# Minimization of Delay Sensitivity to Process Induced Voltage Threshold Variations*

Georges Nabaa Farid N. Najm

Department of Electrical and Computer Engineering
University of Toronto, Toronto, Ontario, Canada
{georges, najm}@eecg.toronto.edu

## ABSTRACT

Threshold voltage variations, resulting from underlying process variations, cause variations in circuit delay that can affect the chip timing yield. We study design techniques and optimization strategies that minimize the effects of threshold voltage variations on circuit delay variability. Specifically, we compare different static circuits (classic CMOS, ratioed logic, and transmission gate logic) and dynamic circuits and evaluate their limitations and benefits in terms of delay variablilty, performance penalty and area overhead. Based on our findings, we also introduce circuit design guidelines and techniques that help mitigate the effects of threshold voltage variations. By reducing delay variability on a per-gate basis, we show how one can build threshold voltage variations-aware gate libraries for use in deep submicron design.

## 1. INTRODUCTION

For more than 30 years, scaling of CMOS technology has resulted in an increase of the overall circuit operating speed and hence overall chip performance. This technology scaling has required a reduction of the threshold voltage ($V_{th}$) and the supply voltage ($V_{dd}$). Supply voltage reduction guarantees an acceptable dynamic power dissipation at the expense of overall performance. In turn, this loss is recovered by a corresponding reduction in $V_{th}$ which restores the current drive of the transistor. However, $V_{th}$ scaling has not been accompanied by a similar decrease in threshold voltage *variations* ($\Delta V_{th}$) at the process level. This is primarily due to random dopant fluctuations which become harder to control in finer geometry processes. As we move into the sub-90 nm technology, these variations are expected to cause significant variation in delay which in turn impacts chip timing yield. Overall, process variations can cause up to 30% variation in chip frequency [1].

While process variations have been an active research topic for decades, the prominence of threshold voltage variations is a fairly recent problem. It is a problem that plagues the memory community to a large extent, but has recently become an issue for logic circuits as well. Process variations (and therefore, $V_{th}$ variations) have a die-to-die and a within-die component. The portion of the variation which impacts every transistor on the die equally is said to be a die-to-die component. One can compensate for the die-to-die component of the $V_{th}$ variations by a die-wide compensation scheme in which body voltage, for instance, is changed to maintain chip speed in spite of the $V_{th}$ variations, such as in [2]. Such chip-wide schemes, however, cannot compensate for the within-die component of the variations. In this work, we are focused on the within-die component of the $V_{th}$ variations. Standard compensation schemes, such as by body bias, cannot be applied to individual gates - a separate compensation circuit for every gate would have a ridiculously high area overhead. Instead, circuits must be made *resilient* to local $V_{th}$ variations by some other means. Our approach is to introduce minimal changes to the circuit design style, and design guidelines, in order to achieve $V_{th}$ variations-aware designs. We will start with a study of the extent of sensitivity of the delay to $V_{th}$ variations, and then proceed to the design guidelines.

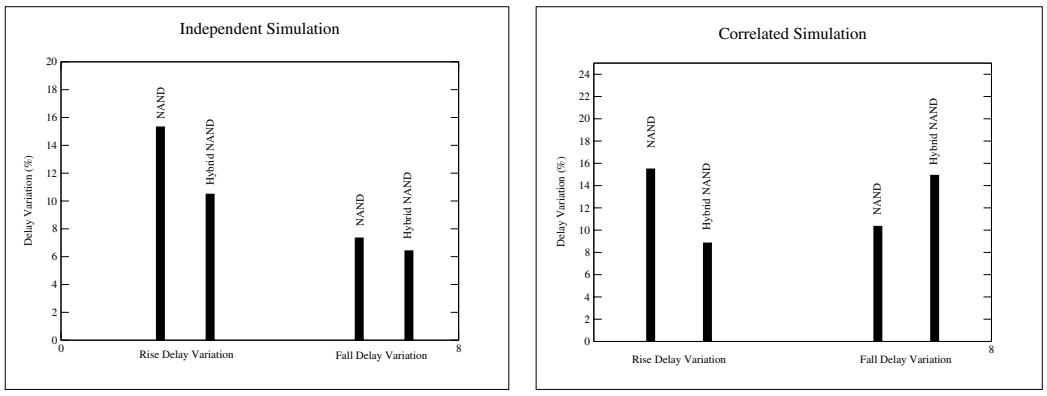
We use a standard Monte-Carlo approach where the threshold voltages of all transistors in a given gate are varied independently according to a normal distribution; correlated normal analysis is also applied to investigate certain corner cases. These statistical approaches are preferred to an overly pessimistic corner-case analysis. All the tests are conducted on the $0.13\,\mu$m platform; this technology does not exhibit nearly as much $\Delta V_{th}$ sensitivity as is projected for future sub-90 nm generations [1]; nevertheless, it constitutes a good starting point for the study.

Using this statistical framework, we explore the variability of the major design families. Classic CMOS tests allow us to compare series and parallel stacks. These results lead us to develop hybrid NAND and NOR gates that add a dummy transistor in their parallel network circuitry. Next, we consider transmission gate circuits and expose some of the properties that make them extremely robust when considering delay sensitivity under $V_{th}$ variations. We also evaluate two dynamic circuits styles: "footerless" and standard domino logic design are benchmarked and compared. Finally we demonstrate how delay sensitivity varies with sizing and gate fan-in.

## 2. STATISTICAL ANALYSIS

We model the $V_{th}$ variations as normally distributed random variables (RVs) and, for each gate, we perform two types of analyses: independent and correlated. In the independent case, the $V_{th}$ RVs within the gate are assumed to be independent. In the correlated case, they are assumed to be totally *positively* correlated, i.e., when one is high the others are also all high. Here, we are referring to the algebraic value of $V_{th}$, not its magnitude. Thus, the correlated case means that $|V_{th}|$ of the n-channel devices is assumed

*This work was supported in part by the Semiconductor Research Corporation (SRC 2001-HJ-901) and by the Canadian Microelectronics Corporation(CMC).




(a) (b)

**Figure 4: 2-input NAND vs Hybrid NAND delay sensitivity**

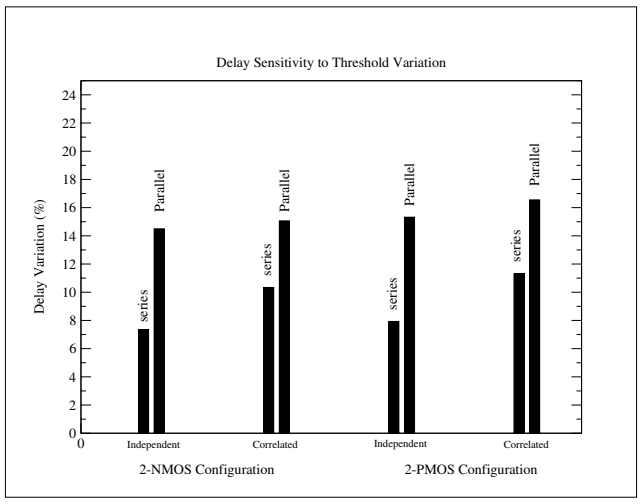



**Figure 1: Series vs Parallel Configuration**

high when $|V_{th}|$ for the p-channel devices is assumed low, and vice-versa. The $3\sigma$ limits of the normal are determined by the worst case DELVTO values from the technology file and scaled according to the law of area [3, 4] (namely: the magnitude of $V_{th}$ variations is inversely proportional to the square root of transistor area). DELVTO is a user-specified HSPICE parameter that shifts the value of the threshold voltage during the simulations. In order to avoid biasing our estimators, we use 1000 sweep points between the plus and minus $3\sigma$ limits of the normal. With the $\Delta V_{th}$ in place, the gates are simulated and the **worst case** rise and fall times are recorded. We use straight-forward averaging to determine the mean rise and fall times:

$$\mu = \sum_{k=1}^{n} \tau_i \tag{1}$$

and we compute the variance using the standard estimator:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - \mu)^2 \tag{2}$$

Using the mean and variance, we compute the following figure of merit, which we refer to as the *delay sensitivity*:

$$\Delta_{\text{rise}} \text{ or } \Delta_{\text{fall}} = \frac{3\sigma}{\mu} \tag{3}$$

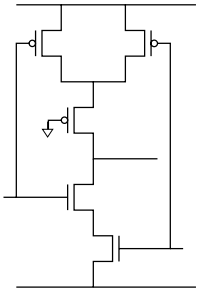


**Figure 2: Hybrid Nand (configuration 1)**

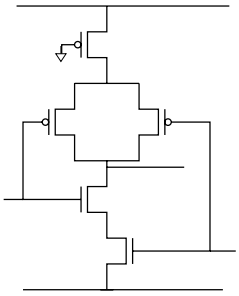


**Figure 3: Hybrid Nand (configuration 2)**

# 3. STATIC GATES

For static CMOS, we considered two different styles and studied the delay sensitivity in each, leading to design guidelines.

## 3.1 Standard CMOS

For standard fully-complementary static CMOS gates, we considered the basic gate types, NAND and NOR. For two-input gates, each simulation consisted of 5 cases: 4 cases where each transistor in the gate is subjected to a $\Delta V_{th}$ in a stand-alone fashion (with all the rest set at nominal), and

**Table 1: Summary of Delay Sensitivity to Independent Threshold Voltage Variations in Major Design Families**

| | NAND | NOR | T-GATE AND | DYNAMIC NAND Standard | DYNAMIC NAND Footerless |
|---|---|---|---|---|---|
| $\Delta_{Rise}$ | 15.32% | 7.93% | 6.76% | N/A | N/A |
| $\Delta_{Fall}$ | 7.34% | 14.50% | 7.95% | 8.35% | 7.46% |
| Rise Delay(ns) | 34.35 | 37.96 | 35.34 | N/A | N/A |
| Fall Delay(ns) | 32.25 | 38.52 | 24.47 | 28.78 | 25.60 |
| Area ($\mu m$) | 1.88 | 2.26 | 1.38 | 2.69 | 1.44× |

**Table 2: Summary of Delay Sensitivity to Positively Correlated Threshold Voltage Variations in Major Design Families**

| | NAND | NOR | T-GATE AND | DYNAMIC NAND Standard | DYNAMIC NAND Footerless |
|---|---|---|---|---|---|
| $\Delta_{Rise}$ | 15.49% | 11.31% | 0.3% | N/A | N/A |
| $\Delta_{Fall}$ | 10.34% | 15.06% | 1.14% | 11.00% | 10.08% |
| Rise Delay(ns) | 34.47 | 38.33 | 35.32 | N/A | N/A |
| Fall Delay(ns) | 32.23 | 38.00 | 24.45 | 31.29 | 25.57 |
| Area ($\mu m$) | 1.88 | 2.26 | 1.38 | 2.69 | 1.44 |

the last case where all the transistors in the gate received the variation simultaneously and independently. Fig. 1 clearly shows that series stacks exhibit less sensitivity than their parallel counterparts. We exploit these findings by introducing a hybrid design style that adds a dummy transistor to the parallel circuitry of standard CMOS NAND and NOR gates.

Two hyrbid gate configurations are proposed (shown for a 2-input NAND gate in Figs. 2 and 3). The difference between these configurations is in the position of the "serializing" dummy transistor. Simulations show that while both configurations have similar delay variabilities, the gate in Fig. 3 exhibits 17% smaller absolute rise delays than the gate in Fig. 2. This is due to the fact that under worst case rise conditions, the internal capacitance of the added dummy transistor is precharged in Configuration 2. In the remainder of this paper, hybrid style refers to configuration 2 for the NAND and its corresponding counterpart setup for the NOR.

Fig. 4(a) demonstrates that these hybrid gates exhibit less sensitivity in their rise times (NAND) and fall times (NOR). However, this comes at some expense of larger delays which can be recovered by a corresponding increase in area and power dissipation. In order for the hybrid NAND to match the nominal delay performance it's area must be 2.1 times the area of a standard NAND gate . This overhead is reduced as the number of inputs increases; correlated simulations show that for a three input hybrid NAND, the area overhead required to match the nominal performance of a standard 3 input NAND gate is 1.5×. The delay variability savings remain higher than in the standard NAND even when a similar scaling is applied (13.68% vs 11% in rise delay variability). In order to minimize this area overhead, we applied a dual-Vt solution to the hybrid-gate by making its dummy transistor a low-vt device. The area overhead required to match the rise and fall times of a standard two-input NAND dropped to 78%. Hence, this hybrid gate with a single low-vt transistor can be a viable alternative to an overly leaky standard NAND gate with all low-vt devices when performance is critical.

Simulations where all transistors in a given gate are subjected to correlated normal $\Delta V_{th}$ variations confirmed the series stack's lower sensitivity compared to its parallel counterparts (Fig. 1). It also validated our hybrid designs with rise delay variations dropping from 15.49% to 12.2% for the hybrid NAND (Fig. 4(b)).

## 3.2 Transmission Gates

Transmission gates display the best delay variability robustness while keeping very small absolute delays under independent and correlated $\Delta V_{th}$ variations. Table 1 illustrates the better delay sensitivity of transmission gates compared to the other design styles. Rise and fall variations are nearly halfed while absolute speed remains acceptable. Column 4 in Table 2 shows that, under correlated variations, delay sensitivity approaches zero for our transmission gate based AND. This result can be explained through the intrinsic structure of a transmission gate: NMOS and PMOS have opposite $V_{th}$ values. Hence, when subjected to a similar $\Delta V_{th}$, the contribution (faster or slower) that results from say the NMOS device is counterbalanced by an equal contribution from the PMOS device and vice versa.

# 4. DYNAMIC GATES

The domino logic NAND gate performed slightly worse than its traditional classic CMOS counterpart. These results can be explained by the fact that dynamic logic adds an additional footer transistor to the pull-down network. This can be rectified by designing footerless domino-logic circuits. By using footerless logic, we were able to reduce absolute delay by 22.36% while achieving the same delay variability. The last 2 columns of Tables 1 and 2 compare the performance of standard and footerless logic in both independent and correlated conditions. Hence from a design perspective footerless dynamic logic performs better than its standard counterpart. Finally, when looking again at Tables 1 and 2, we clearly notice that a footerless dynamic NAND exhibits the same variability as its static Classic CMOS counterpart but offers lower absolute delays (26% faster).

**Table 3: Effect of sizing on Fall Delay variability**

| Size Factor | NAND | | NOR | | DYNAMIC NAND Standard | | DYNAMIC NAND Footerless | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta_{Fall}$ | Fall (ns) | $\Delta_{Fall}$ | Fall (ns) | $\Delta_{Fall}$ | Fall (ns) | $\Delta_{Fall}$ | Fall (ns) |
| x1 | 7.34% | 32.25 | 14.50% | 38.52 | 7.15% | 31.32 | 7.46% | 25.60 |
| x1.5 | 6.33% | 31.31 | 12.86% | 37.33 | 7.06% | 28.98 | 5.73% | 23.15 |
| x2 | 5.95% | 31.68 | 11.60% | 37.44 | 6.78% | 28.65 | 4.77% | 23.20 |
| x2.5 | 6.15% | 32.40 | 11.30% | 38.27 | 7.46% | 28.73 | 4.75% | 23.87 |
| x3 | 5.87% | 33.35 | 10.39% | 38.58 | 7.71% | 28.72 | 4.05% | 24.66 |

**Table 4: Effect of sizing on Rise Delay variability**

| Size Factor | NAND | | NOR | |
|---|---|---|---|---|
| | $\Delta_{Rise}$ | Rise (ns) | $\Delta_{Rise}$ | Rise (ns) |
| x1 | 15.32% | 34.35 | 7.93% | 37.09 |
| x1.5 | 12.51% | 31.61 | 6.68% | 36.02 |
| x2 | 10.84% | 30.77 | 5.92% | 35.87 |
| x2.5 | 9.70% | 30.75 | 5.83% | 36.23 |
| x3 | 9.51% | 31.06 | 5.64% | 37.01 |

# 5. SIZING AND GATE FAN-IN

Using a similar setup, we increase the width of the transistors while respecting the pull-up to pull-down ratio. We perform the tests on the NAND, NOR, dynamic-NAND, and dynamic-footerless-NAND. The results in Tables 3 and 4 show that while there is no significant change in absolute delay, delay variability is inversely proportional to the width of the transistors. Moreover gates with a high fan-in exhibit an expected deterioration in absolute delay but offer a lower delay sensitivity.

# 6. RESULTS

Tables 1 and 2 show in detail the delay sensitivity of the major logic design families to threshold voltage variations. In each case, speed and area overhead are depicted. Fig. 1 clearly demonstrates the stronger delay sensitivity of parallel circuits to $\Delta V_{th}$ over their series counterpart. The trend is shown for both 2-NMOS and 2-PMOS configurations when subjected to independent and correlated input variations. Fig. 3 depicts the topology of our improved NAND gate. The pull up network is "serialized" using a dummy transistor. Figs. 4(a) and 4(b) demonstrate the robustness of our hybrid NAND against its classic counterpart for both independent and correlated simulations. The "serialized" pull up network yields improved rise variation delay, while the fall variations remain mainly unchanged.

# 7. CONCLUSION

In this paper, we studied the delay sensitivity characteristics of the major logic design families with respect to normal independent and correlated $\Delta V_{th}$. Based on our findings, we introduced novel hybrid NAND and NOR gates. We also showed that delay sensitivity is inversely proportional to sizing and gate fan-in. All these results allowed us to express high-level design guidelines for $\Delta V_{th}$ aware libraries.

# 8. REFERENCES


[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De. Parameter variations and impact on circuits and microarchitecture. In *ACM/IEEE 40th Design Automation Conference (DAC-03)*, pages 338–342, Anaheim, CA, June 2-6 2003.

[2] C. H. Kim, K. Roy, S. Hsu, A. Alvandpour, R. K. Krishna murthy, and S. Borkar. A process variation compensating technique for sub-90nm dynamic circuits. In *VLSI Symposium*, Japan, June 2003.

[3] U. Grunebaum, J. Oehm, and K. Schumacher. Mismatch modeling and simulation- a comprehensive approach. In *Analog integrated circuits and signal processing*, pages 165–171, December 2001.

[4] S.J. Lovett, L. Wall, and M. Welton et al. Sensitivity of MOS transistor mismatch to device dimensions and suggestions on how to improve matching performance. In *Improving the efficiency of IC manufacturing, IEEE Colloquium on*, pages 11/1–11/5, London, UK, April 12 1995.