CMOS CHARACTERIZATION, MODELING, AND CIRCUIT DESIGN IN THE PRESENCE OF RANDOM LOCAL VARIATION

by

Benjamin A. Millemon Sr.

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Electrical Engineering

Boise State University

August 2012

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

Of the thesis submitted by

Benjamin A. Millemon Sr.

Thesis Title: CMOS Characterization, Modeling, and Circuit Design in the Presence of Random Local Variation

Date of Final Oral Examination: 9 April 2012

The following individuals read and discussed the thesis submitted by student Benjamin A. Millemon Sr., and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

R. Jacob Baker, Ph.D.	Chair, Supervisory Committee
Vishal Saxena, Ph.D.	Member, Supervisory Committee
Bill Knowlton, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by R. Jacob Baker, Ph.D., Chair of the Supervisory Committee. The thesis was approved for the Graduate College by Dr. John R. Pelton, Dean of the Graduate College.

AUTOBIOGRAPHICAL SKETCH OF AUTHOR

Benjamin A. Millemon Sr. served 5 years on active duty with the United States Army as a Paratrooper and Spanish Linguist in the 82nd Airborne Division from 1994 to 1999. After serving, he attended Boise State University from 1999 to 2003, completing his Bachelor's degree in Electrical Engineering. He worked as an intern at Micron Technology in the DRAM compact modeling group from January 2002 to May of 2003. He started full-time in May of 2003 in the same group at Micron Technology where he has been involved with all aspects of CMOS device modeling, reliability, variability, noise, speed, and interconnects in support of DRAM, SRAM, CMOS Imagers, and NAND products. He is currently the manager of the compact modeling and characterization team for DRAM CMOS development at Micron Technology. His wife of 17 years, Jennifer, and their children, Benjamin Jr., Alyssa, and Savanna, are the pride of his life and reside in Boise, Idaho.

ABSTRACT

Random local variation in CMOS transistors complicates characterization procedures, modeling efforts, simulation tools, and circuit design methodologies in highly scaled CMOS devices. Mismatch is not only a concern for closely matched device pairs in analog circuits; digital circuit designers also have to consider the effects of random variation. Device characterization, modeling, process development, and circuit design engineers have to work together to mitigate the impact of random local variation. This thesis outlines the primary challenges of CMOS characterization, modeling, and circuit design in the presence of random local variation and offers guidelines and solutions to help mitigate and model the unique characteristics that mismatch introduces. Random data sets are generated to demonstrate the statistical transistor and circuit response to random variation across die and process and to demonstrate the challenges in each area of CMOS development.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND KEY TERMINOLOGY

A_{VT}	Slope of the Pelgrom plot for a delta VT mismatch
$A_{VT,local}$	Slope of the Pelgrom plot for an individual device ($A_{VT}/\sqrt{2}$)
CHC	Channel Hot Carrier degradation mechanism
CMOS	Complementary Metal Oxide Semiconductor
IOFF	sub-threshold MOSFET leakage at VGS=0
LER	Line-Edge Roughness
LOCAL	Denotes random variation within a die
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
NBTI	Negative Bias Temperature Instability
OCV	On-Chip Variation, generally random in nature
VBS	Voltage from bulk to source
VDS	Voltage from drain to source
VGS	Voltage from gate to source
VT	MOSFET threshold voltage at low drain to source bias
VTSAT	MOSFET threshold voltage with VDD from drain to source

CHAPTER ONE – INTRODUCTION

1.1 Introduction

Random mismatch in threshold voltage and carrier mobility in complementary metal oxide semiconductor (CMOS) transistors has been present since their inception. Random atomic-level fluctuations cause behavioral differences between transistors such that no two transistors are ever exactly the same. Figure 1 shows an atomistic device level simulation of the surface potential variation along with the random channel dopant fluctuations in a highly scaled CMOS transistor illustrating the discrete nature and random placement of dopant atoms for modern devices [1]. There are fewer than 100 dopant atoms in the channel of most sub 50 nm devices.

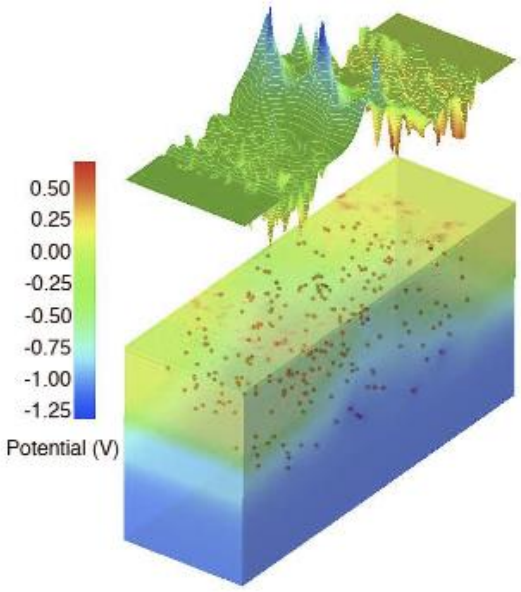

Figure 1. A three-dimensional atomistic simulation showing a statistically rare scenario of dopant atom placement and the corresponding surface potential for a sub-50 nm CMOS transistor [1].

The landmark paper by Pelgrom in 1989, which established a clear relationship between the area of the MOSFET device and the local threshold voltage variation, still holds up quite well on modern CMOS devices [2]. Marcel Pelgrom established that the threshold voltage variation (sigma VT = σ_{VT}) for closely placed devices increases as the device area is reduced and is inversely proportional to the square root of the device area $\left(\sigma_{VT} \approx 1/\sqrt{area}\right)$. This relationship is referred to as the Pelgrom law. Tomohisa Mizuno established a direct relationship to oxide thickness in [3] where the random local σ_{VT} was shown to decrease linearly with decreasing oxide thickness according to Equation 1. This relationship also shows that σ_{VT} is proportional to the fourth root of the number of dopant atoms in the channel. Equation 2 simplifies the process dependent variables into a single variable $A_{VT,local}$. $A_{VT,local}$ can be used to model local *VT* variation for a given process node and is defined in Equation 3 using the pre-factor from Equation 1. $A_{VT,local}$ is generally reported with units of mV-um. The relationship in Equation 1 only explains about 60% of the local variation. The rest of the variation is generally tied up in interface states, charge in the oxide, and poly grain boundary variation.

$$\boldsymbol{\sigma_{VT,local} = \frac{\sqrt[4]{4q^3\varepsilon_{Si}\varphi_B}}{2} \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \frac{\sqrt[4]{N_{tot}}}{\sqrt{L_{eff} \cdot W_{eff}}}} \qquad \textbf{Eq. 1}$$

$$\boldsymbol{\sigma_{VT,local} = \frac{A_{VT,local}}{\sqrt{L_{eff} \cdot W_{eff}}}} \qquad \textbf{Eq. 2}$$

$$\boldsymbol{A_{VT,local} = \frac{\sqrt[4]{4q^3\varepsilon_{Si}\varphi_B}}{2} \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \sqrt[4]{N_{tot}}} \qquad \textbf{Eq. 3}$$

It is common to characterize local variation by measuring the difference between two closely placed devices, which will be discussed in more detail in Chapter Three. The

variance in the difference between two devices is larger than the local variance of an individual device by a factor of 2, which arises from summing the variances of the two identical devices as described by Equation 4.

$$\boldsymbol{\sigma^2_{deltaVT} = \sigma^2_{local,A} + \sigma^2_{local,B} = 2 \cdot \sigma^2_{local}} \qquad \textbf{Eq. 4}$$

This factor of two is not often clarified in literature and can lead to misinterpretation of experimental results. Equation 5 relates the local and delta *VT* slopes. Figure 2 shows an example of a Pelgrom plot with an A_{VT} slope of 3.4 mV-um. Note that this plot is usually reported using A_{VT} from sigma delta *VT* rather than $A_{VT,local}$. It is important to understand which version of slope is being reported in the model provided by the foundry in order to accurately predict the device and circuit response.

$$\boldsymbol{A_{VT} = \sqrt{2} \cdot A_{VT,local}} \qquad \textbf{Eq. 5}$$

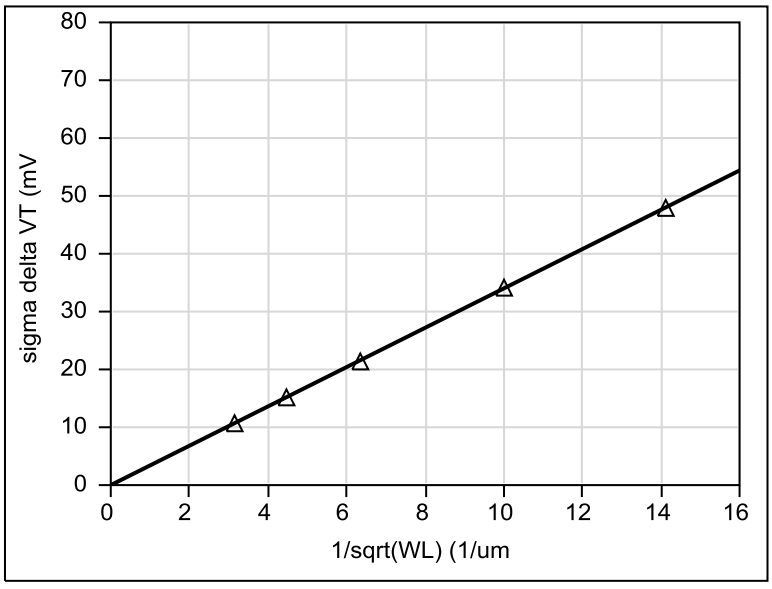

Figure 2. A sample Pelgrom plot showing sigma delta VT plotted against the inverse of the square root of the area for various device geometries with an A_{VT} of 3.4 mV-µm.

Figure 3 illustrates how the threshold voltage varies with the number of atoms in the channel for a highly scaled transistor and shows a modest reduction in variation as the number of dopant atoms decreases, which is consistent with Equation 1.

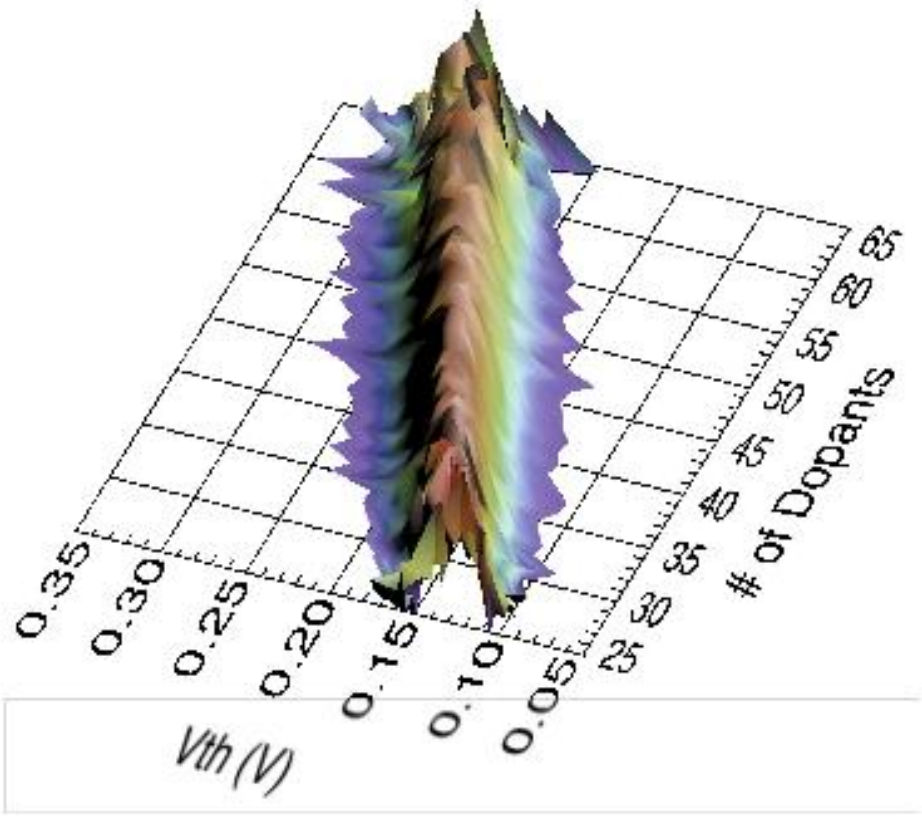

Figure 3. Intrinsic VT and VT variation plotted against the number of dopant atoms in the channel showing how the VT and VT variation are both reduced as the number of dopant atoms decreases [1]

As technologies have scaled, mismatch has actually improved consistently due to a steady decrease in oxide thickness and channel doping to achieve lower threshold voltages [1] [4]. In recent years, however, the oxide thickness and threshold voltage

scaling has slowed down in some technologies due to fundamental limits in reliability and standby power eliminating the benefits they provided. Some CMOS technologies have stopped scaling oxide thickness and threshold voltage and are susceptible to excessive levels of random local variation. In addition, the gate overdrive voltage has been decreasing from reductions in the power supply voltage, which greatly increases the device sensitivity to threshold voltage fluctuations.

Random local variation or mismatch has generally been a concern for analog designers that leverage matched pairs for many applications, however, mismatch is now impacting internal timing margins in digital circuits and is at the forefront of the barriers limiting cutting edge logic design. Transistor and gate-level simulation tools now offer a variety of options for simulating random local variation or on-chip variation (OCV), and foundries are offering the needed statistical models to simulate the effects. Research in the area has exploded over the last 10 years and the IEEE Electron Device Society recently compiled a special issue dedicated to the characterization of nano CMOS variability, much of which was devoted to understanding the implications of random local variation [5]. Advances in process technology such as high-K/Metal gates and un-doped channels for thin body SOI (silicon on insulator) and FINFET ("fin" field effect transistor) technologies have shown significant improvements in mismatch performance. Despite the strides in the technology solutions, mismatch remains a significant source of variability impacting yield in high speed, low voltage CMOS circuit designs.

Device development teams have to comprehend the process parameters that impact mismatch and properly model the device variation. Circuit designers also have to understand how the random variations affect internal timing margins, standby currents,

and ultimately product yield in high volume manufacturing. This thesis will provide insight into the latest challenges and solutions for characterization, modeling, and digital circuit design in the presence of random local variation. This thesis will not go into depth on the impacts of local variation (mismatch) in sensitive matched pairs that have been observed and researched in great detail for more than 20 years.

CHAPTER TWO – MISMATCH SCALING AND ORIGINS

2.1 Gate Overdrive Voltage Scaling

A key factor in determining the impact of threshold voltage variations is the gate overdrive voltage (V_{ov}), which is the difference between the gate to source voltage (*VGS*) and the threshold voltage (*VT*). Higher V_{ov} results in less drain current modulation for a given *VT* shift. Lower V_{ov} causes the drain current to be much more sensitive to *VT* variation, degradation, and power supply fluctuations. This is an important concept to understand and consider when evaluating the impact of threshold voltage variation across supply voltage. At a fixed power supply voltage, lower threshold voltage devices are less sensitive to all of these factors since they operate with greater V_{ov}. It is a good design practice to use as low of a threshold voltage as possible at a given supply voltage without burning too much standby power. Often the supply voltage is set by customer specifications and is not a variable that can be used by a designer to optimize performance. The V_{ov} dependency is made apparent when we examine the drain current Equations for long and short channel devices in saturation. It is evident in the generic square law in Equation 6, for the long channel MOSFET, that the sensitivity of the drain current in saturation (IDS_{sat}) with respect to the threshold voltage decreases as the supply voltage increases and is proportional to the square of the difference between the power supply voltage and the threshold voltage.

$$IDS_{sat} = \frac{\mu \cdot C_{OX}}{2} \cdot \frac{W}{L}(VDD - Vt)^2 \qquad \textbf{Eq. 6}$$

The gate to source voltage (*VGS*) in Equation 6 has been replaced by the power supply voltage, *VDD*. The sensitivity to changes in *VGS* (or *VDD*) is referred to as the transconductance or g_m and is shown in Equation 7. The transconductance with respect to *VT* can be quantified at low or high *VDS* and referred to as $g_{msat,VT}$ in saturation and $g_{mlin,VT}$ at low *VDS* in the linear regime. Likewise, the transconductance with respect to VDD is given by $g_{msat,VDD}$ and $g_{mlin,VDD}$. The transconductance is not necessarily a constant value across device geometry and bias conditions; however, the general relationship can be understood in these simplified expressions.

$$\frac{dIDS_{sat}}{dVt} = g_{msat,VT} = -\mu \cdot C_{OX} \cdot \frac{W}{L}(VDD - VT) \qquad \textbf{Eq. 7}$$

The ratio of the change in IDS_{sat} with respect to *VDD* and *VT* simplifies to Equation 8 and 9 respectively, where the sensitivity to each only differs in polarity.

$$\frac{\frac{dIDS_{sat}}{dVT}}{IDS_{sat}} = \frac{-2}{(VDD - VT)} \qquad \textbf{Eq. 8}$$

$$\frac{\frac{dIDS_{sat}}{dVDD}}{IDS_{sat}} = \frac{2}{(VDD - VT)} \qquad \textbf{Eq. 9}$$

The saturation current, or drive current (I_{drive}), is specified when *VGS* = *VDD* = *VDS* and is normalized per µm transistor width. I_{drive} is a decent indicator of the relative digital speed and can be used to estimate the effective switching resistance in a MOSFET [6]. Figure 4 shows the percent change in I_{drive} as a function of *VDD* for a 10 mV shift in *VT* across *VDD*. *VT* variation generally runs in the 10's of mV's range, therefore it is useful to quote the sensitivity with respect to a 10 mV shift. It is evident that lower V_{ov} results in a higher sensitivity to *VT* or *VDD* modulations.

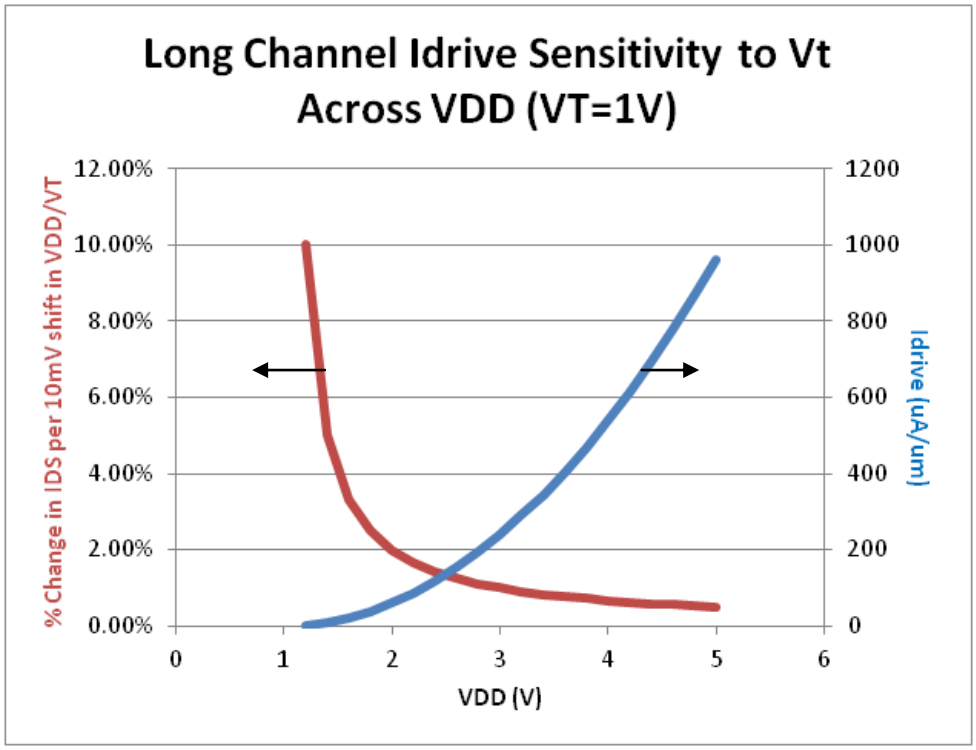

Figure 4. Long channel I_{drive} current and I_{drive} sensitivity to VDD across VDD, illustrating an exponential sensitivity to VDD.

Circuits that operate with higher V_{ov} are more robust to anything that shifts the threshold voltage or power supply voltage. This includes variability, noise, and device

degradation over time. The sensitivity of a circuit to changes in power supply is commonly referred to as *PSS* (power supply sensitivity) and is analogous to changes in the threshold voltage.

The *VDD* and *VT* sensitivity is reduced for short channel devices but remains significant. The short channel saturation current can be expressed by Equation 10 [7]. The saturation current is now linearly proportional to V_{ov}, instead of the square of V_{ov} like it was in the long channel Equation.

$$\boldsymbol{IDS_{short} = vsat \cdot C'ox \cdot W \cdot (VDD - Vt - VDS, sat)} \qquad \textbf{Eq. 10}$$

The sensitivity to *VDD* or *VT* is then simply the saturation transconductance $gm_{sat,VT}$, which is simply the pre-factor $vsat \cdot C'ox \cdot W$ and is given by Equation 11. The ratio of $gm_{sat,VT,short}$ to *Idrive* is approximated by Equation 12. The pre-factor $gm_{sat,VT,short}$ can be measured from standard current-voltage curves (IV curves) from a sweep of *VGS*. When evaluating the impact of *VDD*, it is useful to sweep the gate and drain together in a diode configuration so that Idrive can be evaluated across *VDD*. *Idrive*, shown in Figure 4, represents the diode-connected case. The I_{drive} sensitivity for the short channel devices is shown in Figure 5. It should be evident that lower V_{ov} results in higher sensitivity to changes in *VT* and *VDD* for both short and long channel devices. The overdrive voltage is an important consideration when evaluating the impact of *VT* and *VDD* variations.

$$\frac{dIDS_{short}}{dVt} = -vsat \cdot C'ox = gm_{sat,VT,short} \quad \textbf{Eq. 11}$$

$$\frac{\frac{dIDS_{short}}{dVt}}{IDS_{short}} = \frac{-1}{(VDD-Vt-VDS,sat)} \quad \textbf{Eq. 12}$$

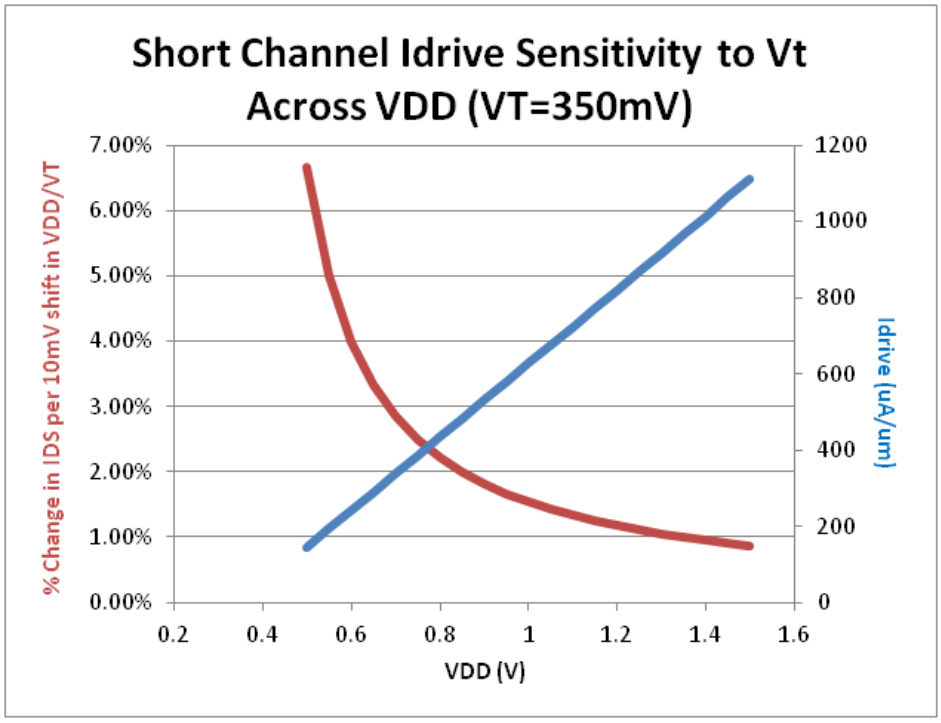

Figure 5. Short channel I_{drive} current and I_{drive} sensitivity to VDD across VDD, illustrating a linear sensitivity to VDD

2.2 Process Parameters Affecting Random Variation

The process parameters available for reducing local variation are generally fundamental to the technology itself and are thus not really variables that can be tuned. They are generally the result of scaling for increased speed and tolerable leakage. Most technologies fall within about 20% of the observed relationship documented in [8] and shown in Equation 13 for NMOS devices. The PMOS relationship is shown in Equation

14. The non-zero slope is not inconsistent with Equation 1, it simply arises from the fact that the channel dopant concentration has tended to increase as T_{ox} has scaled. A_{VT} will tend towards zero when T_{ox} is scaled only if all other factors are held constant.

$$A_{VT,nmos} = 1\ \frac{mV \cdot um}{nm} \cdot T_{ox} + 2\ mV \qquad \textbf{Eq. 13}$$

$$A_{VT,pmos} = 0.75\ \frac{mV \cdot um}{nm} \cdot T_{ox} + 1.5\ mV \qquad \textbf{Eq. 14}$$

Furthermore, the changes in process technology that modulate the local variation away from this trend are fundamental technology metrics themselves. For example, lower channel enhancement implants that produce lower *VT* devices tend to reduce mismatch. The application may require the higher *VT* and higher enhancement dose to keep standby currents under control thereby eliminating that option. Thinner gate oxide thickness also produces better mismatch, but gate leakage, negative bias temperature instability (*NBTI*), and time-dependent dielectric breakdown (*TDDB*) can limit the scaling of the gate oxide thickness for a given power supply voltage. Smaller poly grain size has been shown to reduce mismatch in [9]. Poly depletion and boron penetration have also been shown to impact mismatch in [10]. Increased poly depletion, boron penetration and larger poly grains increase threshold voltage mismatch. The reduction in grain size from amorphous silicon deposition with a furnace anneal down to a poly-silicon deposition with a rapid thermal oxidation showed a drop in A_{VT} from 6.08 to 3.46 for NMOS and 11.2 to 2.85 for PMOS in [8]. These are significant mismatch improvements and good for studying the grain-size effects, but the amorphous process conditions are not a likely candidate for a production-worthy process. The results show that there are process conditions that can greatly increase local *VT* variation but not much

can be done to improve it beyond the general trends in equations 13 and 14. Again, there are limited process parameters that improve mismatch for a given technology. Device and circuit designers need to account for the variation during the development cycle to ensure that process conditions do not cause excessive mismatch.

It is interesting to note that as long as the oxide thickness (T_{ox}) scaling is proportional to the device area or length reductions, the mismatch actually stays relatively constant for a given *W/L* ratio. Figure 6 shows some projected A_{VT} curves as the length is scaled from 250 nm down to 50 nm, while T_{ox} is scaled from 10 nm down to 1.5 nm as the length is scaled down.

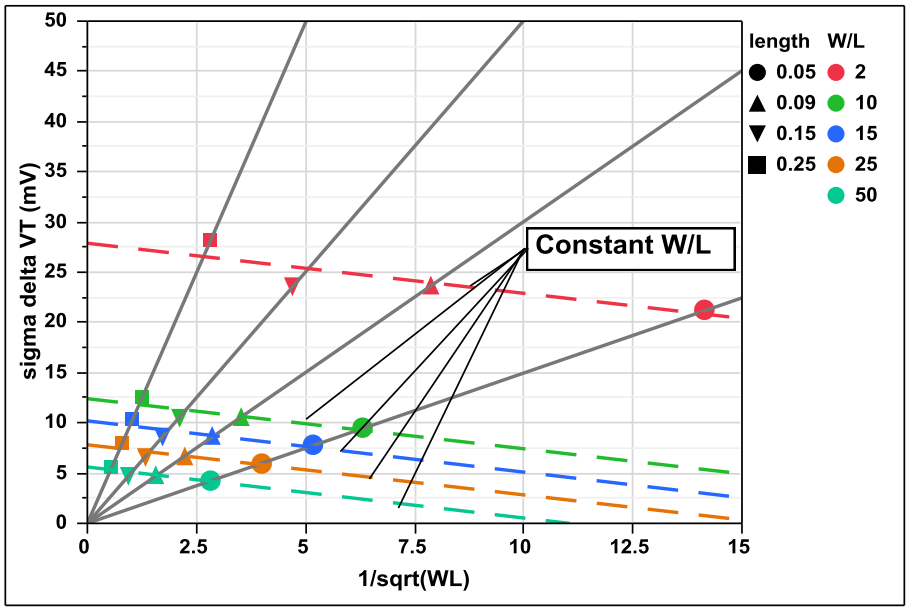

Figure 6. A Pelgrom plot across technology nodes across A_{VT}, illustrating a possible decrease in *VT* mismatch for a constant W/L.

In each case, the mismatch for a given *W/L* actually reduces slightly as we scale. Mismatch can indeed improve as technology scales; however, that does not mean that the

impact of mismatch on circuit performance will improve. Reductions in overdrive voltage will increase the sensitivity to mismatch and circuit margins will likely be tighter. Figure 6 does not consider the effects of decreasing overdrive voltage. If the overdrive voltage is greatly reduced across these technologies, then the same amount of mismatch will produce more I_{drive} variation as previously discussed and shown in Figures 4 and 5. Referring to Figures 4 and 5, it is evident that 10 mV's of *VT* variation produced 0.5% shift in Idrive with V_{ov} = 4V, but the same 10 mV shift produces about 300% more I_{drive} variation on a short channel device operating with 650 mV's of V_{ov}. Again, the reduction in V_{ov} is a root issue that is increasing *VT* sensitivity and causing CMOS integration engineers and circuit designers to evaluate the sources of variation in greater detail. Tighter timing margins are also helping bring each source of variation to the forefront. A lack of T_{ox} scaling coupled with decreasing overdrive voltage will result in excessive intra-die performance variation and should be avoided when possible. CMOS development teams and circuit designers need to work closely to capture the behavior accurately in such scenarios.

In some technologies, the oxide thickness and *VT* have not been scaled as aggressively as the supply voltage for a variety of reasons. Scaling the device width and length without scaling T_{ox} and *VT* will not result in a reduction of A_{VT}, therefore mismatch will increase along the curves shown in Figure 6. Recall that A_{VT} is primarily a function of oxide thickness and the *VT* target. This situation can cause excessive mismatch, which could result in yield loss if not properly characterized, modeled, and simulated during the design cycle.

Line-edge roughness (*LER*) is another source of local random variation that becomes significant at the 32 nm gate length regime [4]. The gate edges can cause variation in the placement of dopant atoms in self-aligned implants. The implants themselves will generally follow a discrete nature even when the gate edges are very smooth, but reducing *LER* is important for highly scaled CMOS devices. Figure 7 shows an atomistic simulation depicting line-edge roughness in a 32 nm CMOS transistor given in a keynote address by Asen Asenov, a pioneer in atomistic simulation technology and modeling [11].

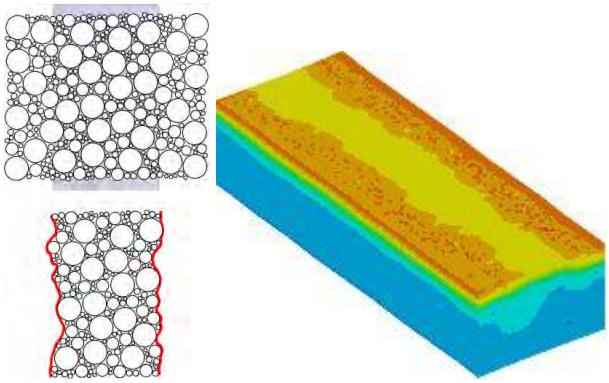

Figure 7. Atomistic cartoon and simulation of line edge roughness (LER) in source/drain dopant atoms due to poly grain boundaries

2.3 Variability Components

From a circuit designer's perspective, there are two primary types of variation that need to be simulated and understood. The first is the chip-to-chip variation, also called inter-die variation, in which all devices on the chip move together at the same time. This is the traditional methodology that has been employed in corner modeling for many years. The typical NMOS and typical PMOS models simulate with the exact same performance for every instance in the netlist. The assumption in this simulation is that

the devices do not have a process gradient of any kind across the chip and that there is no significant random local variation. The fast NMOS and fast PMOS corners also assume that all devices in the netlist behave equally fast. Likewise, the slow corners and combinations of each produce the same behavior for every similar device in the netlist. In each case, the devices are centered at a particular point, but no single instance of the same model is unique. These corner models work perfectly as long as local variation is negligible.

The second type of variation that is important to designers is the random within die or intra-die variation. This is also called mismatch or local variation. These simulations generally have to be simulated with a Monte Carlo analysis in which successive simulations run with each device instance having unique behavior. This can be computationally expensive and time consuming but the response takes into account the joint probability of multiple variables being changed at the same time. This variation is proportional to the A_{VT} slope that arises from local variation for a given device.

Process gradients or systematic offsets within the die are also a concern and could be considered a third concern. Layout dependent offsets associated with device proximity to mask edges, adjacent gates, and *STI* (shallow trench isolation) are all significant sources of variation. Process gradients are significant when the die size is large with respect to the wafer. Reticle field gradients can also impact performance. These effects are all extremely relevant but can generally be minimized with thorough device characterization and proper design rules and will not be covered in detail in this thesis.

There are also two primary categories of circuits to consider, those that are sensitive to local variation and those that are not. The line is not necessarily black and white between them, but many circuit applications fall clearly on one side of the spectrum or the other. It is important to identify these circuits up front because Monte Carlo simulations, which are used to study circuit response to local variation, are not always practical to run on full-chip simulations.

Circuits that tend to be bottlenecks for data transfer such as IO's, sense amplifiers, or differential amplifiers can be highly sensitive to local variation. These circuit blocks will be referred to as 'bottleneck' circuits. Experienced designers have already been considering the effects of local variation or mismatch in addition to the inter-die corners on these circuits for many years. These circuits tend to be analog in nature but that is not always the case. Any circuit block that depends on a single stage or pairs of similar stages can be susceptible to local variation. Two identical logic paths that are required to produce the same delay after a given number of stages are subject to local variation and cannot be expected to produce absolutely identical outputs. A delay chain of combinatorial logic will have slow typical and fast gates due to local variation. The longer paths will tend to have equal number of slow and fast gates that average out the local variation and produce a total delay that is proportional to the average delay. Shorter paths will have more variation in absolute delay through the chain since the number of slow and fast gates will not always be equal. The magnitude of a few highly skewed delays can have a larger impact on the total delay when the path is short. These shorter paths can be categorized as being sensitive to local variation and will be discussed in more detail in Chapter Four. In many ways, they can be considered bottlenecks as well.

Long strings of asynchronous combinatorial logic will tend to average out the effects of local variation. A simple inverter string for example can have a more equal number of fast and slow gates as long as the string is long enough. The greater the local variation of each inverter, the longer the string needs to be to average out the variation. The delay through these blocks can be categorized as having an 'averaging' response. They can be relatively insensitive to local variation. Even the measured delay through a simple ring oscillator with a minimum number of stages can be quite immune to local variation even when using the minimum sized gates with relatively high local variation.

Die or circuit-level standby leakage can also be considered an averaging mechanism since multiple devices contribute to the output at the same time, thus averaging out the local variation within the block. Half of the devices will have a threshold voltage below the mean and the other half above the mean. However, the average value is not centered on the inter-die model (i.e., TT) due to the lognormal nature of sub-threshold leakage. The devices with threshold voltage values lower than the mean will have more weight since a normal Gaussian threshold voltage distribution will produce a lognormal leakage spectrum. Since standby leakage is affected by local variation, it does not fall in one category or the other, but rather somewhere in the middle. It is, however, important to note that the mean value is quite predictable. Instance specific models such as those in a Monte Carlo analysis are not necessarily required to understand the impact of local variation. Models can be built at the local variation mean in order to capture the appropriate leakage or it can easily be hand calculated if the total width at a particular length that contributes to the standby current is

understood. This is an important subject and will be discussed in more detail in Chapter Four.

The inter-die corner models are usually generated for each unique case that needs to be simulated using static models. Model variables can also be parameterized so that circuit designers can simulate performance at various sigmas or at intermediate corners as the application demands. Some may consider these model variants 'statistical' models and in some ways they are; however, every instance of a particular model in the simulation netlist still has the exact same performance. No random intra-die performance is evaluated in this simulation despite the statistical connotation. Parameterized statistical models are still considered inter-die models and are meant to capture the chip-to-chip variation, not the random local variation. A circuit designer needs to understand what the statistical models are providing. These models can also capture various regions of NMOS and PMOS variability behavior like the slow fast (SF) or typical slow (TS) regions as required. As long as the NMOS and PMOS performance is highly correlated, then SS, TT, and FF corners would be all that are needed. The compact modeling engineers will fit the response of high-volume data from the production fabs to set these corners based on the correlation coefficients. As long as the circuit design only uses a few devices, this can be easily provided and simulated. However, many technologies now offer low and high-voltage devices as well as a variety of threshold voltage variants for NMOS and PMOS devices. The number of inter-die corner models required can increase quickly when multiple transistor variants are introduced. If the NMOS and PMOS devices do not correlate well, then the design teams may need to simulate using all 9 permutations of slow, typical, and fast models. If extra devices were introduced,

and again did not correlate to existing devices, then the number of required corners increases rapidly. For example, consider a process that supports low and high-voltage transistors as well as a few threshold voltage variants such that there were 6 unique transistors, all uncorrelated. The corner model name might be STTTTS or FTSFTT and there would be 3^6 or 729 possible corner simulation combinations. Of course, if they all were perfectly correlated there would only be three inter-die corner models needed, namely SSSSSS, TTTTTT, and FFFFFF. The device may tend to share implants and only differ slightly. The required number of inter-die models would fall somewhat higher than 3 but hopefully nowhere near 729. It might also be uncommon for a particular circuit to have all 6 models. A designer would obviously only need to simulate the corners for the devices within the circuit block of interest. Again, these models do very little to support local variation within the die since they are based on the chip to chip variation and are only statistical in nature from a chip-to-chip, wafer-to-wafer, and lot-to-lot perspective.

Consider the case in which a relatively long block of combinatorial logic that is insensitive to local variation is being simulated. The models are being built based on the variation data from individual devices that exhibit a 25% increase in variation due to random local variation. The inter-die corners are incorrectly set based on the total distribution of individual devices, without separating the local and non-local corners. The circuit design would be 25% better in silicon than simulations predict because the extra local variation would be averaged out. There may have been changes that could have been implemented, to reduce die size or save power, that were wasted on efforts to meet specifications at the overly pessimistic slow and fast corners.

Now consider the case in which a circuit response is sensitive to matched timing paths and the same models were used, which did not separate local and non-local variation. The circuit designer may have taken care to match the parasitic and device sizes in the matched paths, but the random variation could have been a significant source of timing mismatch. Every device in the simulation would still be identically matched at all available corners. The slow corner would be exactly the same for both paths and the fast corner would be equally fast in both paths. The total variation would be pessimistic, but the circuit may fail a specification or miss a set up and hold margin on a percentage of the die due to local variation.

A useful statistical model will include the effects of both the random intra-die and systematic inter-die variation. The model-to-model correlations will also be included and the joint probability of multiple random processes would be encompassed in a Monte Carlo or directed Monte Carlo approach. This is no easy task and requires a significant amount of data collection and modeling on very stable silicon before the models can be properly implemented. This can be an even more difficult task while the process is under development during the early circuit design phase.

Failing to recognize and react to local variation can result in both under design and over design. The compact modeling engineers, process integration engineers, and parametric characterization engineers need to work together to develop the proper test structures and a sampling plan to be able to separate the random intra-die variation from the inter-die variation. Circuit designer also need to understand how their circuits are sensitive to each type of variation in order to know which models to simulate and how to interpret the results.

CHAPTER THREE – MISMATCH CHARACTERISTICS

3.1 Characterization Techniques and Challenges

Perhaps the most common way to quantify mismatch or local variation for a given technology is to measure the difference in behavior between two identically matched devices placed next to each other. The difference in behavior across many samples can then be studied across multiple geometries, implants combinations, and process conditions on a given technology. The variation of the difference between these matched pairs is larger than the individual variation by a factor of $\sqrt{2}$ that arises from the difference of two random independent variables. Equations 15 and 16 relate the local variation to the difference between the pairs where the local variation of device A is assumed to be identical to the local variation of device B (an identical pair). This factor of $\sqrt{2}$ is not always accounted for when reporting A_{VT} values in the literature. A_{VT} is generally reported from sigma delta *VT*, but designers should consult the modeling engineers to make sure they are accounting for the variation correctly in simulations.

$$\boldsymbol{\sigma^2_{\Delta VT} = \sigma^2_{localA} + \sigma^2_{localB}} \qquad \textbf{Eq. 15}$$

$$\boldsymbol{\sigma_{\Delta VT} = \sqrt{2} \cdot \sigma_{local}} \qquad \textbf{Eq. 16}$$

Care must be taken to ensure that the layout of the test structure and wire connections do not impact the measured results. A wider device is more sensitive to

interconnect and probe tips resistance than a narrower device; Kelvin style connections can be used to cancel out external resistance effects. Larger area devices have relatively low mismatch, which can be sensitive to instrument resolution and repeatability limitations and give rise to non-zero intercepts for A_{VT} extractions [12]. Circuit designers should be very weary of mismatch data with a non-zero intercept such as that shown in Figure 8. The characterization process could have introduced the offset. Larger samples or improved repeatability might be required on the larger devices in order to accurately predict sigma.

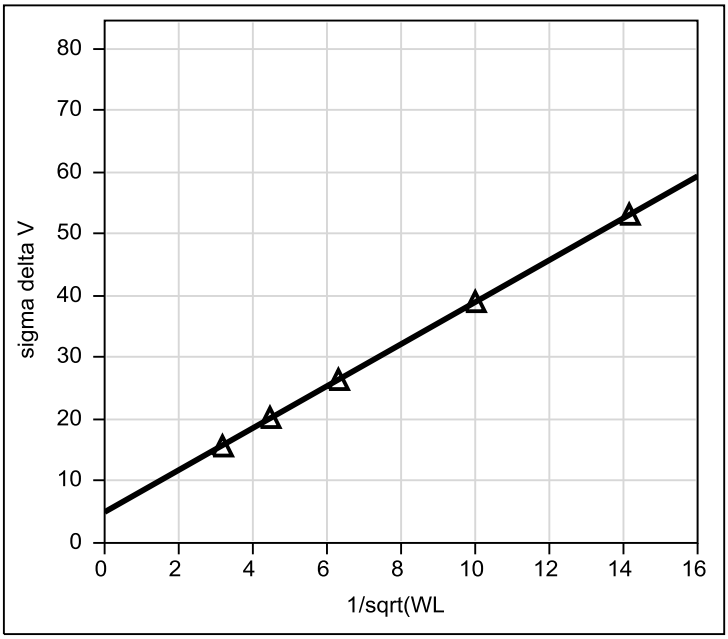

Figure 8. Sample Pelgrom plot showing a non-zero intercept that can arise when the resolution of the largest device is limited.

It is common to run wafer-level experiments to study mismatch, and many other process conditions. Care must be taken to ensure that the sample size is large enough to yield solid statistical data at the wafer level. If intra-wafer mismatch (i.e., center versus

edge) trending is to be studied, then each die must have enough identical samples to yield solid statistical results. This can be accomplished by building and testing multiplexed cores of identical devices. Placing multiple identical devices in multiple array cores also enables parallel testing, which can greatly reduce test times. If the gates are multiplexed and the source and drains are connected directly to adjacent pads, then wire resistance effects can be minimized. Avoiding pass gates on the drain and source will prevent the need to compensate for the body effect and series resistance effects [13]. The device arrays need to be kept small enough so that the total leakage from all devices does not disrupt the threshold voltage extraction routine. Banks of 16 or 32 work well because the off-state leakage is only 16 to 32 times larger than IOFF from a single sample and will not generally interrupt the target device currents. Lower VT devices will have less margin between the bank leakage and the target device current. A common threshold voltage extraction technique for highly scaled CMOS devices is a constant current threshold voltage extraction around 1 to100nA*W/L. If the extra leakage is within about an order of magnitude of the trigger, then the extraction routine can be altered to avoid errors. It might be better to use the common max transconductance extrapolation technique to extract the threshold voltage if the sub-threshold currents are altered by the added leakage; however the max transconductance technique generally has poor repeatability for larger devices due to probe tip resistance variation. This problematic leakage floor can also be reduced by passing an off-state gate voltage that is negative for NMOS devices and above *VDD* for PMOS devices. With sub-threshold slopes in the 80-100 mV/decade range, the leakage can be reduced by an order of magnitude with just a 100 mV gate voltage offset. With 200 to 300 mV's of offset voltage on the unaddressed

gates, the leakage current can be reduced another 1-2 orders of magnitude. If gate-induced drain leakage (*GIDL*) is the limiting factor for sub-threshold leakage, then applying these signals to the unaddressed gates may not provide a reduction in leakage. The increased drain-to-gate voltage will result in an increase in *GIDL*.

Having replicate devices is critical in the presence of random local variation. An experiment designed to look for subtle layout affects or geometry trending can easily be swamped out by the random local variation. When designing the test structures, it is important to anticipate the impact of local variation, and design the sample size accordingly. This is particularly true if the test is designed for detailed bench work where only a few sites can realistically be measured. The devices in a bench testable experiment must be drawn large enough to reduce local variation or utilize multiple devices in parallel to help average out the random variation. When fitting compact models for a width trend, the very narrow devices at nominal lengths tend to be the smallest devices measured and can suffer greatly from excessive local variation.

Suppose that a generic 50 nm CMOS process like the one use in Table 9.2 'CMOS Circuit, Design, Layout, and Simulation' by R. J. Baker [7] is being characterized. Suppose the narrowest device in the test structure width trend was 150 nm with a *W/L* ratio of 3/1. Suppose also that the *AVT* for this process was given by the Equations 13 and 14 in [8] where a 1.4 nm gate oxide thickness is expected to have an A_{VT} of 3.4. The 3/1 device would have a local threshold voltage mismatch variation of 39 mV's at 1 sigma ($\sigma\Delta VT = \frac{A_{VT}}{\sqrt{W \cdot L}}$). The local variation is smaller by a factor of $\sqrt{2}$ so $A_{VT,local}$ is 2.4 mV-um. The local sigma for an individual device is then 28 mV's. Figure 9 shows the expected local variation on this NMOS 50 nm device across width for 100

sites assuming the site-to-site variation is non-existent. Each line connects a possible site on a given wafer or die, and the mean value is 280 mV's. Figure 9 has a constant threshold voltage across width and Figure 10 shows a case with a 30 mV drop in the threshold voltage for narrow devices. The threshold values were generated using a random number generator based on the area of the devices. It is evident that any one site will not give enough resolution to resolve any subtle width trend effects.

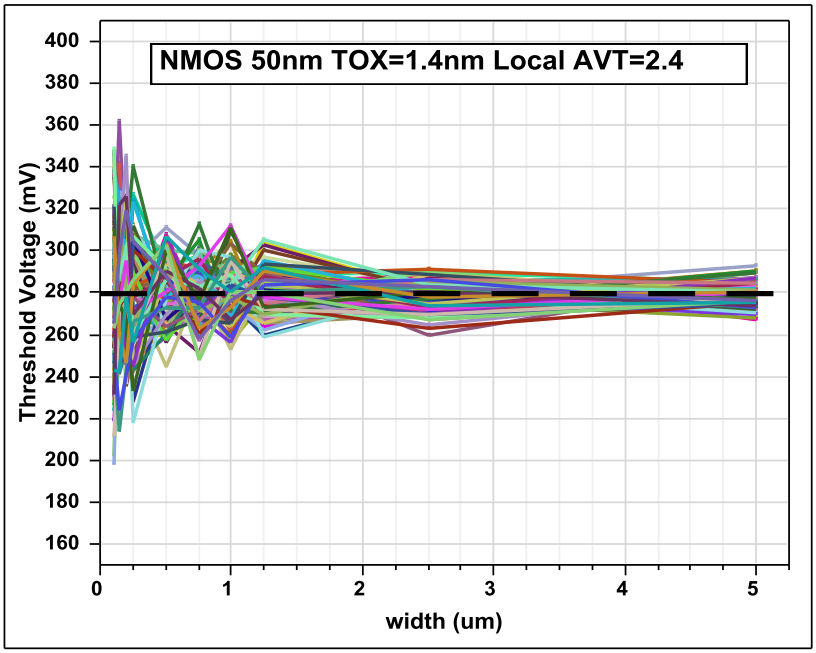

Figure 9. 50nm VT vs. width for 100 samples with a flat width response with an $A_{VT,local}$ of 2.4mV-µm.

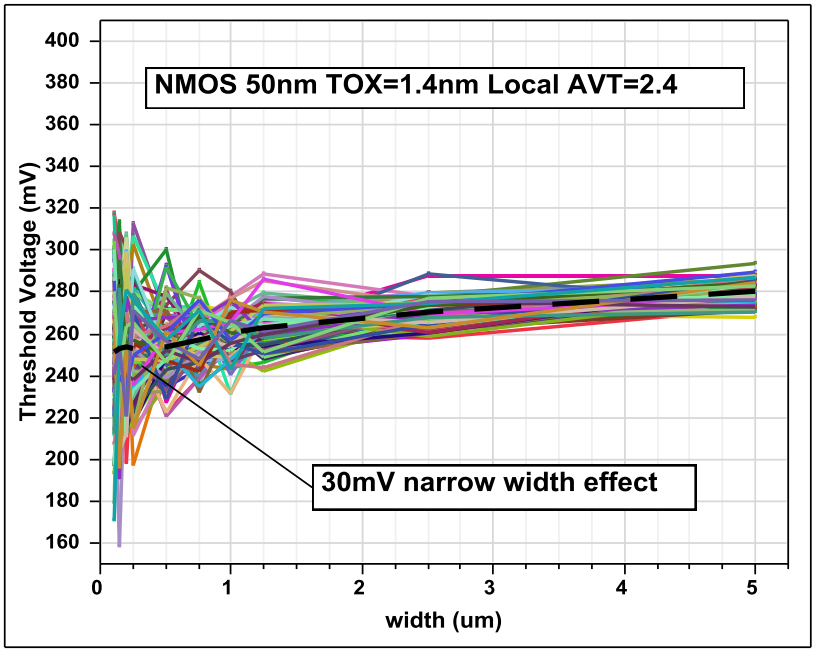

Figure 10. 50nm NMOS VT vs. width for 100 samples with a 30 mV drop in VT across width with an $A_{VT,local}$ of 2.4 mV-µm.

The number of replicate devices required to achieve a given tolerance can be estimated using tradition confidence intervals for a normal distribution. The confidence interval for the mean value of a given sample is estimated as $\Delta\mu = z \cdot \frac{\sigma}{\sqrt{n}}$, where z is the desired sigma interval (i.e., z = 1, 2, 3 corresponds to 68.3%, 95.4%, and 99.7% respectively), n is the number of replicates required and the known random sigma is given by σ. During a compact model fit, it is required to examine the threshold voltage trend across width for a fixed length. The narrowest devices in the trend at nominal lengths can pose a significant characterization challenge. The required sample size or number of replicate devices required to provide 95.4% confidence (z=2) in the mean of the sample would then be calculated as $n = \left(2 \cdot \frac{\sigma_{local}}{\Delta\mu}\right)^2$. The mean threshold voltage in

the typical model provided for the 50 nm process in [7] is 280 mV's. If it is required to resolve the mean threshold voltage of a sample to within 14 mV's or about 5% of the actual mean ($\frac{\Delta\mu}{\mu} = 5\%$), then the number of replicate devices required is 16. This means that we must either measure 16 sites or that we must design the experiment to have 16 replicate devices at each site. Placing 16 replicates at each site is the better option since site-to-site variation across a wafer can also introduce significant variation. Figure 11 shows how the sample size impacts the accuracy of the sample mean for the same 50 nm NMOS device with a threshold voltage of 280 mV's and an A_{VT} of 3.4 [13]. The expected A_{VT} for the 50nm PMOS device in this generic process according to industry trending from [8] is about 2.55 ($A_{VT,local}$ = 1.8) and the mean threshold voltage is also about 280 mV's. The same PMOS trend would require 9 samples to gain 95% confidence in the measured sample mean.

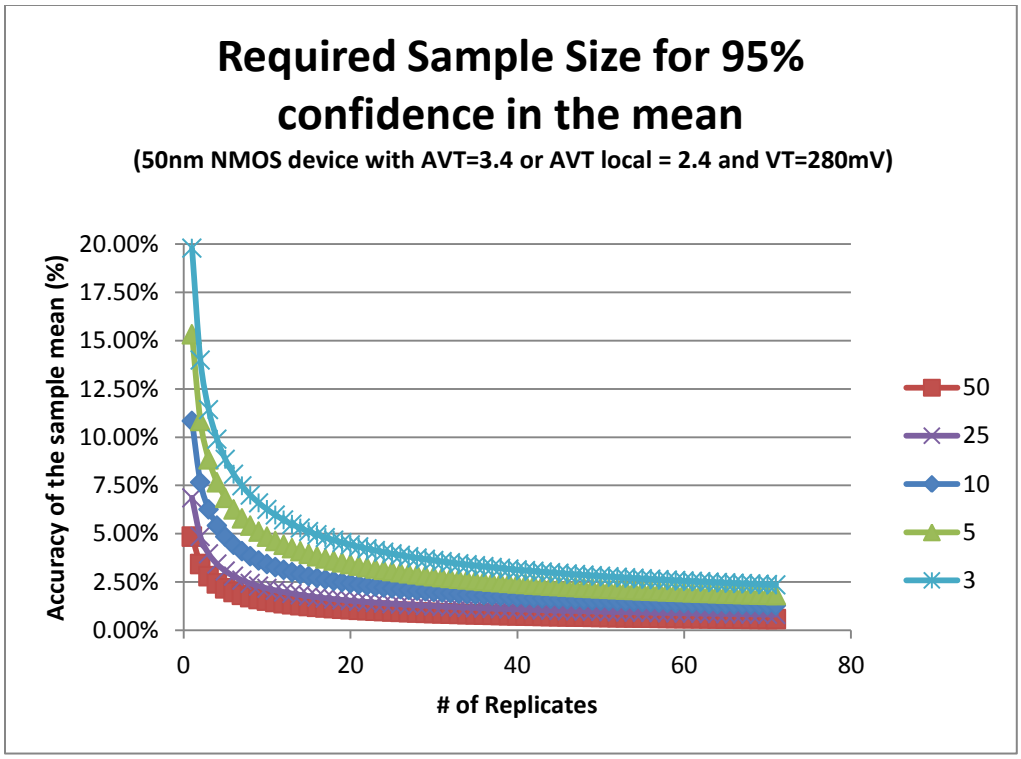

Figure 11. The accuracy of the sample mean across the number of replicate devices per site for various W/L ratios at L=50 nm illustrating increased sample requirements for smaller devices.

It is also interesting to note that a 150 nm length, 80 angstrom NMOS device is expected to have an A_{VT} of 10 [8]. The same 3/1 *W/L* ratio device the local sigma is 27 mV's, which is close to the generic 50 nm NMOS device. However, the mean threshold voltage for the thicker device is likely closer to 600 mV's, therefore 5% of the threshold voltage is 30 mV's. The required sample size for 95% confidence in the mean is then just 4. It does not always make sense to consider a percentage when addressing threshold voltage variation. It might still be desired to have a voltage-based target, such as 10mV, for a confidence interval instead of a percentage-based interval. With a 10 mV expected

tolerance, the same 3/1 50 nm NMOS device would require 32 samples and the 150 nm, 80 angstrom NMOS device would require 30. The allowable tolerance could be based on the expected total variation for the device. However, the 80 angstrom device likely operates with a higher V_{ov}, making it less sensitive to the changes in *VT*.

When characterizing the local and non-local variation, it is important to break out the components of variance correctly and combine the effects appropriately. There are statistical software tools that perform components of variance analysis on sampled data but the main point to consider is that the local variation is independent of the die-to-die variation. The variances can be summed to predict the total variation as shown in Equation 17 below. Equal contributions of variance from local and die-to-die variation result in an increase in the total variation by a factor of $\sqrt{2}$. For example, a 10 mV sigma from each results in a total sigma of $10 \cdot \sqrt{2} = 14.14\ mV$. If the local variation portion of the total is eliminated, 10 mV's of total variation is left, which of course is not a 50% reduction in the variation. The uncorrelated variances are summed; the sigmas cannot be summed.

$$\boldsymbol{\sigma^2_{total} = \sigma^2_{local} + \sigma^2_{die\ to\ die}} \qquad \textbf{Eq. 17}$$

Figure 12 shows a sample set of randomly generated data with 1000 die, each with 1000 threshold voltage values with a sigma of 14.14 mV's at each die and from site-to-site such that the total sigma is 20 mV's. The statistical software tool by SAS, called JMP, was used to generate the data and Figure 12.

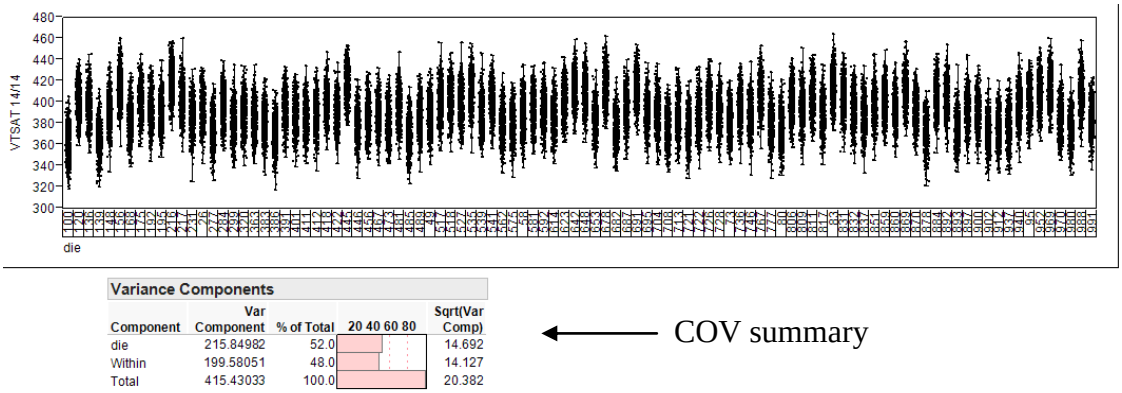

Figure 12. Threshold voltage samples showing local and die-to-die variation along with a components-of-variance analysis with 14 mV of die-to-die and within-die variation.

The summary table below Figure 12 shows that the extracted variance component is indeed 14.14^2 or about 200 mV's. The die-to-die and within-die variations contribute 50% of the total variation each and the sigma is denoted as the square root of the variance component at the end of the table with a total of 20 mV's. Measured data can be fed into an automated tool such as JMP to extract the variance components. Again, eliminating 50% of the variance does not eliminate 50% of the total sigma.

It should be evident now that the random local variation can introduce significant challenges for characterization work. The test-structure layout, design, and characterization plan need to include the impact from local variation. One cannot expect to resolve subtle layout effects or process changes without comparing the required tolerance to the expected local variation and adjusting the test-structure design and sample plan accordingly. Software tools such as JMP can help perform the needed analysis as required.

3.2 Mismatch Across Bias Conditions

Parametric extractions from IV curves such as the threshold voltage and Idrive are good tools for studying device behavior and they attempt to give us points on the curve that help describe the full IV characteristics. The drain, source, gate, and bulk currents respond to sources such as the voltage bias conditions, temperature, dopant atoms, oxide thickness, and interface states (to name a few). The matching behavior is often studied in terms of the extracted threshold voltage, or Idrive. These can be useful but it can be informative to study mismatch for an entire IV curve as well. This is often done when trying to explain the fundamental physical origins of mismatch, which was done in [4] , or when attempting to use back propagation of variance (BPV) techniques to model the mismatch, as was done in [14] . The work required to develop an accurate BPV model is extensive. This method essentially combines the model sensitivity to the measured variation and fits a sigma to each model parameter used. The more model parameters used, the better the fit. The method requires full IV curve mismatch data from multiple geometries for each model. The compact model cannot have bin boundaries within the measured geometry range and the sensitivity to each parameter must be physically accurate. This may sound like an obvious requirement for every compact model, but in reality there are many empirical parameters that are used to nip and tuck the models into place, which can skew the physical sensitivities. If the model sensitivities are accurate and free from bin boundaries, then each model parameter that was used in the analysis is given a unique variable that can be skewed using the extracted values via Monte Carlo analysis to simulate the desired device and circuit response.

The threshold voltage is generally extracted from a sweep of the gate voltage at a particular drain-to-source voltage (*VDS*). It is very common to report mismatch from the threshold voltage extraction with about a 50 mV *VDS*. This threshold voltage is generally referred to as *VTLIN*, denoting that it is measured in the linear region when *VDS* is low. If matching is being considered in digital circuits, then it is more accurate to extract the threshold voltage with *VDS* set to the power supply voltage. The *VDS* voltage is generally at or close to *VDD* when the gate is toggled in digital applications. This threshold voltage is commonly referred to as *VTSAT*. *VTSAT* is generally 0-250 mV's lower than *VTLIN* depending on the device length, technology, and voltage conditions. This difference is commonly referred to as *DIBL* and it is reported in units of mV's of *VT* shift per *VDS* in mV/V. The mismatch of *VTSAT* can be worse than the mismatch of *VTLIN*, particularly when the device is very close to punch-through [3], but in many cases can be negligible [4] even in the presence of substantial *DIBL*. *VTSAT* predicts I_{drive} and gate delays much better than *VTLIN* for short channel devices, therefore *VTSAT* mismatch needs to be considered when studying circuit response in digital applications.

The current factor beta (β) is also a significant source of drain-current variation. Beta mismatch has local and die-to-die components just like all other sources of drain current variation. Beta variation is less significant than *VT* variation for most technologies but not negligible. Beta is expressed in Equation 18 as a function of device geometry, oxide thickness, mobility, and bias voltage.

$$\boldsymbol{\beta = \frac{W \cdot C_{ox} \cdot \mu(VGS,VDS)}{L}} \qquad \textbf{Eq. 18}$$

Beta mismatch needs to be kept separate from *VT* mismatch. Modeling and characterization engineers should specify whether the beta mismatch being reported includes the *VT* variation or if it was decoupled. Mismatch is the drain, and source resistances also play a role in the current factor mismatch by altering *VGS* and *VDS*. It is difficult to separate the mobility fluctuations from the series resistance fluctuations, but Kelvin style test structures can help identify the root sources of variation. Beta variation can be separated from *VT* variation by modulating the gate voltage of an I_{drive} extraction by the shift from the mean *VT*. In this manner, I_{drive} is normalized for a constant overdrive voltage. If I_{drive} is not quantified at a constant V_{ov}, then including the effects of beta variation and *VT* variation would be double counting the effects of *VT* on the drain current.

The substrate or nwell voltage also modulates the mismatch behavior. The threshold voltage increases as the magnitude of the substrate-to-source voltage (*VBS)* increases. The depletion width widens, encompassing a larger region of silicon with independent dopant atoms. This added region increases the *VT* mismatch proportional to Equation 19. Changes in the gate oxide capacitance can also play a role in the substrate voltage sensitivity, but the dopant fluctuations are found to be the dominant source of mismatch [4].

$$\sigma^2_{\Delta Vt,doping} = \frac{t_{ox}^2\sqrt{8q^3\varepsilon_{si}N_A(\varphi_B - V_{BS})}}{3WL\epsilon_{ox}^2} \qquad \textbf{Eq. 19}$$

An empirical model for the body effect of *VT* mismatch was proposed in [4], and shown in Equation 20 where α is a fitting parameter in the range of 0.3 for long channel devices and 0 for short channel devices.

$$\boldsymbol{\sigma_{\Delta VT}(VBS) = \sigma_{\Delta VT}|_{V_{BS}=0} \cdot \left(1 - \frac{VBS}{\varphi_B}\right)^{\alpha}} \qquad \textbf{Eq. 20}$$

The impact of the body bias on mismatch will depend on the device geometry, so fitting α across geometry adds complexity to the threshold voltage mismatch predictions. Understanding how mismatch responds to bias conditions is useful for reporting behavior, but difficult to implement in the dynamic simulation environment. *VTSAT* and Beta mismatch in strong inversion can generally cover the primary behavior in digital and analog circuits and greatly simplifies the modeling and simulation efforts.

These device-level details are useful for understanding root issues and can offer great insight for critical bottleneck circuits that are sensitive to mismatch. Differential amplifiers, voltage regulators, and other mismatch sensitive circuits can be greatly improved by understanding how they behave under different bias conditions. Avoiding a *VBS* potential, for example, can help reduce additional mismatch. However, modeling the dynamic behavior of the mismatch in transistor-level simulations is certainly more challenging.

3.3 Temperature Dependence of Mismatch

The threshold voltages of CMOS devices are higher at cold temperatures and lower at higher temperatures. This begs the question, is the variation itself a function of

temperature or is the variance constant across temperature? The temperature behavior of mismatch was studied and modeled in [15] and [16] show that mismatch at lower temperatures is worse than the higher temperature mismatch for both threshold voltage and drain currents. However, this research also showed that not every device had less *VT* and *IDS* mismatch at higher temperatures. A fraction of the samples had more mismatch at higher temperatures, but the majority of the devices tended towards having a reduction in mismatch, which results in a tighter sigma as temperature is increased. No two devices have exactly the same temperature sensitivity. Figures 13 and 14 show how the *VT* and the delta *VT* might vary across temperature.

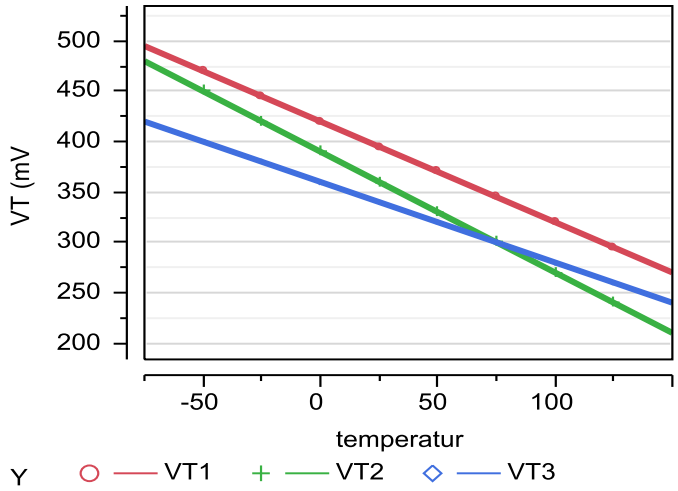

Figure 13. Possible VT variation across temperature for three random samples.

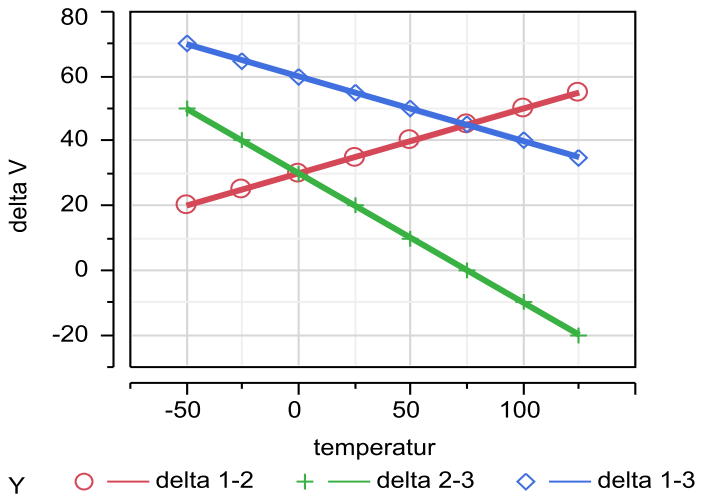

Figure 14. Delta VT across temperature from the devices in Figure 13, showing the statistically rare case with an increase in variation as temperature increases.

The temperature affect in these papers only account for a small fraction of the total mismatch, but this type of analysis needs to be done by the foundry and compact modeling teams in order to better understand the mechanism for a given technology. The 60 nm NMOS device used in [16] had a sigma delta *VT* of 19.5 mV's at 0C, which dropped to 18.6 mV's at 100C. Of the 4 device regions tested (NMOS and PMOS at L=60 nm and 1 um) none moved more than 1 mV. The papers do not mention which threshold voltage extraction technique was used to derive sigma delta *VT*, but did show how the currents across *VGS* changed as temperature increased. This suggests that sub-threshold currents are more sensitive to the temperature changes than the currents in saturation. The max transconductance *VT* extraction method is sensitive to changes in mobility (as well as external resistance); therefore, a reduction in mobility at higher temperatures would result in a lower extracted threshold voltage. Suffice it to say the characterization of the temperature-dependent portion of the mismatch is quite a bit more

challenging to measure successfully than the mismatch itself. The number of samples required to characterize this additional nuance of mismatch is much higher than the number required at a single temperature. This subtle temperature affect may seem insignificant, but it can really hurt sensitive circuits like bandgap references, or other closely trimmed circuits that are designed to meet tight criteria that depends on good mismatch. The temperature affects can set the lower limit for the best achievable behavior in such circuits. Maintaining tight performance as the devices degrades over time is yet another challenge. It is also possible that the temperature dependence changes as the device degrades. This is a possible research subject.

3.4 Reliability Induced Variation

Reliability is another very hot topic in highly scaled CMOS devices. Negative bias temperature instability (NBTI) and channel hot carrier (CHC) degradation are two primary CMOS degradation mechanisms challenging device engineers and circuit designers. Each results in an increase in the threshold voltage over time as the device is used. NBTI occurs primarily on PMOS devices when the nwell, source, and drain are all at VDD and the gate is turned on with 0 V (i.e., after a digital pull-up event). CHC generally occurs during switching when current is flowing from drain to source resulting in impact ionization and the generation of hot carriers that get trapped in the oxide near the drain edge.

It is important to consider reliability when discussing variation because it affects the design space in much the same way that the process variation affects the design space. Device degradation adds another dimension to the variability concerns for two reasons. First, the voltage or use conditions applied to matched devices may not be identical,

which can result in varying amounts of degradation to different devices. Second, even under identical stress conditions, the devices may degrade at different rates and produce additional offsets. Again, no two devices are identical and thus they will not degrade exactly the same under similar conditions. This also poses a challenge for reliability characterization, which will require more samples to determine the mean reliability behavior. A conservative design might consider the max reliability induced variation instead of the mean degradation rate.

Suppose a pair of intrinsic devices could only tolerate 5 mV's of mismatch before a circuit failure and that the devices were sized accordingly to meet the requirement. Suppose also that one of the devices in the pair was held in a stress condition that induced NBTI, perhaps in a standby mode of operation with a DC NBTI stress. The device under the NBTI state would degrade and induce additional mismatch between the pairs. The allowable threshold voltage shift due to NBTI for a device like this would be just a few millivolts. NBTI and CHC tend to follow a power-law relationship as they degrade, therefore a 50 mV lifetime might be met at 10 years but it may have degraded to 5 mV's in just a few weeks of use. A 5 mV NBTI lifetime would be extremely hard to meet under operating conditions if the 10-year specification was indeed 50 mV's.

Suppose now that the devices did see exactly the same stress conditions and that they degraded a fair amount. The rate of degradation for each device will not be the same, and additional mismatch will be introduced [17] [18] [19]. This produces another source of variation to consider during the circuit design phase. It also introduces yet another characterization and modeling challenge. Figure 15 shows an example of what the degradation might look like for a matched pair over time.

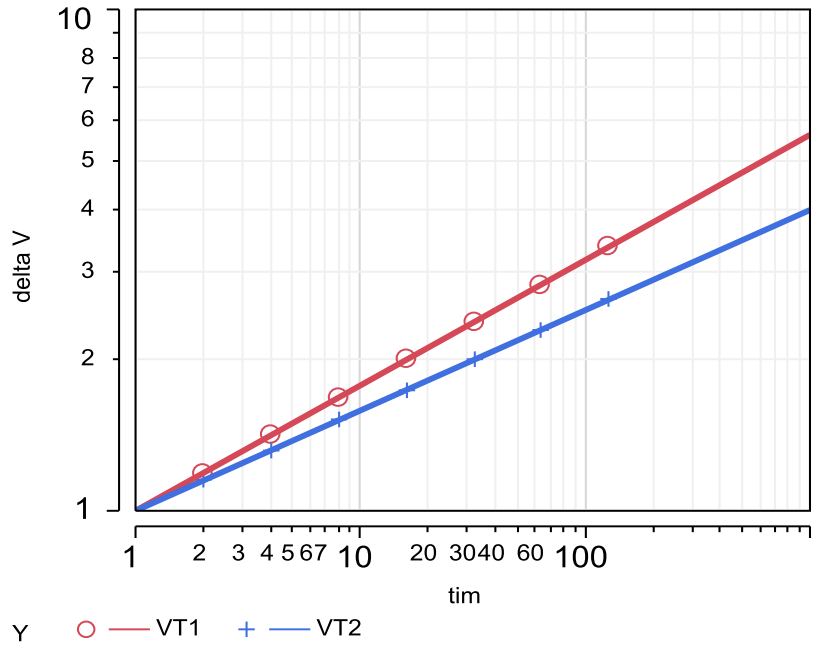

Figure 15. Possible VT shifts over time due to CHC or NBTI degradation for a matched pair of devices illustrating possible divergence.

Mismatch can be characterized on pairs of devices as they are degraded in order to quantify the reliability induced mismatch. This reliability induced mismatch could be very problematic for bottleneck circuits with closely matched or trimmed devices, but are not a likely challenge for averaging topologies such as combinatorial logic blocks.

Reliability or aged models are typically provided by foundry compact modeling teams that model the degraded device performance, but this does not likely capture the impacts of reliability induced mismatch as the degradation occurs. The designer will have to understand the operating condition well enough to place the aged models on the appropriate instances so that they can properly simulate the circuit response with the appropriate devices being degraded. If the designer ensures the operating conditions are equal for matched pairs or matched circuits, then they must

determine if the CHC or NBTI-induced variation provided by the foundry exceeds the circuit tolerance. NBTI is one of the most challenging sources of degradation to deal with because there is very little a circuit designer can do to reduce the degradation since it is not very sensitive to device geometry. The effects of CHC can generally be reduced by increasing channel lengths on nodes with slow rise times in order to reduce the impact ionization at the drain edge. Designers can avoid holding devices in an NBTI state during standby conditions if possible, but even during switching the PMOS devices will be in an NBTI state for a period of time and will degrade. Device degradation adds a time-dependent variability component that has to be considered. Process engineers have to work to reduce these mechanisms at the operating voltages, but circuit designers also have to be diligent in understanding and simulating the weakest links.

3.5 Random Variation in Transistor Noise

Low-frequency drain-current noise is another significant challenge in highly scaled CMOS transistors. This noise generally follows an inverse relationship with frequency and is referred to as 1/f (one over f) or flicker noise. The noise arises from fluctuations in the conductivity or mobility of the channel [20]. The fluctuations in mobility originate from trapping and de-trapping of carriers as they flow from source to drain [20]. The trapping and de-trapping of charge can be modeled as a change in the threshold voltage that modulates the channel conductivity. It was clearly illustrated in [20] that smaller devices have more noise variation from device to device than larger devices. In other words, the noise levels themselves vary greatly between otherwise similar devices. The variation or dispersion of the noise was found to be proportional to

1/sqrt (area), where smaller devices show a much wider range of noise variation than larger devices [20]. No two dielectric interfaces are exactly the same; therefore, we expect differences in noise performance between adjacent devices. Proper sampling techniques need to be followed for flicker noise characterization when the area of the device is small.

CHAPTER FOUR – IMPACT TO CIRCUIT DESIGN

4.1 Simulation Techniques and Challenges

Many CMOS applications require both analog and digital circuits where the use of transistor-level spice simulation tools are needed for increased accuracy and gate-level logic simulators are needed for efficient simulation times. Full-chip simulations using spice can take hours or days to complete; therefore running a Monte Carlo-style analysis can be prohibitive. However, the Monte Carlo analysis provides the needed statistical approach to study the joint probability of multiple random events occurring at the same time. For example, a simple matched pair that is sensitive to mismatch can easily be studied by skewing the performance of the devices individually. A single simulation or just a few manual iterations can uncover the worst case response quite easily. However, consider a larger block of devices with digital and analog circuits combined. Identifying the worst case scenario might not be obvious to the designer; therefore, a Monte Carlo analysis might be the best approach. This is particularly true for random local variation, but can also be useful for die-to-die variation when multiple uncorrelated device types are used in the same simulation. For example, consider the case where multiple uncorrelated device models are used in a circuit block. The devices in the circuit do not have a high probability of being at the slow corner at the same time. Intermediate corners may be needed for each device. As discussed in Chapter Two, the number of corner simulations increases quickly as the number of devices increases. Analysis has to be done to determine which corners are the most likely, but that does not mean that a circuit will not

have marginality at a less probable corner. For example, a circuit could contain a low voltage and high voltage set of NMOS and PMOS devices (4 models). Bringing both NMOS devices to the fast corner at the same time might pass specifications for all PMOS corner variants, but what happens if a marginality occurs when the low voltage and high voltage NMOS devices do not correlate (one is slow and the other is fast). The situation gets much more complicated as the number of unique uncorrelated devices in the netlist increases.

Statistical models such as those developed at IMEC in [21] can bridge the statistical gaps in a traditional corner-model methodology. The simulation methods described in this paper enable a robust statistical approach to circuit design. After thorough characterization, the tool places a voltage source on the gate and a current-dependent current source from source to drain to simulate the threshold voltage and beta variation. The voltage and current source are geometry specific and are unique for local and die-to-die variation. These ideal sources work well around any compact model. Large sets of data containing die-to-die and inter-die samples for each model are fed into the tools. All of the unique device correlations are captured and the random intra-die and systematic inter-die variation can be broken out and studied independently or as a combined global effect. A directed or weighted Monte Carlo analysis is used, for which each input vector has a probability weight associated with it. This allows the simulation to reach 5 and 6 sigma levels without having to simulate millions of vectors to see the tails of the response. This type of modeling approach requires a dedicated suite of test structures and characterization tools, as well as stable silicon. It is more difficult to utilize during the developmental stages of a technology since the silicon data may not be

available. The simulation time is longer than a simple corner methodology, but the benefit from device-to-device correlation and the addition of local variation analysis should outweigh the time hit and add a level of confidence to the simulation results. The methodology can also be applied to logic gate-level simulations, enabling full-chip statistical Monte Carlo simulations. Tool vendors are providing options for simulating local variation with more accuracy and efficiency because the demand from foundries and circuit designers is significant. Many of the digital simulators require a statistical compact model as a baseline for building statistical standard cell libraries. Improved solutions for simulating the local and non-local variation are needed and this is a ripe area of research and tool development, which will likely make great strides in coming years.

4.2 Sub-Threshold and Die-Level Standby Leakage

Leakage currents in MOSFETS (IOFF) follow a lognormal distribution when the threshold voltage varies with a normal Gaussian distribution with a fixed sub-threshold slope. The long channel sub-threshold current is governed by the exponential Equation 21 below, from [6] which describes the diffusion current.

$$I_{sub-threshold} = \mu_{eff} C_{ox} \frac{W}{L} \sqrt{\frac{\varepsilon_{si} q N_a}{4\psi_B}} \left(\frac{kT}{q}\right)^2 e^{\frac{q(V_g - V_t)}{mkT}} \left(1 - e^{\frac{-qV_{DS}}{kT}}\right) \qquad \textbf{Eq. 21}$$

This Equation can be simplified to the root exponential in Equation 22.

$$I_{sub-threshold} \approx e^{\frac{(V_g - V_t)}{m\frac{kT}{q}}} \qquad \textbf{Eq. 22}$$

In short channel device, *VTSAT* should be taken at a full drain to source voltage so that *DIBL* is captured. The variable *m* in Equation 21 is the body effect coefficient, which is proportional to the effective oxide thickness and the max depletion width and typically falls between about 1.1 and 1.4 [6]. A thinner oxide provides better channel control for a given channel doping concentration. The sub-threshold slope is usually reported in units of mV/decade, which is actually the inverse of the slope and sometimes referred to as swing instead of slope. The sub-threshold slope will be reported in mV/decade in this paper. The slope can be derived from Equation 21 and is shown in Equation 23 where a factor of ln(10) is added to convert to $\log_{10}$ (decade portion of mV/decade). With m=1 and T=300K, the ideal sub-threshold slope is 60 mV/decade.

$$\boldsymbol{Slope\left(\frac{mV}{dec}\right) = m \cdot \frac{kT}{q} \cdot \ln(10)} \qquad \textbf{Eq. 23}$$

The sub-threshold slope generally runs in the range of about 80 mV/decade for a reasonable short channel device where m is about 1.3, but can vary from about 70 mV/decade to greater than 100 mV/decade. If sub-surface punch though occurs then Equation 21 is no longer valid and the slope can easily exceed 100 mV/decade.

Figure 16 illustrates the relationship between the threshold voltage and the sub-threshold current near VGS=0V. Notice that the sub-threshold slope is 80 mV/decade and that an 80 mV shift in the threshold voltage shifts IOFF (IDS@VGS=0) by 1 decade in either direction. The IOFF leakage with a 390 mV VT is 1 pA/um, it will be 10 pA/um with an 80 mV drop in VT, and 0.1pA/um with an 80 mV increase in VT.

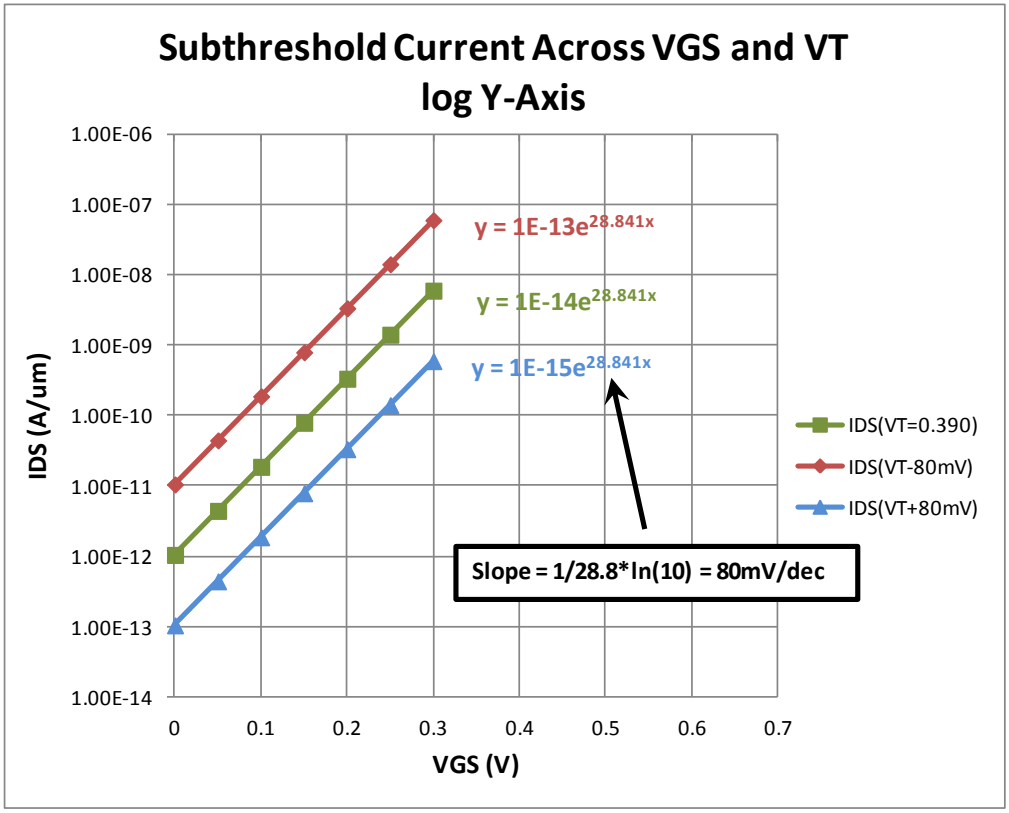

Figure 16. Ideal sub-threshold characteristics with a log Y-AXIS and a sub-threshold slope of 80 mV/decade showing a 1 decade increase and decrease in IOFF as VT shifts by plus and minus 80 mV's.

Figure 17 shows the same data zoomed in on a linear Y-axis showing that the 10X increase in *IOFF* causes a "tail" or a skew towards the higher leakage side as the threshold voltage decreases. This "tail" is the result of the expected exponential relationship with the threshold voltage that produces a lognormal *IOFF* distribution with a normal *VTSAT* distribution.

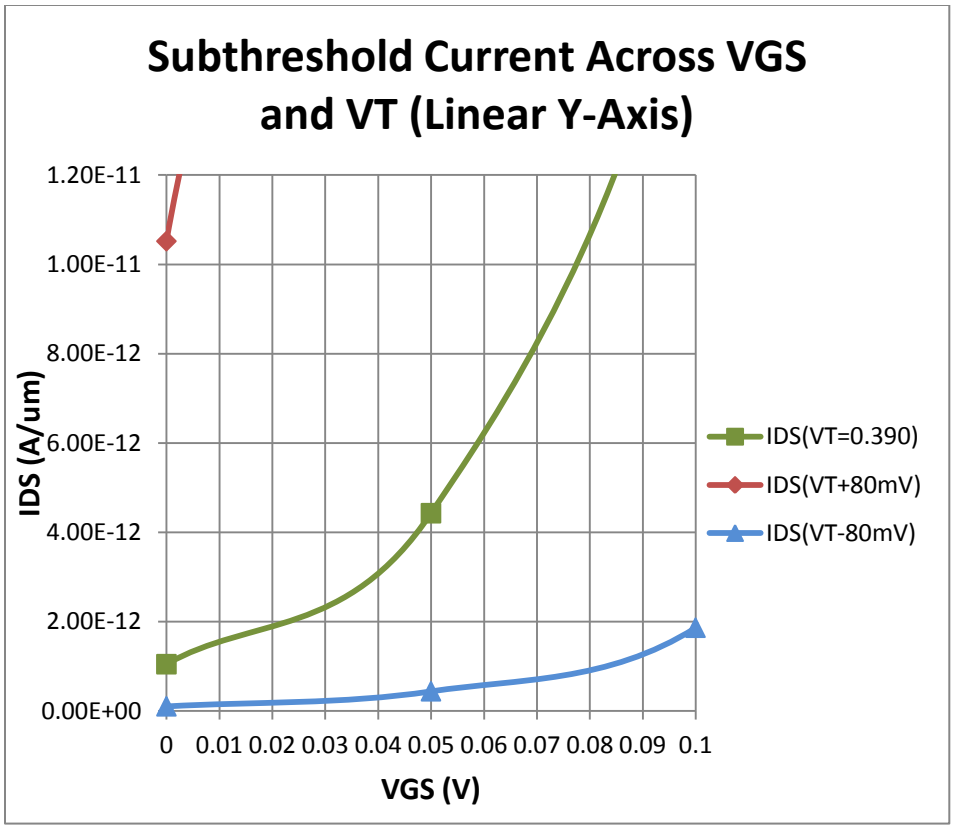

Figure 17. Ideal sub-threshold characteristics repeated from Figure 16 on a linear Y-AXIS, illustrating the exponential behavior of IOFF.

If the slope is constant for a normal *VTSAT* distribution, then a log transformation of the leakage current will also be normal. If the slope is not constant or if the *VTSAT* variation is not normal, then the transformation will be skewed. The log-normal distribution is often described by its sigma and mean in log space after taking the log of each value in the distribution [22]. Figure 18 shows the expected lognormal *IOFF* distribution from the VTSAT distribution centered at 390 mV's with a slope of 80 mV/decade. The transformed *IOFF* (ln(*IOFF*)) is also shown with a mean and sigma of -27.57 and 0.595.

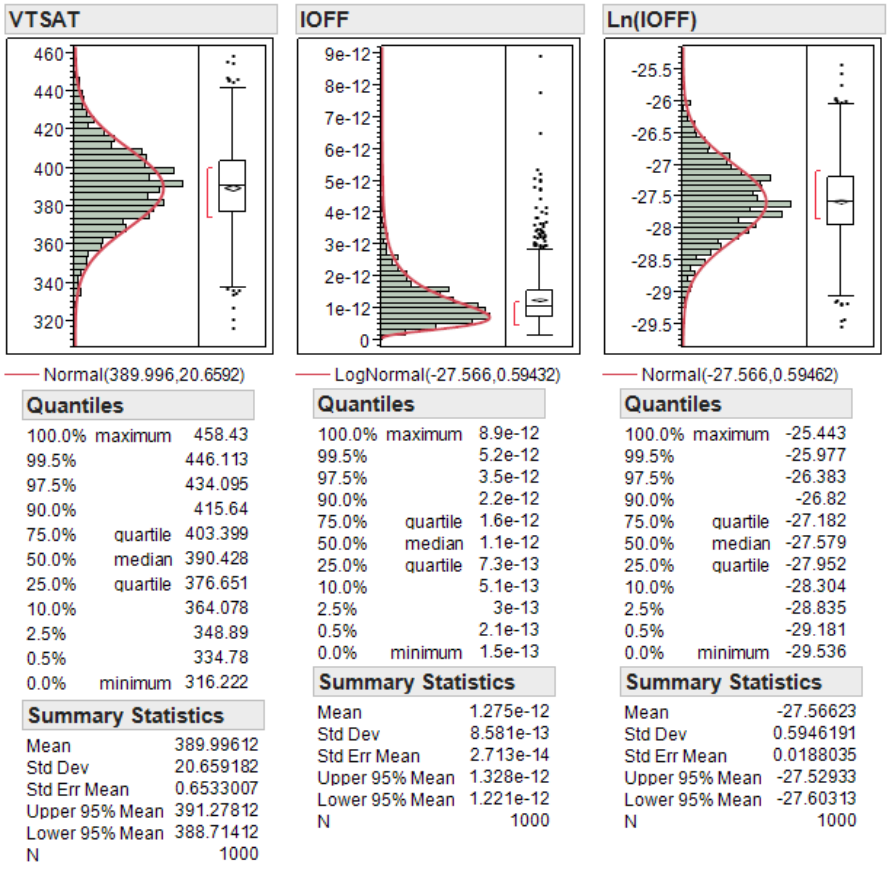

Figure 18. IOFF (center) and Ln(IOFF) (right) distributions arising from a normal VTSAT distribution (left).

Understanding the lognormal nature of *IOFF* with respect to *VT* is important in understanding how random local variation affects standby currents for large samples. It was suggested in Chapter Two that a large number of devices will act to average out the effects of local variation. This is indeed the case, but the average of a lognormal distribution is above the median value and does not occur at the mean of the normal *VT* distribution. The TT model will never predict the mean standby current of a large sample, which is a common misconception. This is always true whether or not local variation is present and is not always intuitive to a circuit designer. Equation 24

determines the mean of a lognormal distribution when the mean and sigma are known from the natural log transformation of the data [22] . Some foundry compact models include a model at the mean *IOFF* point, others rely on the designers to properly predict it.

$$\mu_{IOFF} = e^{M+\frac{S^2}{2}} \; where \; M = \mu_{transformed} \;\&\; S = \sigma_{transformed} \qquad \textbf{Eq. 24}$$

The relationship between the full *VDS* threshold voltage, *VTSAT*, and *IOFF* for a given sub-threshold swing is given by Equation 25.

$$IOFF = A \cdot e^{\frac{-VTSAT}{slope/\ln(10)}} \qquad \textbf{Eq. 25}$$

$$where \; A = \mu_{eff} C_{ox} \frac{W}{L} \sqrt{\frac{\varepsilon_{si} q N_a}{4\psi_B}} \left(\frac{kT}{q}\right)^2 \; and \; slope \; is \; given \; in \; mV/dec$$

Therefore, if we know *VTSAT* and slope, we can determine *IOFF*. Furthermore, if we know how the threshold voltage varies, we can estimate the impact on the mean *IOFF*. Equation 26 can be used to relate the mean *IOFF* to *VTSAT*, and Equation 27 relates sigma to *VTSAT*. Equation 28 gives us the mean *IOFF* value due to *VTSAT* variation.

$$M = \ln\left(A \cdot e^{\frac{-VTSAT}{slope/ln(10)}}\right) \qquad \textbf{Eq. 26}$$

$$S = M - \ln\left(A \cdot e^{\frac{-(VTSAT - \sigma_{VTSAT})}{slope/ln(10)}}\right) \quad \textbf{Eq. 27}$$

$$\mu_{IOFF} = e^{M + \frac{S^2}{2}} \quad \textbf{Eq. 28}$$

The mean *IOFF* for the 20 mV *VTSAT* sigma from Figure 18 resulted in factors of M = -27.57 and S = 0.595 and has an estimated mean value of 1.27e-12 using Equation 28. The IOFF plot in the center of Figure 18 indeed has a mean value of about 1.27e-12 when sigma *VTSAT* is 20 mV's. Figure 19 shows how *IOFF* varies across slope and sigma for the same 390 mV example. Figure 19 appears to suggest that a larger slope results in more variation but that is certainly not the case. The steeper slope is more sensitive to changes in the threshold voltage; Figure 20 illustrates the percent change in the mean *IOFF* as a function of sigma *VTSAT*.

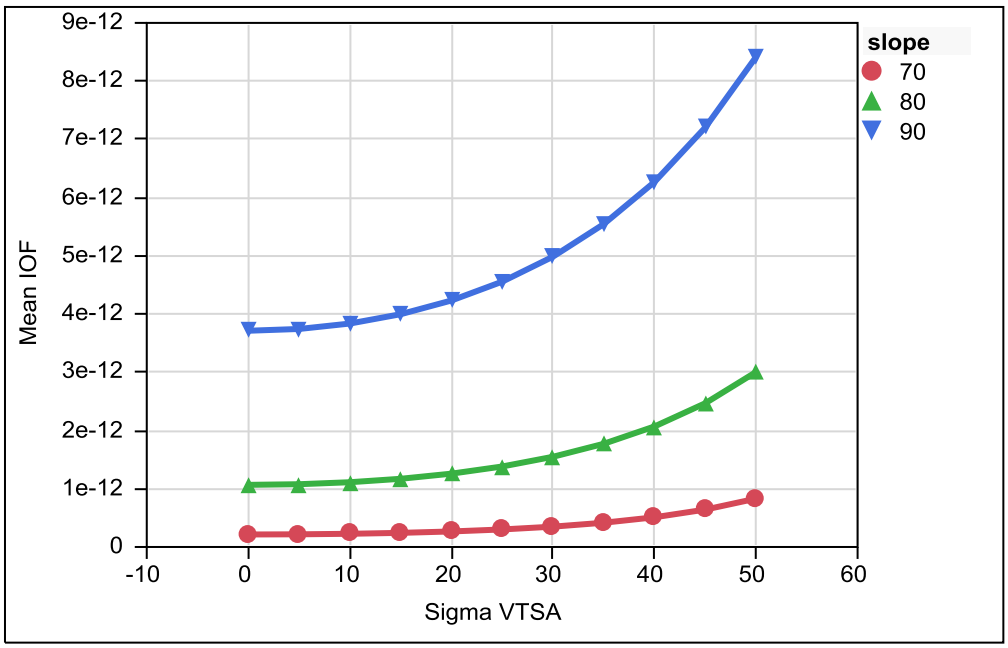

Figure 19. Mean IOFF vs. sigma VTSAT across various sub-threshold slopes.

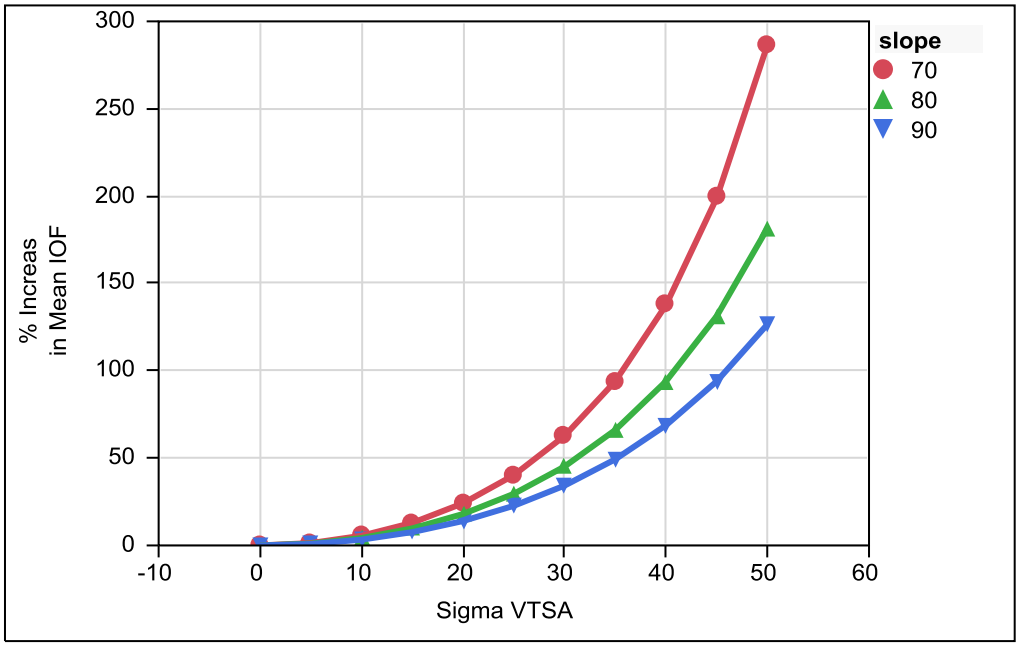

Figure 20. Percent increase in mean IOFF vs. sigma VTSAT across various sub-threshold slopes illustrating that lower sub-threshold slope results in a larger increase in the mean IOFF.

To illustrate the significance of these relationships, suppose there are 2 distributions of 1 million devices; one of the distributions has no variation at all and each device has exactly 1 pA of leakage with a 390 mV *VTSAT*, similar to the samples in Figures 18 through 20. The total leakage for this example is simply 1 uA. The second set has the same number of devices but the local variation gives rise to a 20 mV sigma perfectly centered on 390 mV's with a normal distribution. The mean *IOFF* from the second distribution of devices has a mean value that is 18% higher than case 1, or 1.18uA. If the second set of devices has a 40 mV threshold voltage sigma, then the mean *IOFF* would be 1.94 uA or 94% higher than both the TT case and the case with no variation. The local variation is increasing the mean leakage. To further illustrate the point, suppose the device widths were scaled down by 20%, and 20% more of them were placed on a chip as a natural result of scaling for increased yield. One could argue that the total device width present on each die was identical; therefore, delivering the same median standby performance would result in the same mean standby current. That conclusion would be incorrect. The scaling would most likely result in an increase in the *VT* variation and standby current would increase due to the lognormal nature of *IOFF*. The mean standby leakage is a function of the *VTSAT* variability. This analysis applies similarly to random local variation and inter-die variation. This effect will be further proven in a moment.

The central limit theorem also applies to chip-level standby currents. When the average or mean standby current is measured from a chip with many devices on it, the distribution of the average of these samples will be tighter than the contributions from

each individual device would suggest. In other words, the individual device measurements would show a much wider leakage spread than the die-level standby currents. Non-normal distributions at the device level will tend towards a normal bell curve when measured at the die level due to the averaging affect that the central limit theorem is based on [23].

A random number generator was used to produce 1000 *VTSAT* values on 1000 die. *IOFF* was computed using the *VTSAT* values at each die with a constant slope of 80 mV/decade. The sum of the currents from the 1000 devices at each die was then computed and labeled as the standby leakage. This process was repeated 4 times with the total sigma *VTSAT* equal to 20 mV's in each case. The 20 mV sigma was then altered from having no local variation to being completely dominated by the local variation. In the case where the local variation is 0%, every device on the die has exactly the same leakage. The case with 100% local variation has 20 mV's of random variation within each die and no die-to-die variation. All four cases have exactly 20 mV's of *VT* variation which induces leakage according to Equation 21. The *VT* variation was held constant at 20 mV by satisfying Equation 13 (repeated below for convenience).

$$\sigma_{total}^2 = \sigma_{local}^2 + \sigma_{die\ to\ die}^2 \qquad \textbf{Eq. 13}$$

Figure 21 shows that the mean standby leakage (blue line) is unchanged, but the corners are tighter as the percentage of local variation increases. The median standby

current from 1000 samples per die from the 390 mV *VTSAT* with an 80 mV/decade slope is ~1.04pA, and the mean due to a 20 mV sigma is about 20% higher at 1.25 nA.

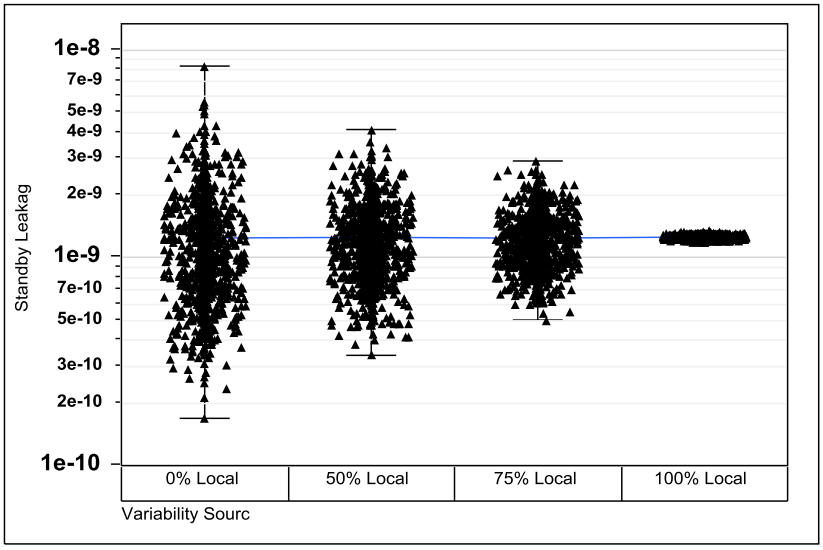

Figure 21. Standby Leakage due to 20 mV's of VTSAT variation as the percentage of local variation is varied from 0% local with 100% die-to-die to 100% local and 0% die-to-die.

Figure 22 shows the mean and median values for *IOFF* as the variation moves from 0% local to 100% local.

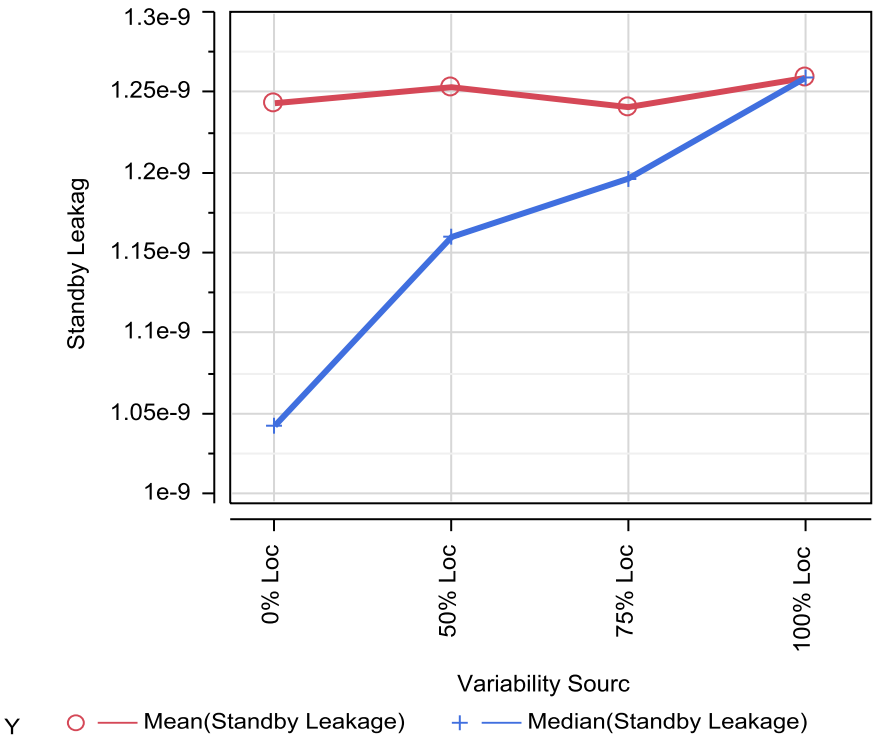

Figure 22. Mean and median sub-threshold leakage due to 20 mV's of VTSAT variation as the percentage of local variation is varied from 0% local with 100% die-to-die to 100% local and 0% die-to-die.

These results are important in illustrating how the local and non-local components of the variation effect die-leakage performance. From a parametric view taken from single-site samples across many die, it might appear that the leakage behavior induced from the 20 mV *VT* sigma is well modeled if it captures the full variation, but the worst case scenario is increasingly pessimistic as the local variation becomes a greater portion of the 20 mV sigma. The assumption here is that the standby leakage from each die has enough devices in parallel to average out the variation at the die level via the central limit theorem (more than about 30). Correlating to silicon without considering the impact of local variation would obviously reveal a gross miss in the leakage corners. Circuit

designers need to be aware of these effects so they can properly predict the leakage currents at the full-chip level. Notice that the mean value is constant in each case and only the corner cases are being impacted by the introduction of local variation. This example used a constant 20 mV sigma for *VTSAT* variation and was useful in showing how a known amount of *VTSAT* variation and an unknown decomposition of the local and non-local variability components can cause significant errors in standby current estimates.

It is also important to consider the case where the local *VT* variation is increasing in the presence of constant die-to-die *VT* variation. The data in Figures 23-27 was generated using a random number generator with the die-to-die or non-local variation set to 14.14 mV while varying the random variation from 14.14 mV, to 30 mV's. Each die has 1000 samples and there are a total of 1000 die in the generated table. Figure 23 shows the *VTSAT* distributions generated in the table along with the sigma for each distribution. The total variation is still given by Equation 13, and the case with both the local and non-local variation set to 14.14 mV's results in a sigma of 20 mV's like the previous example. This case is labeled as the 14/14 case, with the first value being the die-to-die sigma and the second the random intra-die sigma. The other scenarios are labeled similarly. Equation 13 predicts a total variation of 33 mV's when the die-to-die variation is 14 mV with the random local variation at 30 mV. The chart of sigma in Figure 23 agrees.

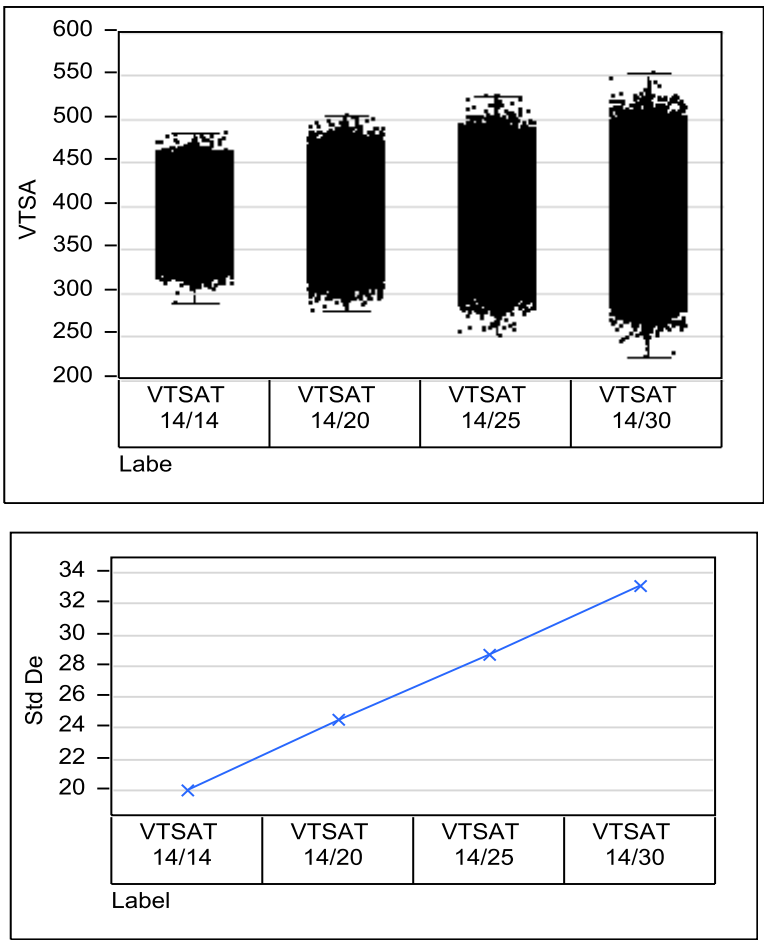

Figure 23. VTSAT variation for fixed die-to-die variation with local variation increasing from 14 to 30 mV's, illustrating how the variance of the two components are summed.

Figure 24 shows that as the local *VT* variation increases, *IOFF* variation also increases. This is again the expected response from the lognormal relationship. Figure 25 shows how the mean and median standby leakage increases. It is important to recognize that the mean and median in Figure 25 represent the die-level sum of the lognormal *IOFF* distributions at each die. Both the mean and median increase because the total variation increases each time the local variation increases. The median of the inter-die variation is also the mean of the intra-die variation.

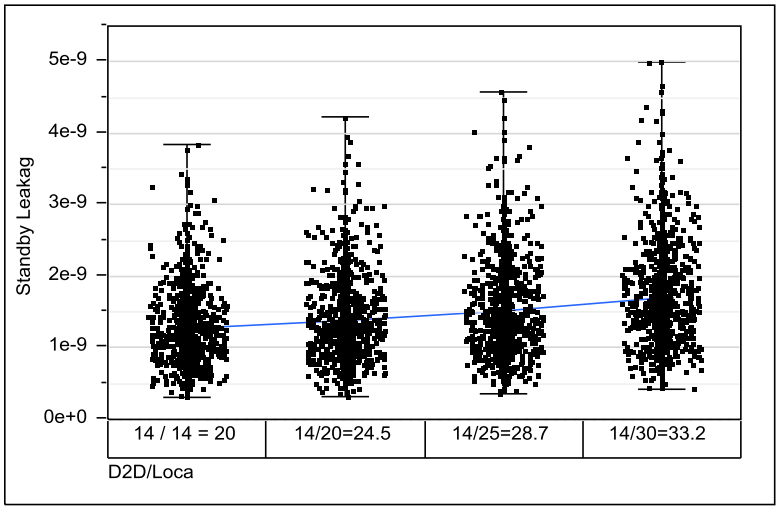

Figure 24. Standby leakage as local VTSAT variation increases from 14 to 30 mV with a constant 14 mV die-to-die variation illustrating an increase in the mean IOFF.

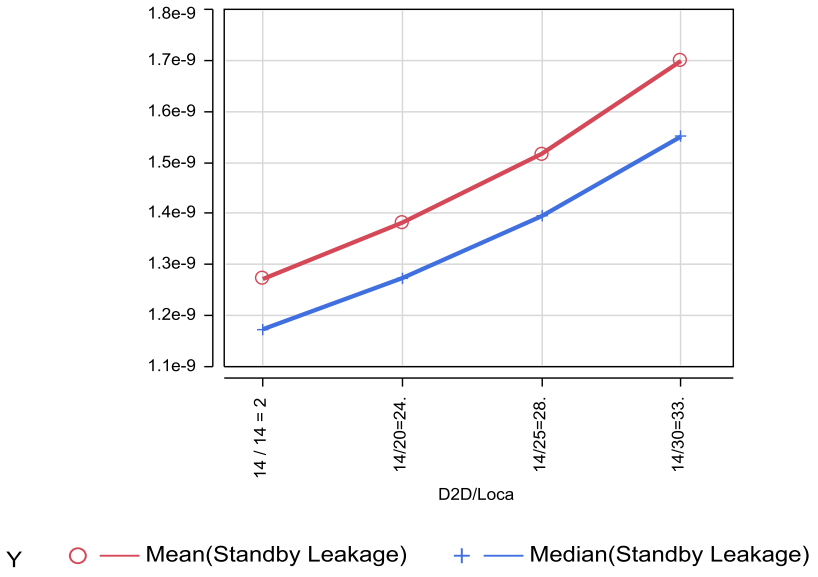

Figure 25. Mean and median IOFF increasing due to increased local variation in the presence of constant die-to-die variation.

Figure 26 shows the *VT* variation across a random sample of 10 of the 1000 die. Each site indeed has a sigma of about 14 mV's. *IOFF* for each of these 10 sampled sites is also shown in Figure 27.

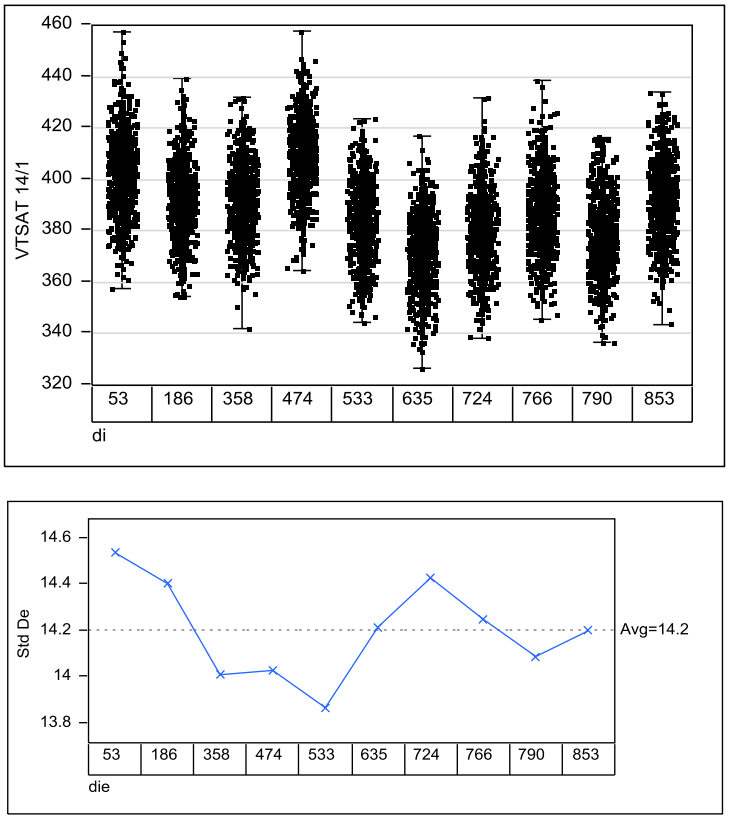

Figure 26. A 10 site sample of VTSAT with 14 mV of local and die-to-die variation.

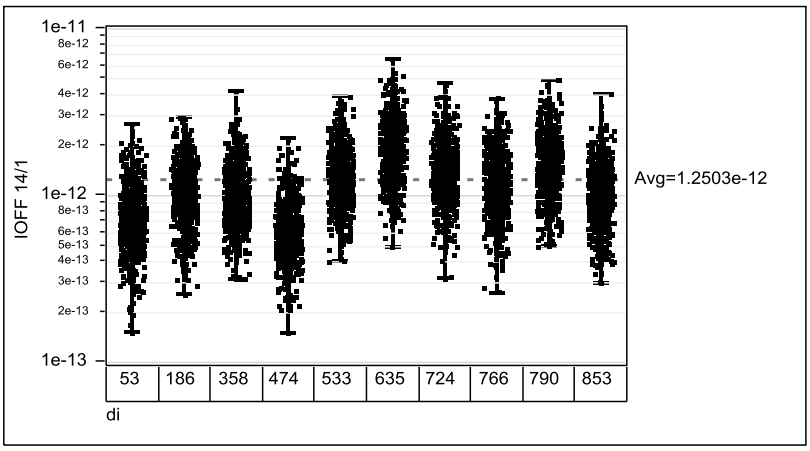

Figure 27. A 10-site sample of IOFF induced from 14 mV of local and die-to-die *VT* variation.

Standby current estimates require careful attention to the local and die-to-die variation components. A failure to predict these properly can result in gross errors in standby leakage estimates.

4.3 Gate Delay and Clock Tree Behavior

Local variation has played a role in CMOS digital logic since the very first CMOS circuit was fabricated. Dopant fluctuations and interface states have always been present but their impact has not always been at the forefront of digital logic design. Shrinking overdrive voltage and tighter timing margins are highlighting the second tier issues such as mismatch or random intra-die variation. The impact of mismatch on digital logic, however, has seen a surge in research over the last 10 years as technologies have pushed the limits for overdrive voltage and device geometries.

The impact of local variation on digital logic gates can be studied by looking at a string of logic gates such as one would find in a clock tree or ring oscillator. Each stage in the string will have a random tendency towards faster and slower delays. A portion of the stages will be faster and a portion will be slower than the mean gate delay. A longer

string of gates will tend to have a more equal number of slow and fast gates such that the delay at the end of the chain will approach $n_{cp} \cdot T_{PD}$ as the length increases, where n_{cp} is the number of stages in the critical path and T_{PD} is the mean propagation delay for a single stage. The ring oscillator can be a valuable characterization tool for studying the delay variation per stage for a given logic gate such as an inverter, nand, or nor gate chain. The difference in frequency between identically placed ring oscillator chains can be measured across a large sample to determine the frequency or delay variation. This by no means replaces device-level mismatch characterization, but can offer a confirmation that the DC device-level data is translating as expected to AC performance. Varying the number of stages can be used to fit the trend where a shorter path will have a greater tendency to differ from its matched pair than a longer path. The longer paths will have a longer delay but the ratio of the standard deviation to the mean delay will decrease inversely proportional to the number of stages in the path; Equation 29 describes the relationship [24].

$$\frac{\sigma_{TPD,CP}}{\mu_{TPD,CP}} = \frac{\sqrt{n_{cp}} \cdot \sigma_{PD_{rand},stage}}{n_{cp} \cdot \mu_{T_{PD,stage}}} = \frac{1}{\sqrt{n_{cp}}} \cdot \frac{\sigma_{PD_{rand},stage}}{\mu_{T_{PD,stage}}} \qquad \textbf{Eq. 29}$$

N_{cp} is the number of stages in the critical path. Notice how sigma increases by a factor of $\sqrt{n_{cp}}$ for each stage added. This is due to the fact that the stages are assumed to be completely independent and random such that the sum of the variance of each stage will give us the total variance of the critical path (*CP*). Equation 30 illustrates this point, and again we sum the variances of the independent sources of variation, not the sigmas.

$$\sigma^2_{CP_{rand}} = \sigma^2_{PD_{rand,stage1}} + \sigma^2_{PD_{rand,stage2}} + \sigma^2_{PD_{rand,stage3}} \cdots$$

$$\textit{for identical stages} \quad \sigma^2_{CP} = n_{cp} \cdot \sigma^2_{\sigma_{PD_{rand}},stage} \qquad \textbf{Eq. 30}$$

The mean path delay simply increases linearly with each added stage. Contrast Equation 29 with the Equation 31 for the systematic case (die-to-die variation) in which all devices move together at the same time [24].

$$\frac{\sigma_{TPD,CP}}{\mu_{TPD,CP}} = \frac{n_{cp} \cdot \sigma_{PD_{sys},stage}}{n_{cp} \cdot \mu_{T_{PD,stage}}} = \frac{\sigma_{PD_{sys},stage}}{\mu_{T_{PD,stage}}} \qquad \textbf{Eq. 31}$$

Notice now how the sigma and mean critical path delay grow proportional to n_{cp}. Note also that the ratio of the sigma to the mean is not reduced by a factor of $1/\sqrt{n_{cp}}$ as it is for the case with random local variation, but remains constant as stages are added. This illustrates the point that die-to-die variation is much more detrimental to the propagation delay variation than local variation since it does not get averaged out. However, this does not mean that the effects of local variation are negligible [24].

Consider a string of 100 gates each with a normalized delay of 1. The same string is repeated 1000 times using a random number generator. In one case the variation is the same for each stage in the string that simulates die-to-die variation, and in the other case each stage has an independent random value for the delay. In each case, the sigma is set

to 0.1. Figure 28 shows the path delay versus the stage number for 1000 paths for both the systematic (die-to-die) case and the random local case. Despite the fact that the random and systematic variation has the same sigma, the path delay is much more sensitive to the systematic die-to-die variation than the local variation, particularly after a fair number of stages, as is illustrated in Figure 28.

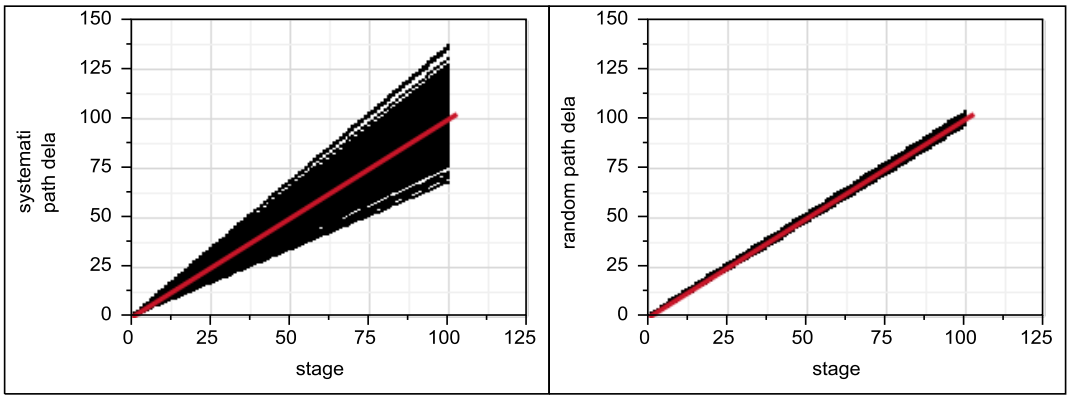

Figure 28. Normalized path delay due to systematic die-to-die and random intra-die from a 10% sigma for each component.

Taking a closer look at the ratio of the sigma to the mean delay per stage (DPS) we can see that the variation is indeed more similar when the number of stages is near 1, but even after the second stage is added the benefits of averaging in the random case begin to show up. This effect is shown in Figure 29.

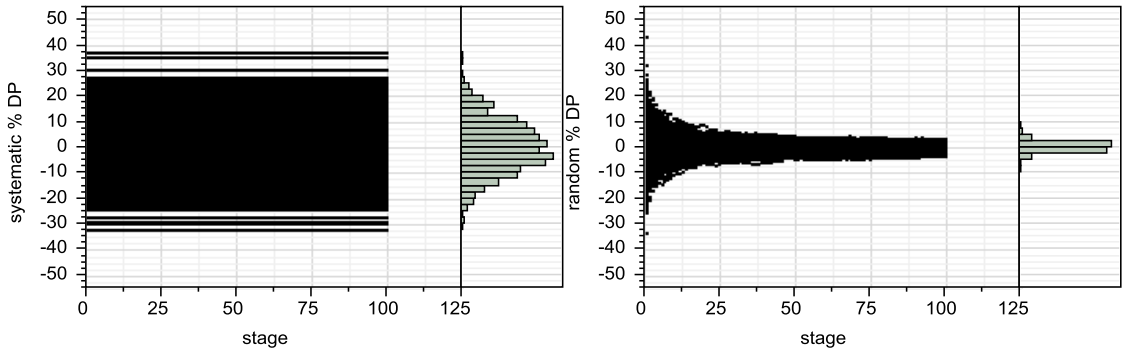

Figure 29. Average delay per stage due to systematic and random variation of 10% as the path length increases from 1 to 100 consecutive stages illustrating how random local variation averages out as the number of stages increases while the systematic die-to-die variation does not.

The expected sigma after 100 stages for the systematic case using Equation 31 is 10. The expected sigma after 100 stages for the random variation using Equation 29 is 1. Figure 30 shows that the sigma does indeed come close to 10 and 1 for the systematic and random variation cases respectively after 100 stages. Figure 30 also shows the total or global variation as a result of both the local and the systematic variation each with a mean of 1 and a sigma of 0.1.

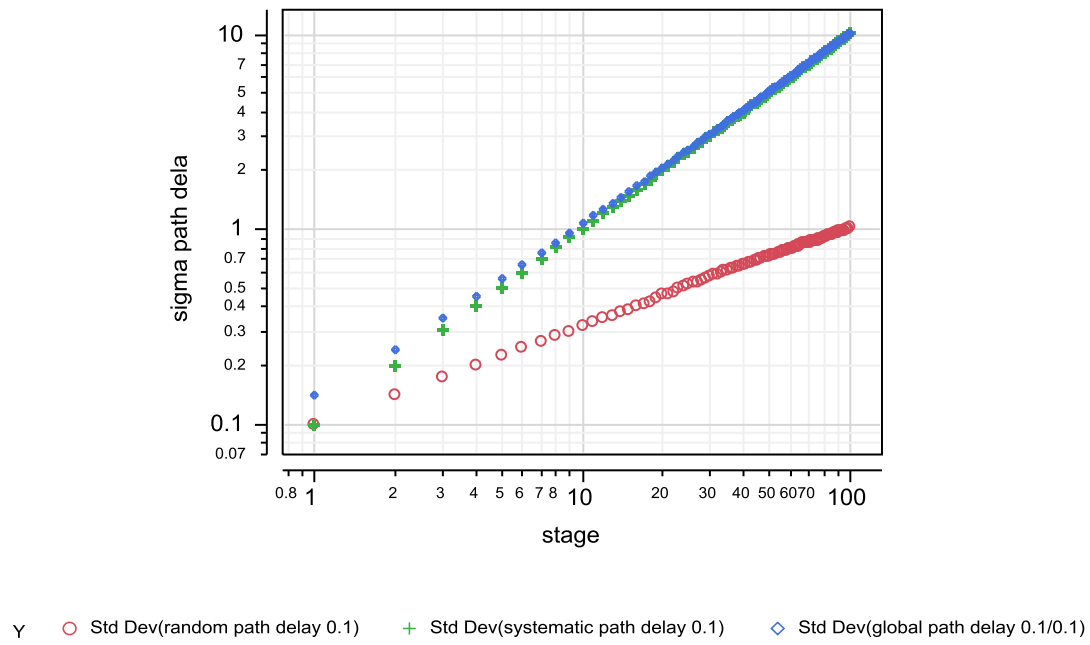

Figure 30. The path delay sigma with a 10% sigma for local and systematic variation along with the combined global variation on a log-log scale showing how the random local variation plays a larger role when the number of consecutive stages is low.

Plotting sigma for the DPS (delay/#stages) better illustrates how the combination of the local and systematic variation is greater when the number of stages is smaller as shown in Figure 31.

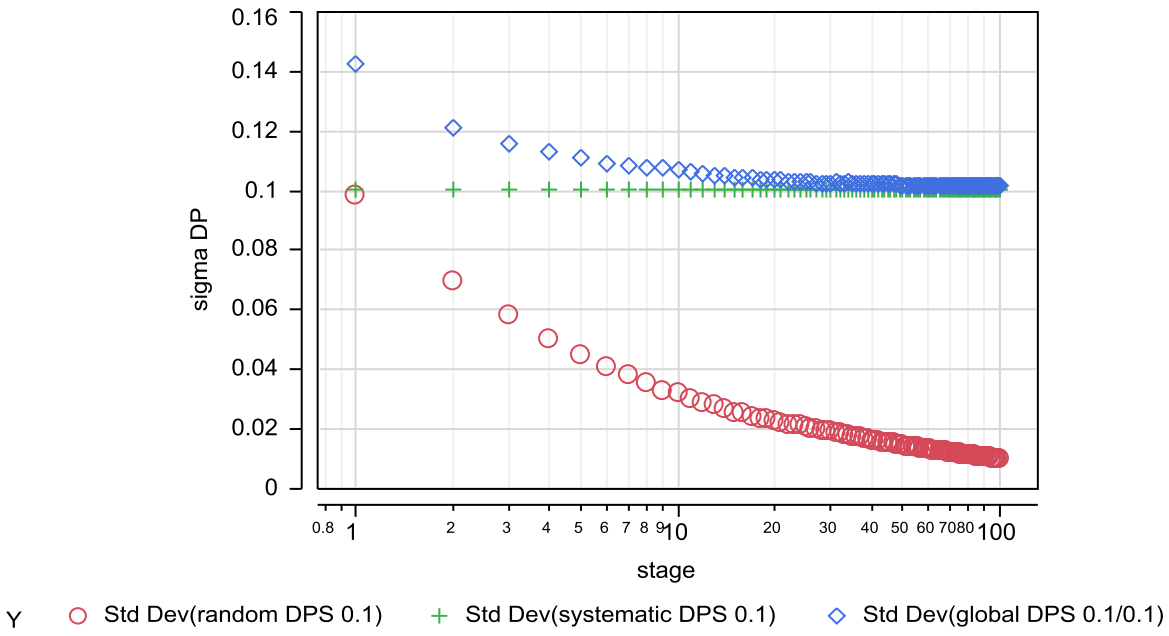

Figure 31. Sigma in delay per stage with a 10% sigma for local and systematic variation along with the combined global variation on a log-log scale showing how the random local variation plays a larger role when the number of consecutive stages is low.

The combination of variance for both the local and the systematic variance is estimated using Equation 32, and Equation 33 shows the sigma. When n_{cp} is equal to one, we expect the global sigma to be 0.1414, which Figure 31 confirms.

$$\boldsymbol{\sigma^2_{CPD_{Global}} = n_{cp} \cdot \sigma^2_{PD_{rand},stage} + n_{cp}{}^2 \cdot \sigma^2_{PD_{sys},stage}} \qquad \textbf{Eq. 32}$$

$$\boldsymbol{\sigma_{CPD_{Global}} = \sqrt{n_{cp} \cdot \sigma^2_{PD_{rand},stage} + n_{cp}{}^2 \cdot \sigma^2_{PD_{sys},stage}}} \qquad \textbf{Eq. 33}$$

If we look more closely at a clock tree and use the fundamental path relationships shown above, we can come up with some basic tradeoffs and design considerations for

clock tree design in the presence of local variation. If we consider only the local variation for a moment, we can determine that a longer path will have less variation than a short path for a given stage delay. However, the added insertion delay may not be worth the benefit of the reduction in variation. The best approach for reducing the local variation for a fixed path length is to increase the size of the devices used in each stage of the path. This will reduce the impact of local variation but will increase power consumption and layout area. Clock tree design, however, is very complex and architecture variants will have unique benefits and the pros and cons will have to be analyzed in detail for a given technology and application. The basic clock tree architecture is shown in Figure 32. It is made up of root and branch stages and can have multiple trunks and branches. The clocked load logic at the end of the branches can be referred to as the leaves of the tree [25]. Local variation in the trunk will affect all branches equally and random variations in the branch devices can cause offsets between the branches. For this reason, it may be a good compromise to use larger devices in the branches to reduce the impact of local variation for branch-to-branch matching [26], but a larger device in the root or trunk will not reduce the difference between the branches since the trunk is common to all branches.

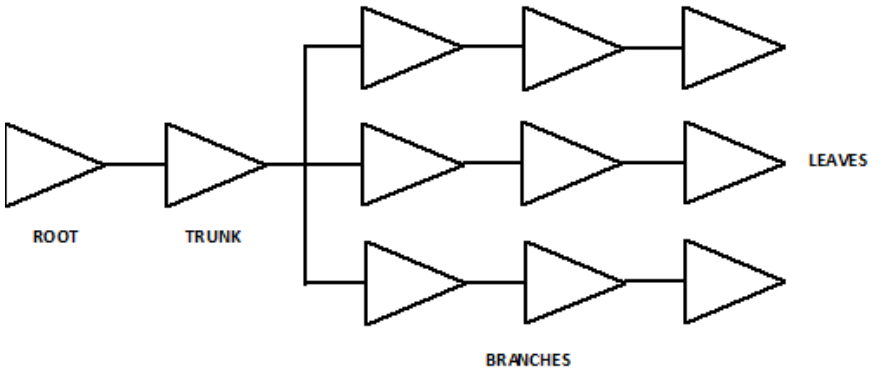

Figure 32. Basic clock tree architecture showing the root, trunk, and branches with loads designated as the leaves.

The difference between branches can be estimated using Equations 34 and 35, where the critical path length starts at the beginning of the branch. If the delay is 1 and the random local sigma is 0.1 for each stage, then a single branch will have a sigma of 0.173. If the branches are identical, then each will have the same local variation and the difference between any two of the branches will be given by Equation 35. The difference between branch outputs is greater than the sigma of any individual branch by a factor of $\sqrt{2}$. Equation 36 shows a more general case in terms of the stage variation from Equation 29. The sigma for the difference between any of the branches with three stages will be 0.245. The systematic die-to-die variation will not produce any delta between branches. However, it is possible that layout dependent offsets could be introduced and cause constant systematic offsets between the branches.

$$\boldsymbol{\sigma^2_{\Delta branch} = \sigma^2_{branch1} + \sigma^2_{branch2}} \qquad \textbf{Eq. 34}$$

$$\boldsymbol{\sigma_{\Delta branch} = \sqrt{2} \cdot \sigma_{branch}} \qquad \textbf{Eq. 35}$$

$$\sigma_{\Delta branch} = \sqrt{2} \cdot \sqrt{n_{cp}} \cdot \sigma_{\sigma_{PD_{rand}}, stage} \quad \textbf{Eq. 36}$$

Clock tree designers should characterize the local variation of each repeater in the tree to predict the difference between branches using Equation 36. The characterization can be accomplished by using a statistical model that contains both the local and non-local variability components.

In an article recently published by Mallik Devulapalli and Yuichi Kawahara from Synopsis Inc, a ‘mesh’ architecture was used to greatly reduce the impact of local variation on clock signal distribution. Figure 33 shows the differences in architecture between the conventional clock tree and a clock mesh [27]. The difference between the branches can be reduced by sharing the nodes at the end of the branches in a mesh of interconnect. The need for such strategies will be specific to the impact of local variation for a given technology, but it is evident that there are circuit design topologies that can help reduce the impact of local variation on circuit performance.

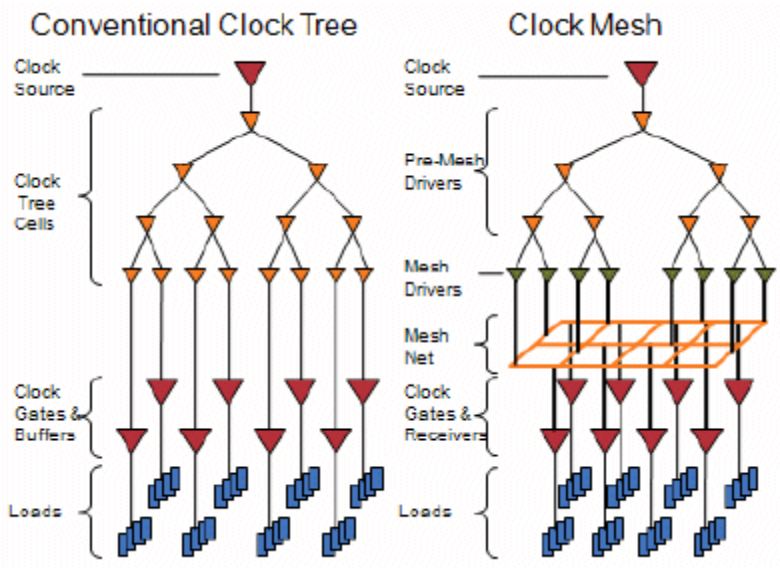

Figure 33. The conventional clock tree is shown on the left and is susceptible to local variation between branches contrasted against the clock mesh on the right, which aligns the local variation at the mesh [27].

CHAPTER FIVE – SUMMARY

5.1 Summary

Random local variation in CMOS devices adds significant complexity to the characterization, modeling, and circuit design processes. The effects of random local variation are most prominent when overdrive voltage is low and when the oxide thickness is not scaled. Random data sets were generated across various combinations of local and non-local variation in order to illustrate characterization, modeling, and design challenges. These data sets were used to predict the statistical response for standby currents and digital gate delays in logic paths and clock trees. The behavior of these circuits depends highly on the significance of local variation for a given technology.

Device development teams have to consider the impact of local variation at all phases of process development. Circuit designers need to understand proper simulation techniques and how random variation affects circuit response. Accounting for random variation is particularly important for estimating standby leakage currents for large blocks in highly scaled CMOS transistors in order to prevent over design. Local variation can also result in significant branch-to-branch variation within clock trees and must be accounted for during the design cycle. Failure to properly account for local variation can result in over design and inefficient layouts as well as under design and possible circuit failures where timing margins are tight.

WORKS CITED

[1] A. Voss, "nanoCMOS Device, Circuit and System Simulations," 17 November 2009. [Online]. Available: http://cnx.org/content/m32874/1.1/. [Accessed 31 March 2012].

[2] M. Pelgrom, "Matching Properties of MOS Transistors," *IEEE Journal of Solid State Circuits,* pp. 1433-1439, 1989.

[3] T. Mizuno, "Influence of statistical spatial-nonuniformity of dopant atoms on threshold voltage in a system of many MOSFETs," *Japanese Journal of Applied Physica,* vol. 35, no. 2B, pp. 842-848, 1996.

[4] Jeroen A. Croon, Willy Sansen, Herman Maes, Matching Properties of Deep Sub-Micron MOS Transistors, Dordrecht: Springer, 2005.

[5] IEEE, *IEEE tranactionson Electron Devices, Special issue on Characterization of Nano CMOS variability by Simulation and Measurements,* vol. 58, no. 8, pp. 2190-2818, 2011.

[6] Yuan Taur, Tak H. Ning, Fundamentals of Modern VLSI Devices, Cambridge: The Press Syndicate of the University of Cambridge, 1998.

[7] R. J. Baker, CMOS Circuit Design, Layout, and Simulation, Hoboken: John Wiley & Sons, 2010.

[8] Ceclilia Maggioni Mezzomo, Aurelie Bajolet, Augustin Cathignol, Regis Di Frenza, Gerard Ghibaudo, "Characterization and Modeling of Transistor Variability in

Advanced CMOS Technologies," *Transactions on Electron Devices, VOL. 58, NO. 8,* pp. 2235-2248, 2011.

[9] R. DiFrenza, J.C. Vildeauli, P. Llinares, G. Ghibaudo, "Impact of grain number fluctuations in the MOS transistor gate on matching performance," *IEEE ICMTS*, pp. 244-249, 2003.

[10] H. Tuinhout, A. Montree, J. Schmitz, P Stolk, "Effects of gate depletion and boron penetration on matching of deep submicron CMOS transistors," in *International Electron Device Meeting*, 1997.

[11] A. Asenov, "Statistical device variability and its impact on design," Device Modeling Group, Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow, 2008.

[12] K. J. Kuhn, M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma and S. Mudananai, "Process Technology Variation," *IEEE Transactions on Electron Devices, VOL. 58, NO. 8*, pp. 2197-2208, 2011.

[13] Sharod Saxena, Hossein Karbasi, Angelo Rossoni, Stefano Tonello, Patrick McNamara, Silvia Lucherini, Sean Minehane, Christopher Dolainsky, Michelle Quarantelli, "Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies," *IEEE Transactions on Electron Devices, Vol. 55, No 1,* pp. 131-144, 2008.

[14] Patrick G. Drennan, Colin C. McAndrew, "Understanding MOSFET Mismatch for Analog Design," *IEEE Journal of Solid State Circuits* , vol. 38, no. 3, pp. 450-456, 2003.

[15] P. Tan, A. Kordesch and O. Sidek, "CMOS transistor mismatch model with temperature effect for HSPICE and SPECTRE," in *Solid-State and Integrated Circuits Technology*, 2004.

[16] Pietro Andricciola, Hans Tuinhout, "The Temperature Dependence of Mismatch in Deep-Submicrometer Bulk MOSFET's," *IEEE Electron Device Letters,* vol. 30, no. 6, pp. 690-692, 2009.

[17] S.E. Rouch III, "The Statistics of NBTI-Induced Vt and β mismatch shifts in pMOSFET's," *IEEE Transactions on Device Materials and Reliability,* vol. 2, no. 4, pp. 89-93, 2002.

[18] Paolo Magnone, Felice Crupi, Nicole Wils, Ruchil Jain, Hans Tuinhout, Pietro Andricciola, Gino Giusi, Claudio Fiegna, "Impact of Hot Carriers on nMOSFET Variability in 45- and 65-nm CMOS Technologies," *IEEE Transactions on Electron Devices,* vol. 58, no. 8, pp. 2347-2353, 2011.

[19] S. Pae, J. Maiz, C. Prasad, B. Woolery, "Effect of BTI degradation of Ttransistor Variability in Advanced Semiconductor Technologies," *IEEE Transacations on Device Materials and Reliability,* vol. 8, no. 3, pp. 519-525, 2008.

[20] Diana Lopez, S. Haendler, C. Leyris, Gregory Bidal, Gerard Ghibaudo, "Low-Frequency Noise Investigation and Noise Variability Analysis in High-k/Metal Gate32-nm CMOS Transistors," *IEEE Transactions on electron devices,* vol. 58, no. 8, pp. 2310-2316, 2011.

[21] M. Miranda, B. Dierickx, P. Zuber, P. Dobrovoln, F. Kutscherauer, P. Roussel and P. Poliakov, "Variability aware modeling of SoCs: From device variations to manufactured system yield," in *ISQED*, San Jose, 2009.

[22] E. W. Weisstein, ""Log Normal Distribution." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/LogNormalDistribution.html," MathWorld--A Wolfram Web Resource, [Online]. Available: http://mathworld.wolfram.com/LogNormalDistribution.html . [Accessed 20 March 2012].

[23] J. L. Devore, Probability and Statistics for Engineers and the Sciences, Pacific Grove: Duxbury, 2000.

[24] Keith A. Bowman, Xinghai Tang, John C. Eble, James D. Meindl, "Impact of Extrinsic andIntrinsic Parameter Fluctuations on CMOS Circuit Performance," *IEEE Journal of solid-state circuits*, vol. 335, no. 8, pp. 1186-1193, 2000.

[25] E. G. Friedman, "Clock Distribution Networks in Synchronous Digital Integrated Circuits," *Proceedings of the IEEE*, vol. 89, no. 5, pp. 665-692, 2001.

[26] Tarun Chawla, Amara Amara, Andrei Vladimirescu, "Yield, Power and Performance Optimization for Low Power Clock Network under Parametric Variations in Nanometer Scale Design," in *IEEE International Midwest Symposium on Circuits and Systems*, San Juan, 2006.

[27] Y. K. Mallik Devulapalli, Yuichi Kawahara, "Clock Mesh Variation Robustness: Benefits and Analysis," Synopsis Inc, [Online]. Available: http://www.design-reuse.com/articles/21019/clock-mesh-benefits-analysis.html. [Accessed 29 3 2012].