and numerous approaches exist for dealing with it (see for example [11]). The model parameters are typically represented by some joint probability density function (JPDF) $N(u, S)$, where u is a vector of means and S is a variance/covariance matrix. Often, we choose to perform a principal component rotation of the parameters to come up with a set of uncorrelated parameters. This has two advantages, it simplifies the analysis algorithms and often reduces the dimensionality of p.

When analyzing the impact of variations, one is often interested in determining the extreme value of z, which we denote by z_{wc} (where the wc stands for worst case), and the conditions under which that extreme value occurs. Assuming without loss of generality that smaller values of z are desirable, this implies finding the value of z_{wc} which bounds (from above) a specified proportion of possible outcomes:

$$Prob(z < z_{wc}) = \zeta \tag{3}$$

The parameters which correspond to this extreme value, z_{wc}, are called the *worst case parameters*, p_{wc}. Since many combinations of parameters can result in the same value of performance, we determine p_{wc} by solving the following maximization problem:

$$\begin{aligned} p_{wc} &= \max_{p} Prob(p) \\ \text{subject to:} \quad & f(p) = z_{wc} \end{aligned} \tag{4}$$

Eq. (3) determines the most probable point (in parameter space) on the $z = z_{wc}$ surface. When f is linear or monotonic in p, this turns out to be an easy computation to find the point on the surface closest to the center of the parameter distribution (see [1] chapter5). For a general function f, the problem is still quite tractable because the dimension of p is small. Figure 1 shows an example where the dimension of p is 2; the ellipses illustrate equi-probability contours of p and the dashed lines denote the values of the performance z.

In principle, given adequate analysis and simulation resources, environmental variations can be predicted by using a circuit simulator to first estimate the power dissipation of individual blocks (see for example [12, 13, 14, 15, 16]). It is then possible to analyze the power distribution network to determine the variations in V_{dd}. It is also possible to model the spatial distribution of power dissipation and use that in a thermal simulator to determine the variations in the temperature T (see for example [17] and [18]).

We are left with the analysis of within-die physical variations. Stine [19] distinguishes four components of physical variations:

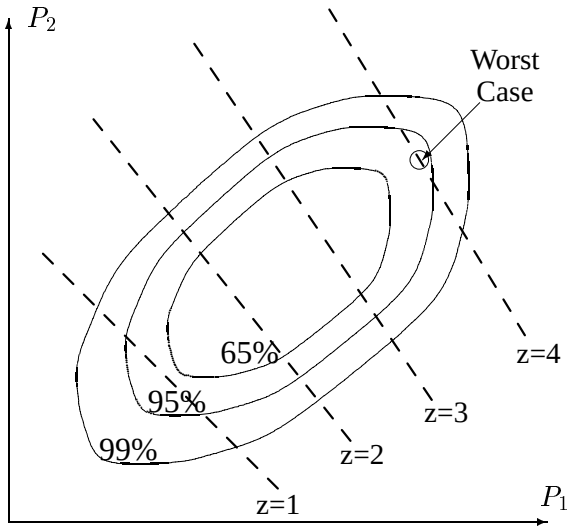

Figure 6: Worst Case Analysis.

- *Wafer level variations* (WLV): these are typically smooth variations across the wafer due to processing non-uniformities such as thermal gradients.
- *Die level variations* (DLV): these are caused by imperfections in the mask-making process, and by the interaction between the lithography and the local layout.
- *Wafer-Die Interaction* (WDI): these are corrections to account for the dependence of the die level component on the location of the chip within the wafer.
- *Random residuals*: are what remains after the first three components are characterized and is assumed to be normally distributed.

We can then express the spatial distribution of a physical parameter as:

$$P(x, y) = P_{WLV}(x, y) + P_{DLV}(x, y) + P_{WDI}(x, y) + N(0, \sigma^2) \tag{5}$$

Observe that within any randomly selected die, the wafer level variation component appears as a bias which is a function of the coordinates of the die within the wafer, and of higher level (e.g. lot-to-lot) variations. Since the designer has no explicit control over the location of the die in a wafer, this bias appears as a random number.

When the layout of a design is complete, it is sometimes possible to model the die-level variation components and thus treat them as a deterministic bias in performing subsequent variability analysis (see for example [20]). Often, however, either (a) the phenomena involved are not well

understood, or (b) the resources necessary to do the modeling are not available, or (c) there is a need to estimate the impact of within-die variability *before* the physical design is completed. In such a case must revert to a distributional approach where we characterize the phenomena statistically: i.e. the random residuals component grows in magnitude to accommodate whatever variations are not modeled. This is the analysis problem we wish to tackle, in which we express the physical parameters simply as:

$$P(x, y) = N(0, \sigma^2) \tag{6}$$

Examples and methodologies for the analysis of within-die variability in this form can be found in [4].

5 Conclusions

The impact of variability on circuit design is (a) increasing as we scale technology further, and (b) changing in character as the proportion of variations which is within-die increase and therefore force designers to deal with their interaction with detailed physical layout design. Near term trends point to an overall increase in total variations, in the portion of these variations that is within-die, and specifically variations in wire performance. Designers and EDA tool developers need to be cognizant of these trends and, where applicable, use them to guide develop improved analysis and design techniques.

References

[1] S. W. Director and W. Maly, editors. *Statistical Approach to VLSI*, volume 8 of *Advances in CAD for VLSI*. North-Holland, 1994.

[2] J. Power, B. Donnellan, A. Mathewson, and W. Lane. Relating statistical mosfet model parameter variabilities to ic manufacturing process fluctuations enabling realistic worst case design. *IEEE Trans. Semiconductor Manufacturing*, Aug 1994.

[3] A. Misaka, A. Goda, K. Matsuoka, H. Umimoto, and S. Odanaka. A statistical critical dimension control at cmos cell level. In *Proceedings of IEDM*, 1996.

[4] S. R. Nassif. Within-chip variability analysis. In *Proceedings of IEDM*, 1998.

[5] M. Eisele, J. Berthold, D. Schmitt-Landseidel, and R. Mahnkopf. The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits. *IEEE Trans. VLSI*, Dec 1997.

[6] *The National Technology Roadmap for Semiconductors*, 1997.

[7] C. Alpert, A. Devghan, and S. Quay. Buffer insertion with accurate models for gate and interconnect delay. In *Proceedings of DAC*, 1999.

[8] M. McKay, R. Beckman, and W. Conover. A comparison of three methods fos selecting values for input variables in the analysis of output from computer codes. *Technometrics*, May 1979.

[9] L. W. Nagel. *SPICE2: A Computer Program to Simulate Semiconductor Circuits*. PhD thesis, University of California, Berkeley, 1975.

[10] Y. Cheng, M. Chan, K. hui, M. Jeng, Z. Liu, J. Huang, K. Chen, J. Chen, R. Tu, P. Kp, and C. Hu. *BSIM3v3 Manual*. EECS Dept, Berkeley, 1996.

[11] C. Michael and M. Ismail. *Statistical Modeling for Computer-Aided Design of MOS VLSI Circuits*. Kluwer, 1993.

[12] A. Dharchoudhury et. al. Design and analysis of power distribution networks in powerpc^{TM} microprocessors. In *Proceedings of DAC*, 1998.

[13] M. K. Gowan, L. L. Biro, and D. B. Jackson. Power considerations in the design of the alpha 21264 microprocessor. In *Proceedings of DAC*, 1998.

[14] G. Steele et. al. Full-chip verification methods for dsm power distribution systems. In *Proceedings of DAC*, 1998.

[15] Y. M. Jiang and K. T. Cheng. Analysis of performance impact caused by power supply noise in deep submicron devices. In *Proceedings of DAC*, 1999.

[16] S. Nassif and J. Kozhata. Fast power grid simulation. In *Proceedings of DAC*, 2000.

[17] G. Digele, S. Lindenkreuz, and E. Kasper. Fully coupled dynamic electro-thermal simulation. *IEEE Trans. VLSI*, Sep 1997.

[18] V. Szekely, A. Poppe, A. Pahi, A. Csendes, G. Hajas, and M. Rencz. Electro-thermal and logi-thermal simulation of vlsi designs. *IEEE Trans. VLSI*, Sep 1997.

[19] B. E. Stine, D. S. Boning, and J. E. Chung. Analysis and decomposition of spatial variation in integrated circuit processes and devices. *IEEE Trans. Semiconductor Manufacturing*, Feb 1997.

[20] C. Yu, T. Maung, C. Spanos, D. Boning, J. Chung, H. Liu, K. Chang, and D. Bertelink. Use of short-loop electrical measurements for yield improvement. *IEEE Trans. Semiconductor Manufacturing*, May 1995.