Mapping Statistical Process Variations Toward Circuit Performance Variability: An Analytical Modeling Approach

Yu Cao
Department of EE, ASU
PO Box 875706
Tempe, AZ 85287-5706
480.965.1472
ycao@asu.edu

Lawrence T. Clark
Department of EE, ASU
PO Box 875706
Tempe, AZ 85287-5706
480.727.0295
Lawrence.Clark@asu.edu

ABSTRACT
A physical yet compact gate delay model is developed integrating short-channel effects and the Alpha-power law based timing model. This analytical approach accurately predicts both nominal delay and delay variability over a wide range of bias conditions, including sub-threshold. Excellent model scalability enables efficient mapping between process variations and delay variability at the circuit level. Based on this model, relative importance of physical effects on delay variability has been identified. While effective channel length variation is the leading source for variability at current 90nm node, performance variability is actually more sensitive to threshold variation at the sub-threshold region. Furthermore, this model is applied to investigate the limitation of low power design techniques in the presence of process variations, particularly dual V_{th} and L biasing. Due to excessive variability under low V_{DD}, these techniques become ineffective.

Categories and Subject Descriptors
B. 7.2 [**Integrated Circuits**]: Design Aids– *performance analysis and design aids*; B.8.2 [**Performance and Reliability**]: Performance Analysis and Design Aids.

General Terms
Performance, Design, Reliability, Experimentation.

Keywords
Process Variations, Delay, Variability.

1. INTRODUCTION

1.1 Process Variations in VLSI Design

The rapid scaling of CMOS technology has introduced drastic variations of process and design parameters, leading to severe variability of chip performance in the nanometer regime [1-3]. Among many process variation sources, the most important ones continue to be variations in effective channel length L and threshold voltage V_{th}, as a result of extreme difficulties in the precise control of lithography and channel doping [2]. Considering both inter-die and intra-die components, the variance of L and V_{th} can be more than 30% and 10%, respectively [3]. Other process effects, such as isolation oxide strain, transistor orientation, etch loading, etc., further spread out their values from the nominal points. Due to their stochastic nature or difficulties in process modeling, we intend to treat these process variations as statistical numbers with appropriate distributions. For instance, a popular distribution function is the Gaussian model. On the other hand, for the uncertainties in design parameters, such as supply voltage V_{DD} and temperature, they are usually treated as corners of the operation condition, rather than random parameters. In case supply voltage is scaled down for power reduction, increasing delay sensitivities are exposed and thus, the role of V_{DD} should also be correctly analyzed in variation-aware designs.

To handle the impact of process variations in VLSI design, corner based methodology is traditionally utilized to design to a slow timing corner so that full operational frequency can be met for all dies. As greater variability of circuit delay must be accounted for, the guard-band of design frequency (i.e., the difference between the target and post-fabrication performance) has to increase for both maximum delay constraining paths and minimum delay constraining (hold time) paths. Design resource may be wasted in this approach when process variations increase.
Recent efforts have been active in the area of statistical static timing analysis to overcome this barrier [4-7], which focus on the propagation of gate delay variability along paths and probabilistically solve timing margin. Although these approaches aim to precisely predict the guard-band in the presence of statistical variations, their accuracy strongly depends on the models of gate (and interconnect) timing variability that are employed as the basis. Without accurate and efficient models that map fluctuations in process parameters to timing variability at the circuit level, a statistical timer is not capable of correctly analyzing the distribution of path delay and thus, may not be better than corner based methodologies [8].

1.2 Circuit Performance Variability Analysis

Conventionally, gate delay variability can be studied through Monte-Carlo circuit simulations. Although the results are valuable, in practice such a methodology is often computationally prohibitive. An alternative solution is to build response surface models (RSM) that linearly expand circuit performance around nominal process values, which is usually combined with principal component analysis to decouple statistical variation sources. Yet for short-channel transistors, this approach is constrained due to complicated parameter correlations and high non-linearity,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
DAC 2005, June 13–17, 2005, Anaheim, California, USA.
Copyright 2005 ACM 1-59593-058-2/05/0006…$5.00.

particularly the V_{th} dependence on L [9]. Moreover, design insights for variation-aware circuit optimization are limited from either Monte-Carlo or RSM based approaches.

To accurately and efficiently predict delay variability at the circuit level, a desirable solution is an analytical performance model that is generic for digital circuits and directly links process parameters to performance metrics. Such a model should embody key physical phenomena in deep sub-micron technology, in order to achieve scalability of variability studies. Based on previous efforts, we observe two main reasons that support the feasibility of this idea:

1. It is possible to develop a generic performance model for a variety of digital circuits. A widely used example is the gate delay (T_d) model based on the Alpha-power law [10]:

$$T_d = \frac{CV_{DD}}{I} = \frac{KV_{DD}}{(V_{DD} - V_{th})^{\alpha}} \quad (1)$$

   Even though accurate predictions of variability are difficult due to its empirical nature, this model can quantitatively explain timing variation behavior during design optimization [11].

2. Physical understanding of short-channel effects is available in contemporary MOSFET models, such as the industrial standard BSIM models [12] or [13].

1.3 Scope of This Work

In this study, we will couple the Alpha-power law based timing model (Eq. (1)) with physical considerations in [12], particularly short-channel effects, to develop a physical yet compact performance model for predictions of circuit variability. This new model describes all regions of operation both above and below threshold), which provides an excellent basis for variation-aware design of both high-performance and low-power applications. Sec. 2 presents the derivation of this model. Under the assumption of separated normal distributions of L and V_{th} variations, Monte-Carlo simulations verify model predictions over a wide range of process and design conditions. Sec. 3 further elaborates the impact of variations and correlations in L and V_{th} on circuit delay variability. As both gate delay and delay variability can be obtained in closed-form from the model, it is very convenient to employ this approach for variation-aware design decisions. Application to dual-V_{th} and L biasing techniques are demonstrated in Sec. 4.

2. PERFORMANCE MODEL DERIVATION

Eq. (1) is simple and accurate for nominal timing calculation, yet includes insufficient modeling of short-channel effects leading to significant prediction errors of variability. In this section, our derivation begins from the examination and modeling of short-channel effects, improving model scalability and minimizing the parameter fitting effort. Then, these effects are incorporated into a unified output current model for performance calculation in both saturation and sub-threshold operating regimes.

2.1 Short-Channel Effects

Short-channel effects, such as V_{th} roll-off and velocity saturation, are critical for transistor behaviors in the deep sub-micron regime and need to be appropriately incorporated [12].

2.1.1 V_{th} Dependence on L and V_{DD}

As channel length L becomes shorter, V_{th} exhibits a greater dependence on L and drain bias (*DIBL*). Larger V_{DD} and smaller L usually lead to sharp degradation in V_{th} (i.e., V_{th} roll-off) and thus, shorter gate delay. Accurate modeling of V_{th} dependence on L and V_{DD} is important for accurate circuit analysis. Based on the physical derivations in BSIM [9], we simplify the model for V_{th} roll-off as:

$$V_{th} = V_{th0} - V_{DD} \cdot \exp(- a_{Vth} \cdot L) \quad (2)$$

where V_{th0} is the long-channel V_{th} and a_{Vth} is the DIBL coefficient. Both values can be extracted from transistor characteristics. Note that for some technologies where heavy Halo implantation is employed, V_{th} roll-up can also be apparent. In that case, another term in the order of $L^{-1/2}$ should be added to Eq. (2).

Experimental data shows that the sub-threshold swing (S) is also a function of L, sharing similar exponential dependence as the DIBL effect. To the first order, we can represent this as:

$$S = S_0 \cdot [1 + \exp(- a_S \cdot L)] \quad (3)$$

2.1.2 Velocity Saturation

In short-channel transistors operating in the saturation region, carrier velocity saturates resulting in sub-square dependence of output current on gate bias. Instead of the empirical exponent term in Eq. (1), a physical formula is introduced [9]:

$$I \propto (V_{DD} - V_{th})^{\alpha} \Rightarrow \frac{(V_{DD} - V_{th})^2}{1 + (V_{DD} - V_{th})/(E_{sat} L)} \quad (4)$$

where the value of E_{sat} can be estimated from model files.

2.2 A Unified Formula for Drive Current

For digital circuits, signal delay is proportional to $(C \cdot V_{DD})/I$, where I is the effective drive current during the switching period. Therefore, a physical and scalable formula for I is crucial for calculations of delay and delay variability. In the Alpha-power law based Eq. (1), I is proportional to $(V_{DD}-V_{th})^{\alpha}$, where α and V_{th} are empirically fitted parameters for saturation current. To correctly capture the dependence on L, V_{th}, and V_{DD} for short-channel transistors, we utilize Eq. (4) with V_{th} defined as Eq. (2).

To extend the model to the sub-threshold region, where I is an exponential function of V_{DD} and V_{th}, the following mathematical relation smoothly links I from both operation regions [12]:

$$\ln(1 + e^x) = \begin{cases} x, & \text{when } e^x >> 1 \\ e^x, & \text{when } e^x << 1 \end{cases} \quad (5)$$

By employing this expression to Eq. (4), a unified current formula is obtained:

$$I \propto \frac{\left\{ \ln\left[1 + \exp\left(\frac{V_{DD} - V_{th}}{2S}\right)\right]\right\}^2}{\left\{1 + \ln\left[1 + \exp\left(\frac{V_{DD} - V_{th}}{E_{sat} L}\right)\right]\right\}} \quad (6)$$

where S is the sub-threshold swing as shown in Eq. (3).

Table 1. Analytical Formulas for Circuit Performance

Nominal Gate Delay	
$$T_d = \frac{CV_{DD}}{I} = \frac{K \cdot V_{DD} \cdot \left\{1 + \ln\left[1 + \exp\left(\frac{V_{DD} - V_{th}}{E_{sat} L}\right)\right]\right\}}{\left\{\ln\left[1 + \exp\left(\frac{V_{DD} - V_{th}}{2S}\right)\right]\right\}^2}$$	
Gate Delay Variability	
$$\frac{\sigma_T}{T_d} = \sqrt{\left(\frac{\partial \ln T_d}{\partial L}\right)^2 \cdot \sigma_L{}^2 + \left(\frac{\partial \ln T_d}{\partial V_{th}}\right)^2 \cdot \sigma_{Vth}{}^2}$$	
DIBL:	$V_{th} = V_{th0} - V_{DD} \cdot \exp(-a_{Vth} \cdot L)$
Sub-threshold swing:	$S = S_0 \cdot [1 + \exp(-a_S \cdot L)]$
Loading:	$K = (k_0 + k_1 \cdot L \cdot C_{load} + k_2 \cdot L^{a_k})/W$

2.3 Delay and Variability Calculations

Based on the results above, a new gate delay model is developed. The complete performance model is summarized in Table 1. The $(V_{DD}-V_{th})^{\alpha}$ term in Eq. (1) is replaced with the unified formula for driving current (Eq. (6)). The parameter K, which is expressed as a polynomial function of L and loading capacitance, represents the dependence of gate delay on loading and is normalized to (W/L). For designs solely in one region or the other, the delay model can be simplified as:

$$T_d \propto \begin{cases} \frac{V_{DD} \cdot [1 + (V_{DD} - V_{th})/(E_{sat} L)]}{(V_{DD} - V_{th})^2}, & \text{at saturation} \\ \frac{V_{DD}}{\exp\left(\frac{V_{DD} - V_{th}}{S}\right)}, & \text{at sub-threshold} \end{cases} \quad (7)$$

In this new model, the values of V_{th0} and L are specified by the process, rather than fitting parameters. Furthermore, the other parameters can be efficiently extracted as: first, coefficients in the models of V_{th} and S are obtained from transistor behaviors; then for a specific circuit, coefficients in K are extracted from simulation data under a variety of V_{DD}, L, and C_{load}. Minimum fitting is needed to improve the accuracy, while strong physical meanings are preserved in the formula. Thus, such an analytical model is capable of accurate predictions of delay variability, without resorting to Monte-Carlo circuit simulations.

In fact, if the statistical variations in L and V_{th} follow the Gaussian distribution and transistors within a small range of gate size are strongly correlated, we can directly solve the variability of gate delay, which is defined as (σ/μ) of T_d, in closed-form from the timing model in Table 1:

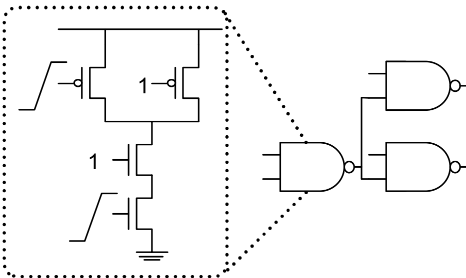

Figure 1. A test circuit of a NAND gate (FO=2).

Table 2. Technology Specifications at 90 nm Node.

	Effective *L* (nm)	V_{th0} (V)	V_{DD} (V)
Nominal (μ)	55	0.35	1.0
Variation (σ)	5.5	0.01	--

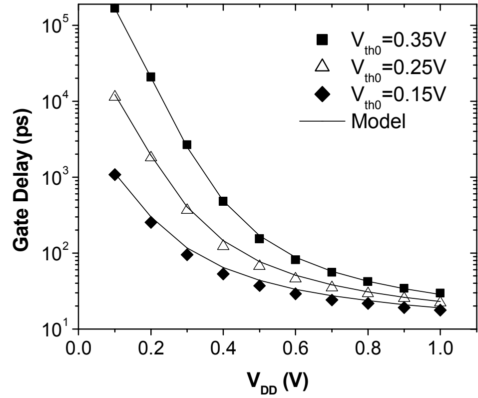

Figure 2. The dependence of NAND gate delay on V_{DD} and V_{th}.

$$\frac{\sigma_T}{T_d} = \sqrt{\left(\frac{\partial \ln T_d}{\partial L}\right)^2 \cdot \sigma_L{}^2 + \left(\frac{\partial \ln T_d}{\partial V_{th}}\right)^2 \cdot \sigma_{Vth}{}^2} \quad (8)$$

For a larger circuit, in which spatial dependence of correlation (ρ) is not negligible, we can first partition it into smaller grids (where $\rho \approx 1$), apply Eq. (8) to each grid to calculate the variability, and then sum them together with appropriate correlation function to obtain the total variability [7].

2.4 Model Verification

To verify the accuracy of this new analytical model, we use a Fanout=2 static NAND gate as the benchmark circuit. The circuit structure is illustrated in Fig. 1. An industrial 90 nm technology is employed for this study with process parameters adjusted according to technology projections in [1, 3]. The specifications are listed in Table 2. Strong correlation among transistors is assumed for such a small circuit. Model parameters are extracted based on generic formulas in Table 1.

Fig. 2 presents the verification of nominal gate delay (T_d) as a function of V_{DD} and V_{th0}. The unified model accurately predicts these dependences both above and below threshold, as delay changes more than four orders of magnitude. Fig. 3 further demonstrates the excellent model scalability to a wide range of L and V_{DD}. The accurate results prove that the physical interactions among major L, V_{th}, and V_{DD} are sufficiently and correctly modeled. Therefore, it is plausible that this analytical approach is also appropriate for studies of delay variability.

This is confirmed in Fig. 4, which compares predictions of delay variability with Monte-Carlo circuit simulations, assuming the variations in L and V_{th} are normally distributed. Note that model results are directly solved from a closed form, which is much more

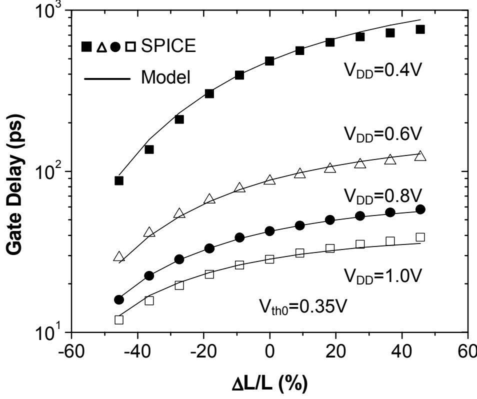

Figure 3. The dependence of nominal delay on *L*.

efficient than Monte-Carlo simulations and also suitable for early stage design exploration and tool implementation. Due to the non-linear response of gate delay to V_{DD} and V_{th}, delay variability increases dramatically at lower power supply. This trend was quantitatively explained by Eq. (1) in [11], while our model precisely predicts the behavior. Fig. 4 also shows that variation in L is the dominant source for delay variability, through both direct impact and V_{th} (i.e., DIBL).

3. IMPACT OF VARIATIONS ON DELAY VARIABILITY

The new analytical approach not only eases the calculation of variability, but also generates physical insights into the variation effects. Using the NAND gate as a benchmark circuit, we illustrate these comprehensions in this section.

3.1 Decomposition of Physical Mechanisms

As Fig. 4 reveals the relative importance of different process variation sources, it is worthwhile to further decompose variability to causal physical mechanisms, as shown in Fig. 5. Based on the assumptions of variations in Table 2, it is observed that DIBL and fluctuations in V_{th0} are the primary mechanisms at low V_{DD}, since the drive current has a stronger dependence on V_{th} under that condition. On the other hand, in the above threshold region, the K factor, which includes the dependencies on both sizing (W/L) and loading capacitance, is the dominant source. These variations are contributed by L and they are relatively independent of V_{DD}. Other

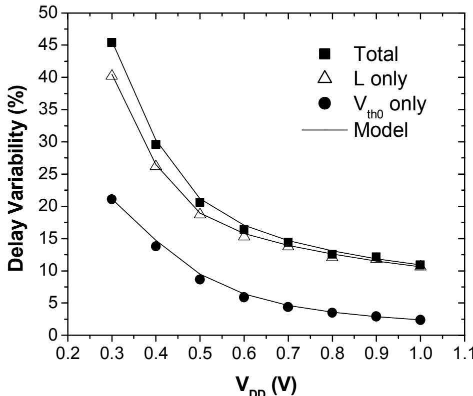

Figure 4. Model accurately predicts variability of delay.

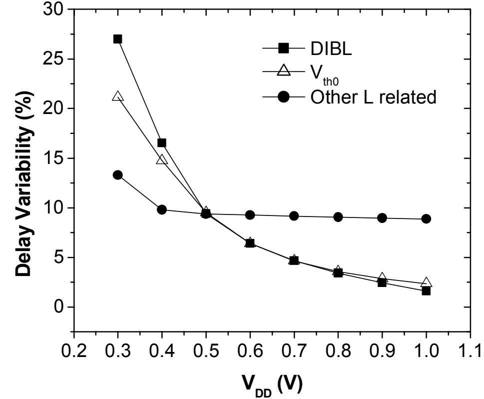

Figure 5. Contributions of individual physical mechanisms to delay variability.

mechanisms, including velocity saturation and the effect of sub-threshold swing (S), are second order to the above ones. Their contributions in this case are about 1-2%, although in the sub-threshold region, variation in S can increase to about 5%.

3.2 Sensitivity of Delay Variability

The timing variability (i.e., σ_T/T_d) of the NAND gate is shown in Fig. 4 for both V_{th} and L. If we further normalize the process variations (σ) by the mean values of V_{th} or L, respectively, we can identify the importance of process control (σ/μ) on variability, as shown in Fig. 6. Between V_{th} and L, their contributions are nearly equal at V_{DD}=0.7V. Actually, the normalized V_{th} change, directly affecting the gate overdrive and particularly the sub-threshold current, is more important than that of L at lower voltages. In current 90nm technology, L is still the dominant source of variability due to its larger amount of absolute variation (Table 2).

3.3 Spatial Correlations

For a realistic critical path, the situation is more complicated as the correlation among gates depends on their separation distance [7]. Different paths may exhibit strong or weak spatial correlation. For instance, ALU's and multipliers, where they fully comprise a critical path, exhibit strong spatial correlation due to their small size. Cross-chip paths, such as those spanning memory blocks or long data-paths, are weaker in correlation. Fig. 7 shows the measurement data of intra-die spatial correlation of L for a 130nm process, in which the spatial correlation can be modeled as a linear

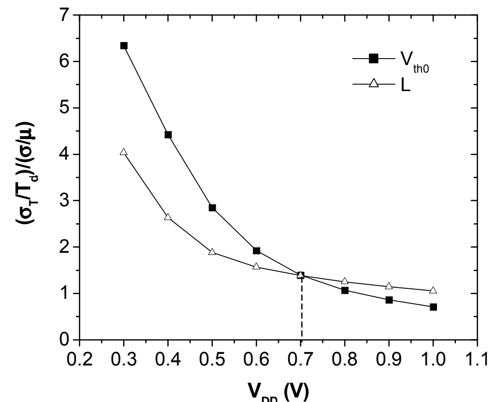

Figure 6. Sensitivity of delay variability to normalized *L* and V_{th} variations.

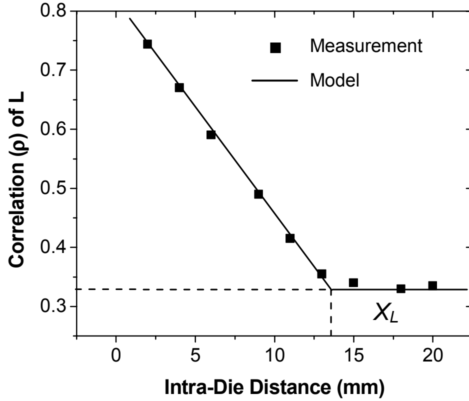

Figure 7. Model for intra-die correlation of *L*.

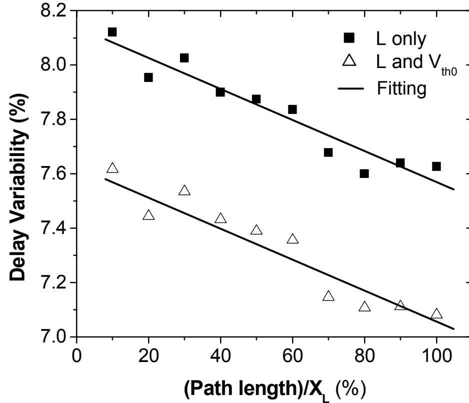

Figure 8. Impact of spatial correlations on variability.

function of the distance [14]. As expected, the highest correlation occurs at the shortest distances while the poorest correlation occurs over long distances. The correlation does not peak at 1, however, since the measured devices still exhibit purely random variations even when very close. Consequently, we can assume a background random component affecting even the best spatially correlated devices, i.e., those that are adjacent. While high correlation results in improved matching for devices that are in spatial proximity, it does not improve delay variability for the overall path. In fact, high correlation does not allow cancellation of variability, as is shown in Fig. 8. Here, the correlation was varied from 0.8 to 0.33 for a 10 gate path (Fig. 7), with the relative distance spanned by the path running from 10% to 100% of the correlation distance X_L. The V_{DD}=1.0V. It should be noted that for smaller die all paths will be within the correlated distances, i.e., the apparent systematic variation becomes larger. The addition of V_{th} variation affects the delay variability primarily by shifting the curve to a larger value, as a result of the weak correlation among V_{th} [15]. Although a weak correlation favors the reduction of path delay variability, it is noted that, on the other hand, a weak correlation worsens the mismatch in the designs of memory cells and the clock network.

4. DESIGN APPLICATIONS

Excessive amount of process variations pose significant limitations on the achievable design space. Thus, the awareness of variability is critical in contemporary VLSI design, in conjunction with timing and power. In this section, we apply the variability model to two popular low power design techniques, dual V_{th} and L biasing, and illustrate the consequence of process variations.

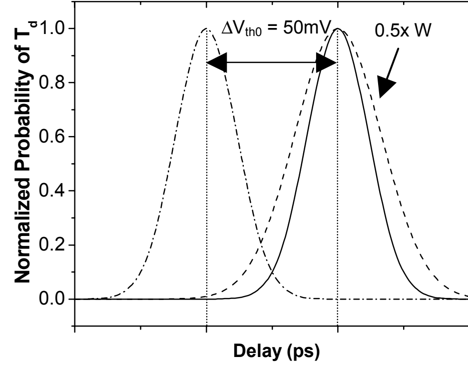

Figure 9. Overlap in gate delay due to variations.

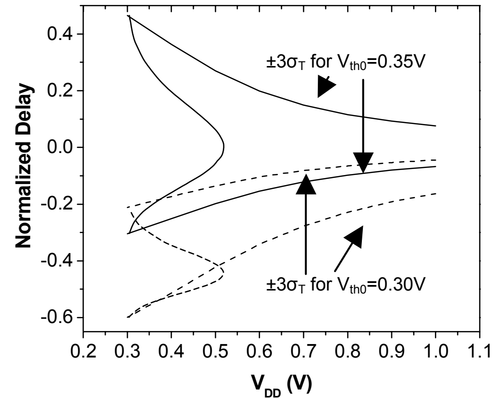

Figure 10. Delay overlap worsens under lower V_{DD} (normalized to delay at high V_{th}).

4.1 Dual V_{th} Assignment

To reduce power consumption, we can apply higher V_{th} on non-critical paths, while low V_{th} transistors are used on critical paths for speed considerations. Such a dual V_{th} technique can be realized by tuning channel doping or via body bias control [16]. In both cases, high and low V_{th} are tuned independently and thus, they are rarely correlated. The use of dual V_{th} produces extra process corners and exaggerates variability. Specifically, if the high and low V_{th} are not sufficiently far apart, some high V_{th} transistors may be faster than some low V_{th} transistors in the presence of variations. Consequently, the insertion is unsuccessful at changing the delay and becomes ineffective to reduce power consumption, especially the leakage. This overlap is shown in Fig. 9, with dual V_{th} at 0.35V and 0.30V assigned to the path of NAND gates. Considering an increase in V_{th} variation if narrower width is employed at non-critical paths [16], the potential gate delay overlap becomes even worse.

The dependence of this overlap at different V_{DD} is investigated in Fig. 10, based on the specifications in Table 2. Without the inclusion of L variation, there is less than a 1% likelihood of delay overlap. If a 5% intra-die L variation is included (i.e., 1/3 of the total L variation [3]), the likelihood that an individual insertion is unsuccessful rises to 5.6% at V_{DD}=0.3V. The possibility of path delay overlapping will increase if there are additional random variations contributed by inter-die components.

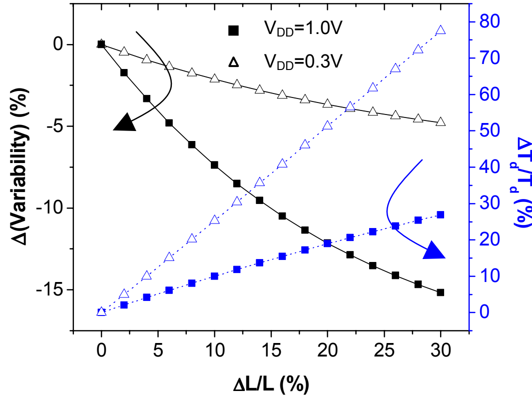

Figure 11. Efficiency of *L* biasing at different V_{DD}.

4.2 *L* Biasing

Similarly to dual V_{th} assignment, dual channel lengths can also be used for power efficiency. Delay variability naturally reduces with increased channel length, since the absolute magnitude of the L variation is almost unaffected and thus, (σ/μ) of L decreases. In contrast, delay variability increases with higher V_{th} due to the non-linear response of Eq. (7). Considering the tradeoffs between delay penalty and variability reduction, increase of L by 15% is beneficial at V_{DD}=1.0V, as shown in Fig. 11: at that point, delay only increases by less than 20% while delay variability is reduced by 15%. At V_{DD}=0.3V, however, biasing L is inefficient due to severe delay penalty.

Moreover, similar as the dual V_{th} technique, the concern of delay overlap also limits the application of L biasing at low V_{DD}. Since delay variability increases substantially in sub-threshold, a different L choice is nearly impractical at V_{DD} =0.3V, as the variations will produce a 64% likelihood of failure insertion (as illustrated in Fig. 12).

5. SUMMARY

We have presented a physically based, analytical delay variability model that is appropriate for early design exploration and simulation. The model includes short channel effects (e.g., DIBL and velocity saturation), and is accurate over a wide V_{DD} operating range, including sub-threshold. The model derivation and parameter extraction have been described. We have shown the relative importance of V_{th} and L variation on the overall digital circuit delay variability and explained the underlying physical phenomena based on the model. Design applications include early prediction of circuit level delay variability effects including the efficacy of dual V_{th} and dual L design and spatial correlation of L. Specifically, while both methods are effective at high V_{DD}, they are shown to be ineffective at very low V_{DD} due to the excessive path delay variability.

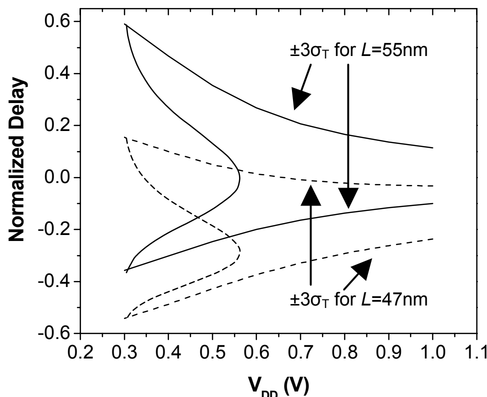

Figure 12. *L* biasing is not practical due to severe variations in both *L* and V_{th0}.

REFERENCES

[1] International Technology Roadmap for Semiconductors, 2003.

[2] K. A. Bowman, S. G. Duvall, J. D. Meindl, "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *JSSC*, vol. 37, no. 2, pp. 183-190, Feb. 2002.

[3] D. Boning and S. Nassif, "Models of process variations in device and interconnect," *Design of High-Performance Microprocessor Circuits*, Chapter 6, pp. 98-115, IEEE Press, 2000.

[4] C. Viswesvariah, "Statistical timing of digital integrated circuits," *ISSCC*, 2004.

[5] M. Orshansky and K. Keutzer, "A general probabilistic framework for worst-case timing analysis," *DAC*, pp. 556-561, 2002.

[6] H. Chang, S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," *ICCAD*, pp. 621-625, 2003.

[7] A. Agarwal, D. Blaauw, V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations*," ICCAD*, pp. 900-907, Nov. 2003.

[8] S. Nassif, N. Hakim, D. Boning, "The care and feeding of your statistical static timer," embedded tutorial, *ICCAD*, Nov. 2004.

[9] M. Orshansky, J. C. Chen, C. Hu, "Direct sampling methodology for statistical analysis of scaled CMOS technologies," *TSM*, vol. 12, no. 4, pp. 403-408, Nov. 1999.

[10] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *JSSC*, vol. 25, no. 2, pp. 584-594, Apr. 1990.

[11] M. Eisele et al, "The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits," *TVLSI*, vo. 5, no. 4, pp. 360-368, Dec. 1997

[12] BSIM4.2.1 MOSFET Model – User's Manual, 2001.

[13] K. A. Bowman et al., "A physical Alpha-power law MOSFET model," *ISLPED*, pp. 218-222, Aug. 1999.

[14] P. Friedberg et al., "Modeling within-die spatial correlation effects for process-design co-optimization," *ISQED*, Mar. 2005.

[15] P. A. Stolk, F. P. Widdershoven, and D. B. M. Klaassen, "Modeling statistical dopant fluctuations in MOS transistors," *TED*, vo. 45, no. 9, pp. 1960-1971, Sep. 1998.

[16] J. Tschanz et al., "Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage," *JSSC*, vol. 1, pp. 422-478, Feb. 2002.