

# Influence of Leakage Reduction Techniques on Delay/Leakage Uncertainty

Yuh-Fang Tsai, N. Vijaykrishnan, Yuan Xie, and Mary Jane Irwin  
Dept. of Computer Science and Engineering  
Penn State University, University Park PA  
{ ytsai, vijay, yuanxie, mji}@cse.psu.edu

## Abstract

One of the main challenges for design in the presence of process variations is to cope with the uncertainties in delay and leakage power. In this paper, the influence of leakage reduction techniques on delay/leakage uncertainty is examined through Monte-Carlo analysis. The techniques investigated in this paper include increasing gate length, stack forcing, body biasing, and  $V_{dd}/V_{th}$  optimization. The impact of technology scaling and temperature sensitivity on the uncertainty reduction are also evaluated. We investigate the uncertainty-power-delay trade-off and suggest techniques for designs targeting different requirements.

## 1. Introduction

With the continuing technology scaling, the need for high performance and low power design has driven the scaling of supply voltage and threshold voltage. As a result, leakage power has become one of the major concerns for current and future technologies. Many leakage reduction techniques have been proposed and their effectiveness and scaling trend have been evaluated [1][2]. Another challenge for technology scaling is to cope with the uncertainties in the presence of increasing process variations. The variations (such as transistor channel length and transistor threshold voltage) as a percentage of their nominal values increase when technology advances. In [3], it is shown that the leakage current can vary from the target leakage current by 6.5x when considering process variations. The measurement of chips in 0.18um technology shows that 30mV variation in threshold voltage can result in 20x difference in leakage power and 30% variation in frequency [4]. The above mentioned works prove that the performance and power consumption of actual silicon may significantly deviate from the targeted design specifications due to variations. After manufacturing, for a chip to be accepted, it must meet a minimum frequency, and at the same time, the accepted die must meet the maximum power consumption requirement. Any die that exceeds the maximum leakage power must either be binned to operate at lower frequency or discarded. The delay/leakage uncertainties due to process variations can worsen the binning distribution and cause yield loss and reliability issues. Moreover, the rising leakage and leakage uncertainty have made the  $I_{ddq}$  test a challenge [5].

Our Hspice simulation (1000 runs of Monte Carlo analysis) results shown in Fig.1 predict the increasing uncertainty in leakage while the leakage grows as technology advances. In Fig. 1, the random distribution of leakage is quantified by the parameter “uncertainty” which we define in this paper as the standard deviation of delay (leakage) divided by its mean value (S.D./mean). There is a need to control the uncertainty.

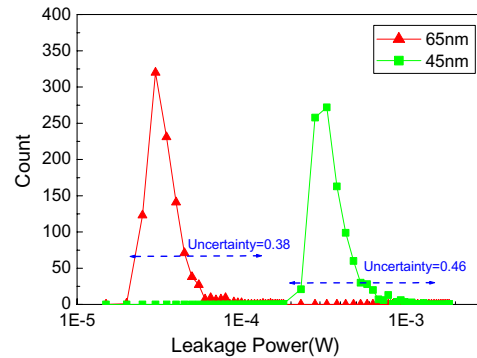


Figure 1. Leakage power distribution for 65nm and 45nm technologies.

Many active leakage power reduction techniques are based on adjusting transistor physical parameters or reducing the effect caused by these physical parameters (such as Short Channel Effect (SCE) and Drain-Induced Barrier Lowering (DIBL)). Their impact on the delay and power uncertainties should be considered when optimizing yield, performance, and power. Moreover, process-invariant design styles will be necessary for future technologies. Some works have shown the effectiveness of leakage reduction techniques in the presence of process variations [3] [6]. The impact of process variations on the effectiveness of stack forcing is presented in [3]. In [6], the effectiveness of power and variability reduction of adaptive body bias and adaptive supply voltage is compared. The authors evaluated the techniques that adjust the body bias and supply voltage as post-silicon tuning to compensate intra-die variations. However, there is no study that compares different leakage reduction techniques with respect to their ability to control the leakage uncertainty. In [7], the authors discussed the impacts of the dynamic power management techniques on the delay variations and fabrication yield. Nevertheless, leakage power and leakage power reduction techniques are not considered in their

In this paper, we examine the mechanisms behind the mainstream active leakage reduction techniques that are used in design phase and evaluate the yield, delay and leakage power uncertainties under the presence of process variations when using these mechanisms. As both leakage reduction techniques and statistic timing analysis emerge as standard for smaller technologies, we examine how the leakage reduction techniques impact the statistic timing analysis and suggest how these emerging techniques can be implemented simultaneously. The influence of technology scaling and temperature sensitivity is also studied. This paper presents a comprehensive study of the influence of active leakage control techniques on the delay/leakage uncertainties in the presence of process variations.

The rest of the paper is organized as follows. In the next section, the sources of variations and the hints for possible uncertainty reduction techniques are presented. In Section 3, the review of the mechanisms for active leakage reduction techniques is presented. The experimental setups and results for exploring the influence on uncertainty are given in Section 4. Finally, the conclusions are provided in Section 5.

## 2. Process Variation Sources and Hints for Reducing the Impacts

– The electrical performance of an integrated circuit is impacted by two distinct sources of variation:  
 – Environmental factors and Physical factors [8]. In this paper, we focus on the variations of physical process variations. Among the variations in transistor parameters, variations in gate length and threshold voltage are found to have most significant impacts on circuit performance and power consumption [4][8]. For devices at technologies below 100nm, the variations in threshold voltage is caused by the doping fluctuation and thus influenced by the transistor geometry. This indicates that the effective variations in the threshold voltage is dependent on variation in gate length and can be explained by the following effects:

• **First order effect:** In [9], the standard deviation of the intrinsic threshold voltage for long channel devices is analytically modeled as:

$$\sigma V_{th} = \left( \frac{q}{C_{ox}} \right) \sqrt{\frac{N_{EFF} W_{DEP}}{3WL}} \quad (1)$$

Where  $q$  is the charge,  $C_{ox}$  is the gate oxide capacitance,  $N_{EFF}$  is the weighted doping concentration,  $W_{DEP}$  is the channel depletion width, and  $W$  and  $L$  are the channel width and gate length, respectively. This equation is derived for long channel devices which do not exhibit short channel effects. As shown in equation (1), the variation in threshold is caused by the doping un-uniformity and is proportional to the square root of doping concentration and inversely proportional to the square root of device gate length and channel width. From circuit design standpoint, the variations can be reduced by increasing the channel width or gate length of devices. However, the effectiveness

is limited due to the weak dependence as expressed in equation (1).

• **Second order effect:** The second order effect that causes the variations in transistor parameters is the result of threshold roll-off and DIBL. In nanometer technologies, the shorter device gate length reduces the effective threshold voltage to:

$$V_{th} = V_{th0} - \Delta V_{th}(V_{th\_roll\_off}) - \Delta V_{th}(DIBL)$$

Where the  $V_{th0}$  is the intrinsic threshold voltage and  $\Delta V_{th}(V_{th\_roll\_off})$  and  $\Delta V_{th}(DIBL)$  are the drop in threshold voltage due to short channel effect and DIBL, respectively. The relation between the gate length and threshold roll-off and DIBL is exponential and thus when there are variations in gate length, the net effect is increased variation in threshold voltage. Due to the exponential dependence, by increasing the gate length, the variations in both gate length and effective threshold voltage can be efficiently controlled

## 3. Review of Mechanisms for Reducing Active Leakage

In this section, we review the mechanisms behind mainstream active leakage reduction techniques whose influence on delay/leakage uncertainty we will expect in this work. The techniques considered include increasing gate length [10],  $V_{dd}/V_{th}$  optimization [11], body biasing [6], and stack forcing [12].

### - Increasing Gate Length

From equation (1), increasing the gate length not only reduces the leakage power but also reduces the leakage and delay uncertainties. Besides controlling the first order effect of variations, longer gate length also reduces the second order effect by lowering threshold roll-off and DIBL.

### - $V_{dd}/V_{th}$ Optimization

Due to the strong dependence of both dynamic and leakage power on power supply voltage, lowering supply voltage is used in most low power designs. Meanwhile, as multi-threshold processes are increasingly common, power optimization through tuning supply voltage to threshold voltage ratio ( $V_{dd}/V_{th}$ ) is used to achieve power-delay trade-off.

### - Body Biasing

The body effect causes threshold voltage roll-off and in turn higher leakage power. By reverse biasing the substrate of a transistor in sleep mode, the leakage current can be reduced. For post-silicon optimization, body bias is used to tune the threshold voltage back to target value.

### - Stack Forcing

The idea of “Stack Forcing” is to break a single transistor into a stacked transistor pair and thus the DIBL of the stacked transistor pair is reduced which in turn mitigates the leakage.

## 4. Exploring Influence On Uncertainty

To evaluate the impact of the leakage power reduction techniques on the uncertainty under the presence of process

variations, Monte-Carlo Hspice simulations are done in 65nm and 45nm technologies. In this section, the experiments and results are presented.

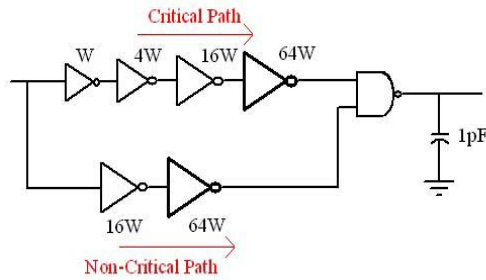


Figure 2. A branched inverter chain circuit.

Table 1. Simulation conditions.

	65nm	45nm
$L_{gate}$ (nm)	65	45
$ V_{th} $ (V)	0.2	0.16
$V_{dd}$ (V)	1.0	0.7
Temp. (°C)	85	85

#### 4.1 Experiment setup

A branched inverter chain shown in Fig. 2 is used as the target circuit in our exploration. This circuit is modified from canonical test inverter chain and sized to optimize the delay and emulate critical and non-critical paths. The simulation conditions are summarized in Table 1. The technology files used are 65nm and 45nm Berkeley Predictive Technology Model (BPTM) [13] where subthreshold leakage and gate leakage are captured. Note that the 3-sigma variation of gate length and intrinsic threshold voltage is set to be 10% and 1000 Monte-Carlo runs are simulated for each technique. The setup for each mechanism is discussed in the following.

##### 4.1.1 Gate length biasing

The optimized gate length is the trade-off between the power, delay and area. Simulation results in Fig. 3 show the achievable leakage power savings and leakage/delay uncertainties. In our experiment, we increase the gate length by 10% as point A shown in Fig. 3. It can be seen that increasing the gate length by 10% of minimum gate length achieves 85%, 55%, and 30% in the leakage savings, leakage uncertainty reduction and delay uncertainty reduction, respectively.

##### 4.1.2 $V_{dd}/V_{th}$ optimization

To find the optimized operating point of  $V_{dd}/V_{th}$  tuning, the design space is explored and the results are shown in Fig. 4. In Fig. 4, the trends of dynamic power, leakage power, leakage power uncertainty, and delay uncertainty at various  $V_{dd}$  and  $V_{th}$  levels are shown. These plots show that when the power supply voltage is lower, the delay and leakage uncertainties are larger. Another observation is that the delay uncertainty is smaller with higher  $V_{dd}/V_{th}$  ratio while leakage uncertainty is larger at this condition. These result in different optimized point for leakage power, delay uncertainty, and leakage uncertainty. This implies that, in nanometer technologies where both leakage power and

uncertainty present major challenges, the optimal  $V_{dd}/V_{th}$  ratio to use can change. Thus when selecting  $V_{dd}/V_{th}$ , special care is needed for the variation analysis. In our experiment, we select  $V_{dd}/V_{th}=0.7V/0.2V$  and  $V_{dd}/V_{th}=0.5V/0.16V$  for 65nm and 45nm technologies, respectively. The decision is made to minimize the leakage power so that the influence of leakage optimization on the uncertainty can be evaluated.

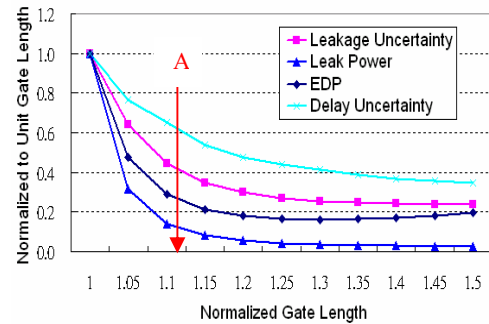


Figure 3. Design trade-off for optimized gate length.

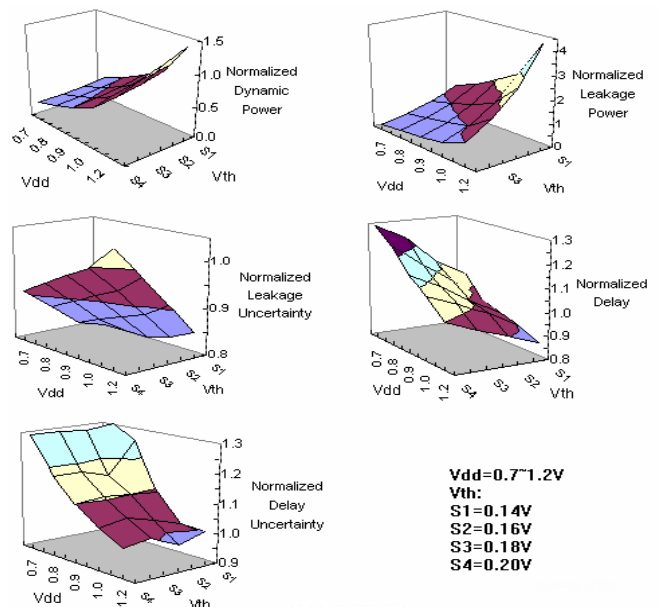


Figure 4.  $V_{dd}/V_{th}$  optimization for dynamic power, leakage power, delay, leakage power uncertainty, and delay uncertainty. The numbers in z-axis are normalized to the case  $V_{dd}/V_{th}=1.0V/0.2V$ .

##### 4.1.3 Body biasing

To employ this mechanism to reduce active leakage at design time, adaptive body bias is proposed. In this design, forward bias is applied to circuit in active mode (for high speed) while reverse bias is used in sleep mode (for low leakage power). With forward biasing to set to low threshold voltage in active mode, the default gate length can be made longer and thus causes smaller threshold roll-off and DIBL. The impact of uncertainty is similar to that of increasing gate length discussed above, and consequently not presented in more detail.

Except the mechanisms discussed above, we also evaluated “Stacking Forcing” given its significant effectiveness. Techniques combining multiple mechanisms are also investigated as optimization through simultaneously tuning multiple design parameters (i.e. tuning threshold voltage, supply voltage, stack forcing, and gate length assigning) are popular. Table 2 summarizes the techniques evaluated. Note that due to the large delay penalty of stack forcing, we evaluate two techniques: one that applies stack forcing to all elements and another that applies it to only the non-critical path.

Table 2. Leakage reduction techniques evaluated.

<i>Orig</i>	Original design without any optimization.
<i>SF</i>	Stack forcing to both PMOS and NMOS.
<i>LB</i>	Gate length biasing by increasing gate length by 10% of the minimum gate length.
<i>VOpt</i>	Tuning $V_{dd}/V_{th}$ to optimize leakage power.
<i>SFNC+LB</i>	Stack forcing on non-critical path and gate length biasing
<i>SF+LB</i>	Stack forcing and gate length biasing
<i>VOpt+LB</i>	$V_{dd}/V_{th}$ optimization and gate length biasing
<i>SLV</i> ( <i>SFNC+LB+VOpt</i> )	Combining stack forcing on non-critical path, gate length biasing and $V_{dd}/V_{th}$ optimization

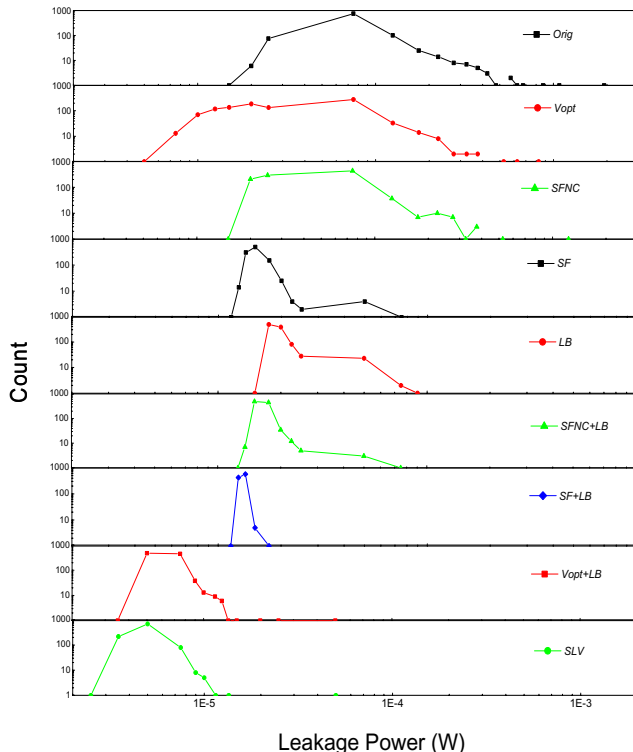


Figure 5. Leakage distribution of evaluated techniques.

## 4.2 Effect of Technology Scaling on Uncertainty

The power and delay simulation results of the test circuit are shown in Table 3 and can be used for evaluating the impacts of active leakage reduction techniques. Throughout this paper, the uncertainty is defined as the standard deviation divided by the mean value (S.D./mean). By comparing the data of 65nm and 45nm, we can see that both the delay and leakage power uncertainties increase with technology scaling. Note that the leakage uncertainty is larger than delay uncertainty across technologies. This is due to the stronger dependence of leakage (exponential) on threshold voltage than the dependence of delay (logarithmic) on threshold voltage.

Table 3. Basic simulation results of the test circuit.

Technology	65nm	45nm
Leakage Power (W)	3.43e-5	2.53e-4
Leakage Uncertainty	0.38	0.46
Dynamic Power (W)	2.74e-4	3.07E-04
Delay(S)	2.15E-10	1.5E-10
Delay Uncertainty	0.12	0.23

## 4.3 Effect of Leakage Reducing Scheme on Uncertainty

The leakage histograms and cumulative percentage of leakage distribution are plotted in Fig. 5. The data shown is for 65nm technology. It can be seen that, for all the evaluated techniques, not only the worst case leakage is reduced but the distribution is sharper (note that the x- and y-axis are in logarithmic scale), which means better yield control. Another observation is that *SF* achieves most leakage savings and results in least leakage uncertainty. However, *SF* is only recommended in non-critical paths due to its large delay penalty. From Fig. 5, we can see that combining *LB* and *SF* in non-critical path (*SFNC+LB*) achieves additional leakage savings and provides further uncertainty reduction. *SFNC+LB* achieves similar leakage savings and leakage uncertainty as *SLV*, which includes tuning  $V_{dd}/V_{th}$ . However, *SLV* offers better delay uncertainty and lower dynamic power consumption. For designs that delay is not the primary constraint, it is recommended to use *SF+LB* hybrid technique to control the yield while achieving minimum leakage power. The results of 65nm technology are summarized in Table 4. The numbers are normalized with respect to the corresponding values for the base case where no leakage reduction technique was used (*Orig*). All the evaluated techniques reduce dynamic/leakage power at the cost of delay penalty. Due to different optimization points for reducing leakage power and delay/leakage uncertainty through  $V_{dd}/V_{th}$  tuning, when it is tuned to minimize leakage power, the leakage and delay uncertainties could deteriorate as can be seen from Table 4. The lower supply voltage incurs larger leakage uncertainty while lower  $V_{dd}/V_{th}$  ratio helps to reduce the leakage uncertainty. On the other hand, both conditions result in larger delay uncertainty.

The fifth row of Table 4 shows the yield achieved for each technique and can be used for evaluating the impact

Table 4. Results in 65nm technology.

	<i>Orig</i>	<i>SF</i>	<i>LB</i>	<i>VOpt*</i>	<i>SFNC+LB</i>	<i>SF+LB</i>	<i>VOpt+LB</i>	<i>SLV</i>
Leakage Power	1.00	0.08	0.14	0.45	0.11	0.02	0.07	0.05
Leakage Uncertainty	1.00	0.45	0.45	0.91	0.52	0.25	0.42	0.48
Dynamic Power	1.00	0.84	0.89	0.44	0.64	0.80	0.41	0.29
Yield (%)	100	0.2	86.8	88.5	77.1	0	1.6	0.1
Delay	1.00	3.30	1.43	1.29	1.88	4.70	1.99	2.67
Delay Uncertainty	1.00	0.71	0.64	1.27	0.62	0.55	0.78	0.71

\* For *VOpt* in 65nm Technology,  $V_{dd}/V_{th}$  is set to be 0.7V/0.2V.

on the yield given the increase in delay and delay uncertainty in some of the techniques. Note that the yield presented is defined as the percentage of the cases, out of the 1000 simulation runs, in which the resulting delay is within the worst-case corner delay of *Orig* (shown as point *W* in Fig. 6) where no optimization technique is used. The worst-case corner delay time of *Orig* is the delay time assuming all the worst case conditions happened at the same time (both a 10% increase in threshold voltage and gate length). While this worst-case corner rarely happened, this is what the corner case simulation assumed in the conventional static timing analysis. The results show that increasing the gate length by 10% achieves 86% savings in leakage power and 86.6% yield. In comparison, optimizing  $V_{dd}$  and threshold voltage ratio achieves 55% leakage savings and 88.5% yield. As a result, even though the average delay time increases when using leakage reduction techniques, based on corner simulation results, moderate yield can still be achieved while leakage power is reduced. We would like to point out that only the transistor delay is evaluated in this experiment.

To study the influence of the leakage reduction techniques on delay uncertainty, the delay distribution of each technique evaluated is plotted in Fig. 6. *VOpt* technique increases the delay uncertainty while *SF* and *LB* have narrow spread of delay distribution. In *VOpt*, since the  $V_{dd}/V_{th}$  used for the non-critical path has a higher delay uncertainty than that of the critical path in the presence of process variation, the non-critical path can actually turn out to be the delay bottleneck. Consequently, the scope for the leakage reduction using  $V_{dd}/V_{th}$  optimization for non-critical paths reduces in the presence of process variation. Given the dependence between the *VOpt* technique and delay uncertainty, and as statistic timing analysis becomes standard to contend with the within-chip variations, power optimization through  $V_{dd}/V_{th}$  tuning needs to be done statistically to achieve maximum power savings while preventing the build-up of critical paths.

Both *SF* and *BL* reduce leakage and delay uncertainties efficiently and their effectiveness increases for smaller technologies where SCE is more pronounced. This is due to their ability to reduce the second order effect of process variations discussed in Section 2.

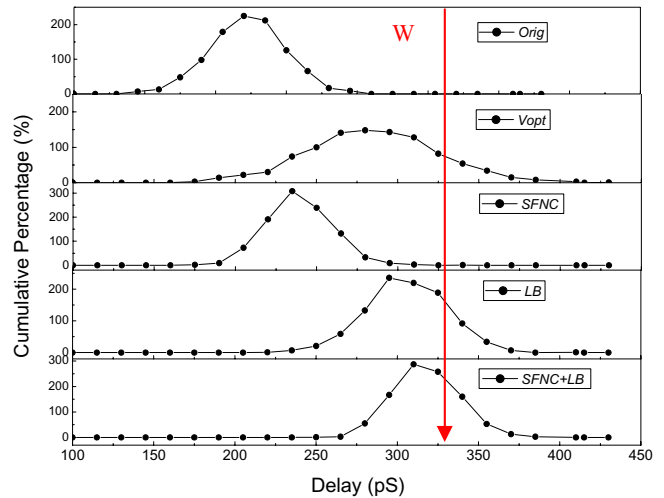


Figure 6. Delay time of test circuit for each technique. *W* point is the worst-case corner delay of *Orig*.

#### 4.4 Temperature Variation Sensitivity

Due to the exponential dependence of threshold voltage on temperature, we also evaluate the temperature sensitivity of delay and leakage uncertainties. Monte Carlo analysis is done at temperature from 25°C to 125°C. The uncertainty shown in Fig. 7 is the deviation in the delay/leakage divided to the mean value of leakage/delay at 25°C. The uncertainty is then normalized to the numbers of *Orig* for comparison. From the data shown in Fig. 7, we can see that both delay and leakage uncertainties increase with temperature. This can be due to the increasing leakage and larger delay at high temperature. However, the behaviors of leakage uncertainty and delay uncertainty across temperature variations tend to be different across techniques. Employing *LB* and *SF*, the variation in leakage uncertainty with temperature is larger as compared to *Orig* and *VOpt*. This is because the mean leakage is lower in the cases of *LB* and *SF* as compared to *Orig*. However, we also see that the uncertainty of *LB* is larger than that of *SF*, whose mean leakage is smaller. This is due to the higher equivalent threshold voltage when employing *SF*. When applying *LB*, gate length is increased and thus SCE is reduced, which in turn raises the equivalent threshold voltage. For *SF*, the equivalent threshold voltage is further increased due to the reduced DIBL. In the case of *VOpt*, since the optimized condition tunes the  $V_{dd}$  only, it is similar to *Orig* where is no leakage reduction technique is

applied. For delay uncertainty, *Vopt* exhibits similar behavior as *Orig* while *LB* and *SF* exhibit larger delay uncertainty. For *SF*, the delay uncertainty doubles from room temperature to 125 °C. We explain the different behavior for delay uncertainty as follows. Fig. 8 illustrates the reverse logarithmic relation between delay time and gate driving voltage. The operating range of the gate driving voltage for each technique is also shown. *Vopt* operates in the same range as *Orig* due to unchanged threshold voltage when it applies. As a result of the above mentioned raised equivalent threshold voltage, designs with either *SF* or *LB* applied operate at lower gate driving voltage. As illustrated in Fig. 8, the same amount of variation in threshold voltage, the change in delay time of *SF* is larger than the other techniques and this explains for its larger delay uncertainty than that of other techniques.

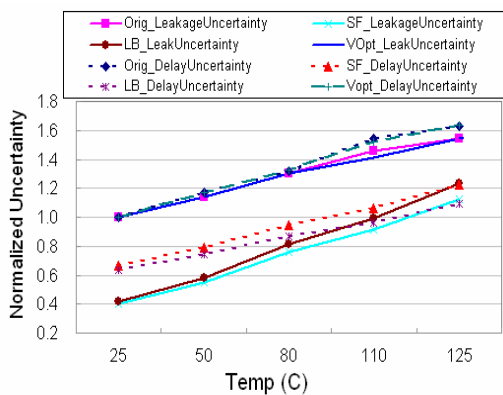


Figure 7. Temperature sensitivity of uncertainties of the evaluated techniques.

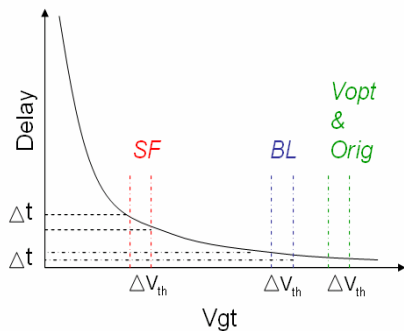


Figure 8. Illustration of delay time vs gate driving voltage.

$$V_{gt} = V_{gs} - V_{th}$$

## 5. Conclusions

The influence of leakage reduction techniques on delay and leakage uncertainties is evaluated. It is found that the technique using lower power supply level or high  $V_{dd}/V_{th}$  increases both delay and leakage power uncertainties. Special care is needed to control the yield when optimizing delay and power trade-off through tuning  $V_{dd}/V_{th}$ . Stack forcing and increasing gate length reduce the delay and leakage uncertainties at the cost of delay and area penalties. We suggest that  $V_{dd}/V_{th}$  tuning should be done in a statistic fashion while gate length tuning and stack forcing can be assigned with static timing analysis. To further exploit the

trade-off space, we suggest some hybrid techniques. By increasing gate length of every transistor and forcing stack of transistors in non-critical paths, the delay and leakage power uncertainties can both be controlled while achieving noticeable leakage savings. Both delay and leakage uncertainties are expected to increase with technology scaling. Fortunately, the effectiveness of controlling the uncertainty through stack forcing and increasing gate length also increases. Another observation is that increasing the gate length or forcing stacked transistors, the temperature sensitivity of both delay and leakage uncertainties increase.

In this paper, we point out the importance of considering the delay and leakage uncertainties when applying leakage reduction techniques and quantify the uncertainties caused by process variations.

## References

- [1] Y-F. Tsai, et al, "Implications of Technology Scaling on Leakage Reduction Techniques", Design Automation Conference, pp.187-190, Jun. 2003
- [2] B. Chatterjee, et al, "Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies", International Symp. On Low Power Electronics and Designs, pp. 122-127, Aug 2003
- [3] S. Nerendra, et al, "Full-Chip Subthreshold Leakage Power Prediction and Reduction Techniques for Sub-0.18-um CMOS", IEEE Journal of Solid-State circuits, VOL. 39, No. 2, pp. 501-510, Feb. 2004
- [4] Bokar, et al., "Parameter Variations and Impact on Circuits and Microarchitecture", Design Automation Conference, pp.338-342, June 2003
- [5] A. Keshavarzi, et al, "Intrinsic Leakage In Deep Submicron CMOS ICs – Measurement-Based Test Solutions", IEEE Trans. On VLSI Systems, VOL. 8, No. 6, pp. 717-723, Dec. 2000
- [6] T. Chen, et al, "Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and leakage Under the Presence of Process Variation", IEEE Transaction of VLSI Systems, Vol. 11, No. 5, pp.888-899, Oct. 2003
- [7] Y. Cao, et al, "Yield Optimization with Energy-Delay Constraints in Low-Power Digital Circuits", IEEE Conference on Electron Devices and Solid State Circuits .pp. 285-288, Dec 2003
- [8] Nassif, S.R., "Modeling and forecasting of manufacturing variations", Asia and South Pacific Design Automation Conference, pp. 145-149, Feb. 2001
- [9] Takeuchi, K, "Channel engineering for the reduction of random-dopant-placement-induced threshold voltage fluctuation", International Electron Devices Meeting, pp. 841-844, Dec. 1997
- [10] N. Sirisantana, et al, "High-Performance CMOS Circuits using Multiple Channel Length and Multiple Oxide Thickness", International Conference on Computer Design, pp. 227-232, Sep. 2000
- [11] L. Wei, et al, "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits", Design Automation Conference, pp. 489-494, Jun. 1998
- [12] Ye Y., et al, "A New Technique for Standby Leakage Reduction in High-Performance Circuits", Symp. On VLSI Circuits, pp. 40-41, Dec 1998
- [13] UC Berkeley Device Group, <http://www-device.eecs.berkeley.edu/~ptm>