2008

Global Interconnects in the Presence of Uncertainty

Ibis D. Benito
University of Massachusetts Amherst

GLOBAL INTERCONNECTS IN THE PRESENCE OF UNCERTAINTY

A Thesis Presented

by

IBIS BENITO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL AND COMPUTER ENGINEERING

February 2008

Department of Electrical and Computer Engineering

GLOBAL INTERCONNECTS IN THE PRESENCE OF UNCERTAINTY

A Thesis Presented

by

IBIS BENITO

Approved as to style and content by:

_____

Wayne Burleson, Chair

_____

Maciej Ciesielski, Member

_____

Sandip Kundu, Member

_____

C.V. Hollot, Department Head

Department of Electrical and Computer Engineering

To my husband, sisters and parents.

ACKNOWLEDGMENTS

Thank you to my advisor, Professor Wayne Burleson for his continuous guidance and support. Thank you for the opportunity of joining your group to begin my journey as a graduate student at the University of Massachusetts.

I would like to thank the Northeast Alliance for the Graduate Education and the Professoriate (NEAGEP) for giving me the opportunity of visiting UMass for the first time in 2002 and for their support during the first year of my Master's studies. Thank you for giving me the opportunity of being a spokesperson for underrepresented groups in the sciences and engineering fields and for incorporating me into your recruiting programs.

I would like to thank the Semiconductor Research Corporation (SRC) and Intel for their support during my Master's thesis and for supporting this work.

Finally, I would like to give a special thank you to Vishak Venkatraman, Jinwook Jang, Sheng Xu and all of the members of the VLSI Circuits and Systems Group for their support and collaboration.

ABSTRACT

GLOBAL INTERCONNECTS IN THE PRESENCE OF UNCERTAINTY

FEBRUARY 2008

IBIS BENITO

B.S., UNIVERSITY OF PUERTO RICO MAYAGUEZ

M.S.E.C.E., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Wayne Burleson

Global interconnect reliability is becoming a bigger issue as we scale down further into the submicron regime. As transistor dimensions get smaller, variations in the manufacturing process, and temperature variations may cause undesired behavior, and as a result, compromise performance. This work makes an effort to characterize the effects of such variations, to provide designers with a guideline for making designs tolerant to these variations while benefiting from tighter design margins.

Since interconnects contribute to most of the delay and power on a chip, interconnect performance becomes a primary issue in design. One of the main concerns when considering physical transistor dimension variations is the effect on delay. Due to smaller transistor dimensions, the photolithographic process may produce transistors with significant variations from the ideal physical dimensions. Such variations cause delay uncertainty which can lead to over or underestimation in the design phase. This work examines interconnects to establish a guideline of the effect that process variations have on delay. A repeated

interconnect is analyzed and the effects of physical device variations on delay are observed. Given the delay distribution in the presence of L_{eff} variation, a supply voltage assignment technique is proposed to correct the observed deviation from the nominal delay on a long, repeated interconnect. This technique results in a significant reduction of the delay distribution, with a negligible power overhead.

After looking at static variation effects on interconnect performance, this thesis addresses thermal variations on global signals, which cause delay degradation and may lead to timing failures. Given the presence of a large thermal gradient along a clock signal in a data path clocked by two leaves of an H-tree, several thermal scenarios which can compromise timing are discussed. A buffer-based skew compensation technique is proposed to correct the effect of thermal and manufacturing variations on this system.

Having characterized repeated interconnect performance under process variations, the bandwidth of the line can be more effectively utilized by using a technique called phase coding. Phase coded interconnects are introduced in the context of using them once an interconnect has been adequately modeled in the presence of variations.

With guidelines quantifying the effects of process variations on interconnect techniques and careful characterization, designers can factor these considerations into their design process, reducing the variation from the nominal expected behavior and allowing for smaller design margins. This will lead to more reliable products as we advance into future technologies and transistor dimensions get smaller.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As the CMOS era approaches the use of technologies that challenge manufacturing and design standards, designers struggle to find solutions that will not compromise timing and cost budgets. Moore's Law predicted with remarkable accuracy that the number of transistors on a chip doubles roughly every 2 years. Smaller technologies, however, bring up the issues of silicon and optical limitations that make this guideline harder to achieve. Furthermore, variations in nominal transistor dimensions and environmental variations that occur at run-time may compromise design timing requirements and hinder performance. The need for robustness makes it essential to have designs that are scalable into future CMOS technologies. In an effort to predict CMOS behavior, the International Technology Roadmap for Semiconductors [1] provides edition reports every two years (with an update every other year) to set up targets for performance, power, and other relevant measures for future CMOS technologies. Even if design solutions are not known to meet a given requirement, ITRS gives insight into what designers should take into consideration in current technologies and those to come.

1.1 Scaling Effects on Interconnects

Device scaling has proven to be a difficult challenge for designers looking to meet performance and power budgets for current and future technologies. ITRS sets design

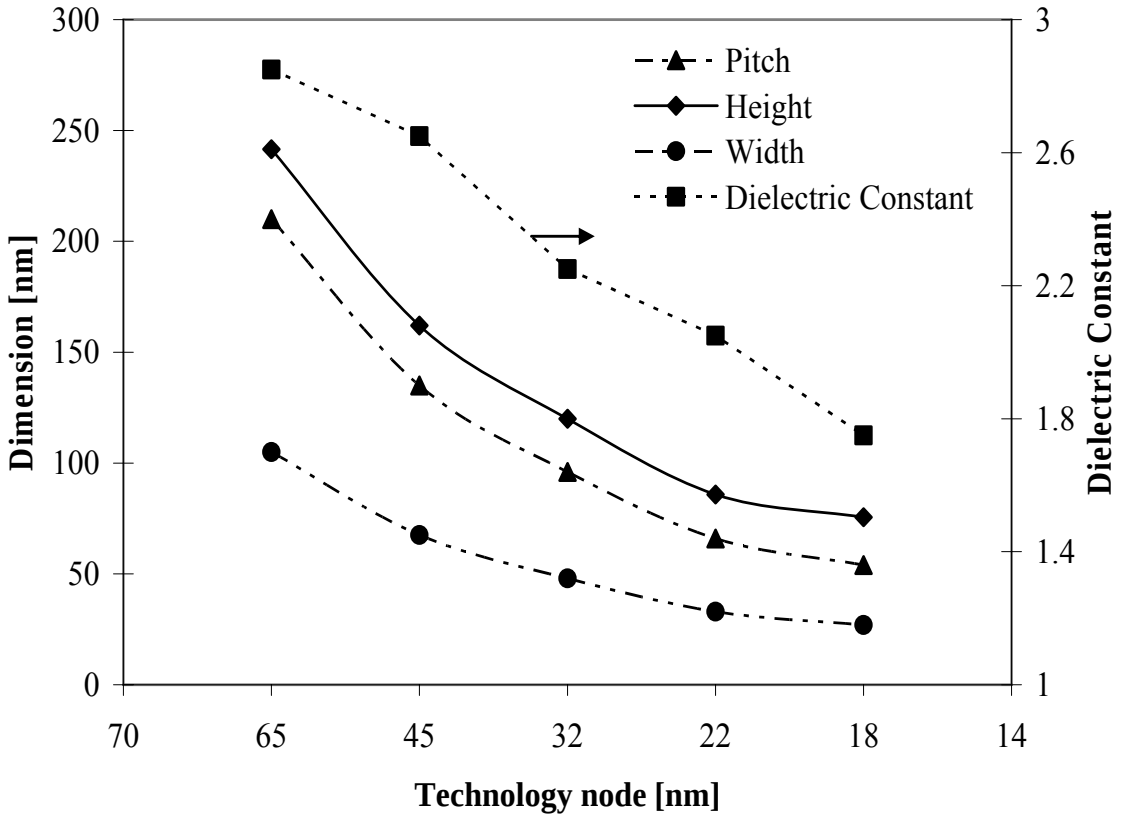


Figure 1. ITRS projections for interconnect dimensions and dielectric constant.

goals and gives insight into manufacturing and design techniques to meet them. ITRS currently predicts CMOS technology will reach up to 14nm for year 2020, however, as we delve deeper into the roadmap, challenges such as device characteristic uncertainty and wire dimension variations become very important issues that must be taken into account when designing for current and future technologies. With the advent of high-k dielectrics and optical solutions for the small dimensions that must be accurately produced in the chip manufacturing process, Moore's prediction for CMOS has been accurate so far. However, as physical limits are reached, research into new non-CMOS technology emerges, with nanotechnology quickly becoming a key area of interest for researchers who wish to explore beyond the CMOS regime.

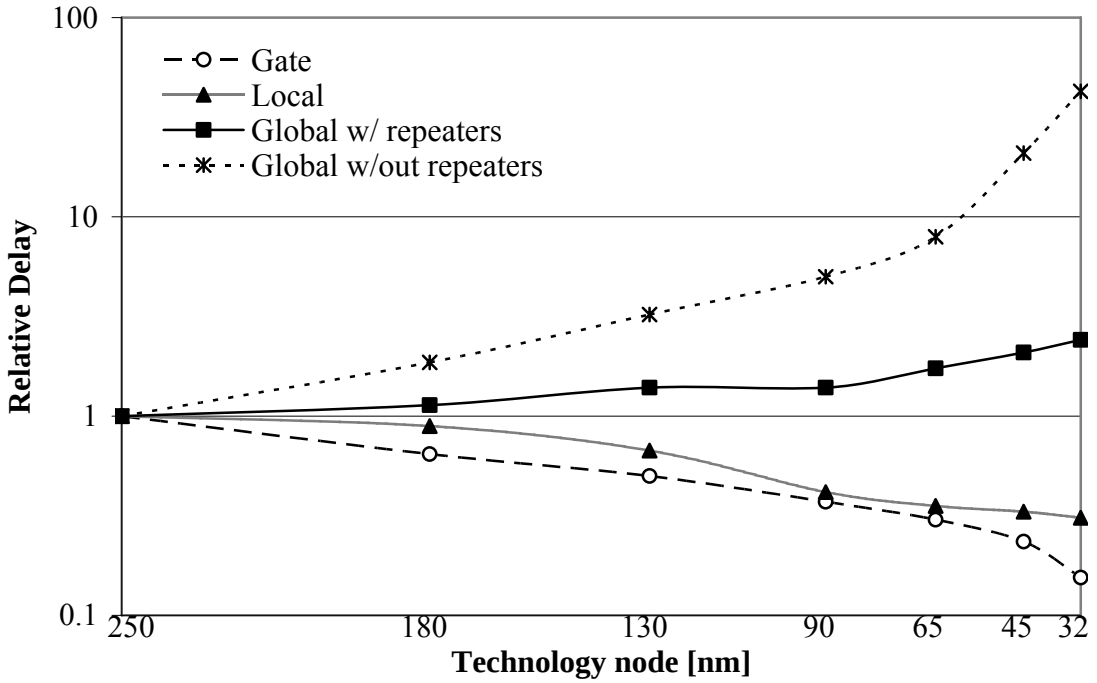


Figure 2. On-chip interconnect relative delay across technology nodes from [1].

With smaller technologies, transistor dimensions shrink and therefore the transistor operation states become harder to control. In addition, leakage becomes a bigger problem since the channel length of the transistor is so small, that a significant amount of current leaks from drain to source, more so in smaller technology nodes. Leakage becomes a dominant component in the total power of the chip and is projected to dominate total power in the future.

Figure 1 shows the trend for various interconnect dimensions and the dielectric constant values, as documented by ITRS. These predictions do not take into account the possible adoption of high-k dielectric materials in future technologies, hence the downward scaling trend across technologies. High-k dielectrics are expected to minimize the effect of leakage by increasing gate capacitance. The implementation of such materials has been a

recent topic of research in the industry [2]. Interconnect dimensions are expected to steadily decrease with technology scaling, as seen in the figure. As dimensions approach limits of optical masking technology, the effect of inaccuracies in interconnect dimensions is greater and can be seen in the form of unexpected behavior in performance and power. It is widely known that with scaling, interconnects become the limiting factor of on-chip performance. Interconnect scaling increases delay since resistance increases as interconnect dimensions get smaller (interconnect resistance is inversely proportional to interconnect cross-sectional area). Relative delay for a gate with fan out of four (FO4), a local interconnect, a global interconnect without repeaters, and a repeated global interconnect is shown in Figure 2 for current and future technology nodes. The increase of delay with technology scaling is observed on a global interconnect without repeaters. This delay increase is diminished by using repeater insertion. Repeaters are inverters inserted along the interconnect to divide it into equal segments in order to bring the dependence of delay on the length of the interconnect from quadratic to linear. Repeater sizing is chosen such that the signal quality and delay is acceptable at the end of the long interconnect. Even though repeaters provide a performance benefit, they consume a lot of power, which is increasingly undesirable in current and future technologies. Repeater insertion optimization algorithms [3,4] have been widely researched and developed to minimize the penalties paid by inserting repeaters, while keeping their benefits. Repeaters are also prone to variations in transistor dimensions during the manufacturing stage, as well as environmental changes that may affect the expected outcome of the logic. Many researchers have looked at repeater insertion in the presence of process variation as well [5,6]. Since interconnects account for a large part of the delay on a

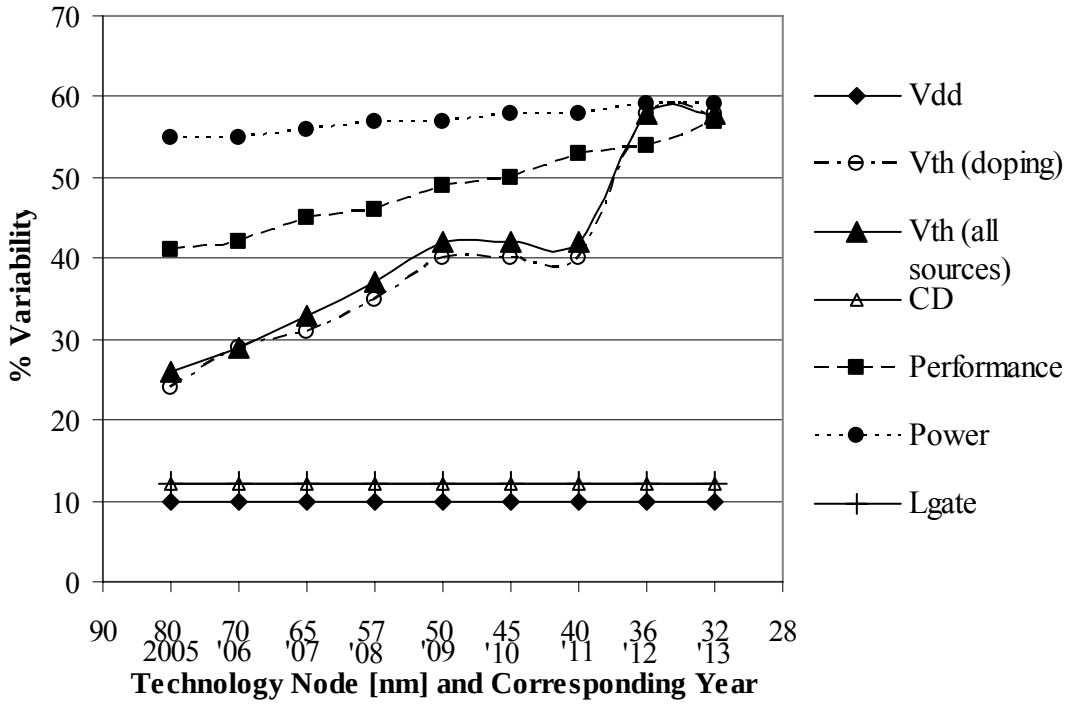


Figure 3. Variability percentage for critical interconnect dimension and performance across technology nodes.

chip in today's microprocessor designs, it is crucial for designers to explore options that do not compromise timing budgets, while still reducing the impact that interconnects have on the overall chip delay.

Two sources of uncertainty that directly impact chip performance are physical and environmental factors. These uncertainties in drawn transistor dimensions, interconnect dimensions and environmental changes are identified as process variations. Process variations are classified into intradie and interdie variations. Intradie variations are those that occur spatially within a die, while interdie variations constitute parameter deviations from the nominal value across nominally equal die [7]. In general, variation trends get worse as we scale down to future technology nodes. Figure 3 shows the percentage of variability of several process parameters for current technologies and that expected for future technologies,

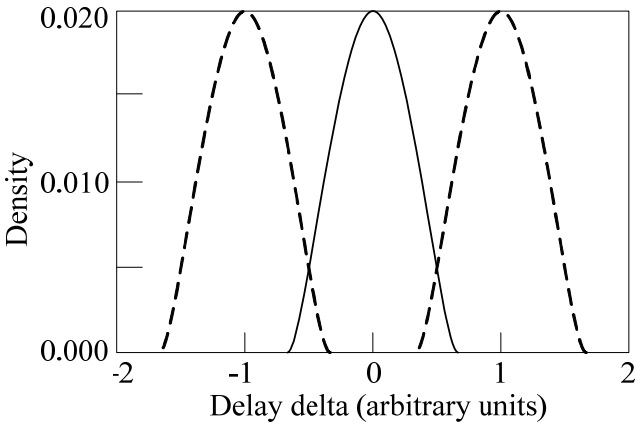


Figure 4. Probability function of distribution widths of within-die delay variation (regenerated from [8]).

according to ITRS. In terms of percentage, the threshold voltage shows the sharpest increase in variability across technologies. The control of the number of dopants in the depletion region of the transistor makes it increasingly difficult to achieve threshold voltage accuracy, and is therefore one of the factors that contribute to threshold voltage variability. Threshold voltage variations and various other sources of variation are considered in this work. In this work, intradie variability is addressed. A sketch of the intradie variation probability distribution of gates is shown in Figure 4 (regenerated from [8]), where a Gaussian distribution is used to describe the deviation from nominal parameters due to variation. The relative probability density of gates at a given difference from the nominal is plotted against the delay delta for a 90nm technology. The model shows that variation can occur such that the nominal distribution is shifted to either slower delays or faster delays. A wider distribution is certainly a worst indication of unsystematic variability and is harder to control or correct, given the uncertain nature of the distribution. To further understand the impact of process variations on global interconnect, the following subsections give a brief overview of the physical and environmental variations that are explored in this work.

1.2 Physical Variations on Interconnects

Maintaining circuit reliability in the presence of process variations becomes a harder task as we scale down further into the submicron regime. Manufacturing process issues become harder to deal with at such small dimensions, and accuracy is lost. In the lithographic process, masking is used to selectively add transistor functionality to a wafer. With technology scaling, the inherent shrinking of transistor and interconnect dimensions results in less accuracy in achieving transistors with "ideal" dimensions. These disparities may subsequently lead to circuit errors due to unexpected material overlap (short circuit) or gaps between metals that should be ideally connected (open circuit). The effect of this imprecision can be directly translated into uncertain behavior in terms of delay. This makes designers more concerned with having adequate margins that will allow for their given performance budgets. Such constrained designs can bring costly overheads in area and power, however, if the effects of the variations are not correctly characterized, there is no other way of guaranteeing the design will meet the specified budgets. To limit overdesign and still meet timing budgets, low-cost design solutions are proposed in this work to compensate for the physical variations seen at the interconnect level.

One physical parameter that can significantly affect the output current of a device is the effective channel length. This measure is a function of the drawn length (usually the feature size or technology node) and the drain to source overlap region, as shown in Equation 1.1. Manufacturing uncertainties lead to variation in the effective channel length, which in turn have a direct effect on the delay of a repeated interconnect, thus the significance of considering channel length variation as an important parameter to model.

$$L_{eff} = L_{drawn} - 2 * L_{overlap} \tag{1.1}$$

Another significant source of variation is uncertainty in the interlayer dielectric thickness. During the chemical mechanical polishing (CMP) process of chip manufacturing, certain areas of the chip polish faster than others because of uneven surface of the die. Sparse regions will polish faster than dense regions, thereby creating uneven dielectric thicknesses across the die. Metal height is an interconnect dimension that can also suffer from variations due to the die's uneven surface. Long interconnects can see much of this variation since they traverse a significant area of the chip, therefore it is relevant to consider these occurrences in an effort to improve interconnect performance at the design phase.

1.3 Environmental Variations on Interconnects

Environmental variations take place when the circuit is in the execution phase and is affected by environmental circumstances around it. Nothing can be done at the design phase to correct these effects, since it is hard to assert the exact environmental conditions the circuit will operate in once it has become a final product. An important example of an environmental phenomenon that has a substantial effect on chip performance is temperature. About 50% of integrated circuit failures is attributed to temperature [9], making this topic very important in design for circuit reliability.

With scaling, the effect of thermal issues on interconnects become more significant. Due to the dynamic nature of temperature variations, predicting thermal behavior on-chip is not an easy task. Temperature has a linear dependence on resistance, and since interconnect resistivity increases with scaling, the increase in temperature in smaller technology nodes causes degradation in the interconnect performance [10]. Hot spots on a chip are a function of the placement of high activity blocks, and long interconnects can see much of this thermal

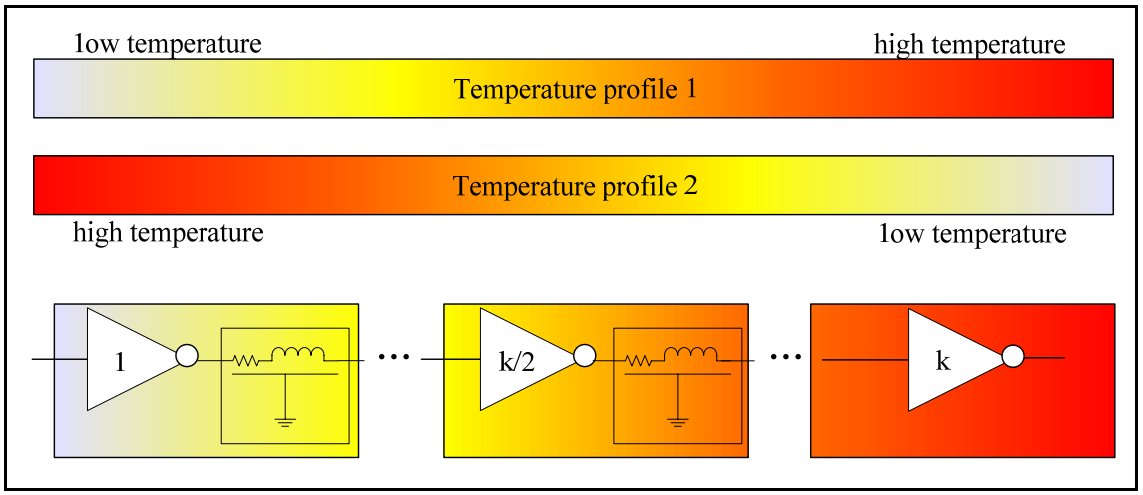


Figure 5. Increasing and decreasing spatial temperature profiles for a repeated interconnect.

variability. Therefore, it is very important that designers take thermal effects into consideration, and accurate thermal modeling becomes essential. While thermal impact on microprocessors has been discussed in the past by [9], this work will address the thermal impact on interconnect in terms of delay for 65nm and 45nm technology nodes. A characterization similar to what was proposed for repeated interconnects under the effect of L_{eff} device variations will be conducted to quantify thermal effects on repeated interconnects and determine how each component in the interconnect contributes to the total performance degradation.

Thermal modeling has become an issue of interest in variation research, since it is relevant to know how temperature varies on a chip, and how much of this variation can a single, long interconnect experience. Methodologies for studying thermal impacts on interconnects have been discussed in [10] and [11] with the objective of providing designers with rules to compensate for such variations. Thermal profile assumptions for wires are discussed in detail by [11], with two types of profiles being discussed at length: spatial (nonuniform heating) and temporal (uniform heating). For a temporal temperature profile,

the temperature is assumed to be the same along the length of the interconnect and is only assumed to change uniformly with time. A spatial temperature profile provides a more realistic scenario which implies that different regions of a long interconnect will encounter different temperatures at a given time, since a long interconnect can traverse a significant area of the chip. Figure 5 illustrates this concept with increasing and decreasing temperature profiles on a repeated interconnect (only the increasing profile is pictured along the interconnect, however, both variants are depicted on top, for clarity). The temperature gradients are illustrated as continuous, but for the purposes of modeling and simulation of this work, the gradients have been discretized along the line. The thermal modeling analysis described will be used in Chapters 3 and 4, to characterize the effect of thermal variations on repeated interconnects and phase coded interconnects, respectively.

1.4 Interconnect Modeling

Repeater insertion is the most widely used interconnect technique, because of its reduction of propagation delay dependency on interconnect length from quadratic to linear. Because of its known nature and popularity, repeater insertion will be the primary technique analyzed in this work. It is appropriate then to introduce the interconnect model assumptions that will be used throughout. Table 1 shows the interconnect assumptions for a global interconnect layer as obtained from [12]. The data for 45nm and 32nm technology nodes has been scaled from the available typical global interconnect dimensions. The scaling factor "1/s" used is the ratio between the smaller technology node and the previous node. Figure 6 illustrates a repeated interconnect with the segmenting assumptions that are used throughout this work, with L being the total interconnect length, H the repeater size (with respect to a

Table 1. Interconnect assumptions for a global interconnect layer [12].

	2007	2010	2013
Tech node [nm]	**65**	**45**	**32**
Vdd [V]	1.1	1	0.9
Width [nm]	450	310	220
Spacing [nm]	450	310	220
Height [nm]	200	138	98
Dielectric Constant	2.9	2.5	2.3

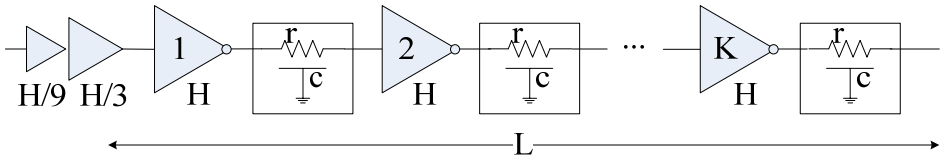


Figure 6. Repeated interconnect model used throughout this work. Each interconnect segment box represents a 5-pi model. R, c and l should be scaled according to the number of segments and length of the line.

minimum sizing assumption), K the number of repeaters along the line, c, r and l the capacitance, resistance and inductance per unit length of the corresponding wire segment, respectively.

In order to determine an adequate number and size of repeaters for a given wire length, an experimental delay-optimal technique was used. This consisted in obtaining the H and K that will produce the lowest delay for a wire of length l, such that the signal has a slew rate of 50%. This ensures that we are not just picking the properties that result in the lowest delay, but that we are also considering the quality of the arriving signal. Table 2 summarizes the experimental findings for 65, 45 and 32nm global interconnects using wire dimensions and transistor model cards from PTM. As expected, the number of repeaters is linear with respect to the wirelength, while the size does not follow a particular trend. Therefore, a good measure of comparability is the *effective* size of the repeaters, namely the number of repeaters times the size of the repeater. Nonlinearity is due to the restrictions set within the

Table 2. Experimental results for repeater number and size for 1mm - 4mm interconnects.

tech[nm]	wirelength[mm]	#repeaters	size[*minsize]	delay[ps]	power[mW]	energy[fJ]
65	1	2	45	46.78	0.097	4.6
	2	4	45	102.2	0.212	21.7
	3	6	46	155.4	0.327	50.9
	4	8	45	210.2	0.439	92.2
45	1	2	47	50.47	0.062	3.1
	2	4	53	105.3	0.139	14.7
	3	6	53	161	0.215	34.6
	4	8	50	218.4	0.282	61.5
32	1	2	88	53.99	0.052	2.8
	2	4	81	118.9	0.108	12.8
	3	6	118	199.7	0.219	43.7
	4	10	51	250	0.187	46.9

experimental method, which favor certain scenarios of delay and skew combinations over others. Integer effects may also account for the nonlinearity of the results.

Although repeater insertion is shown to work well for long interconnects, the analyses assume ideal transistor behavior. However, in real life this is not the case. Designers must consider error margins that allow for unpredictable circuit behavior, without compromising on the actual performance. This translates into a penalty in design, sometimes geared towards worst-case considerations, that prevent the end user from taking advantage of the circuit's full potential. With this in mind, researchers have attempted to establish process variation-aware techniques to allow more circuit flexibility at the design phase. With this in mind, this work characterizes variations on repeated interconnects, setting the importance of variation impact on performance in current and future technologies, and providing designers with guidelines to follow to guardband their designs against variation while avoiding overhead in performance.

1.5 Document Structure

Now that an introduction of the various types of variation considered in this work has been given, Chapter 2 will propose the first technique used at the design phase to correct the effect of effective channel length (L_{eff}) variation on delay of a global interconnect. Chapter 3 will consider several process variation sources to account for correlations amongst certain parameters and the fact that even if the parameters are not correlated, they are common variations seen in today's technologies. Once these variations have been characterized, Chapter 4 will discuss the idea of phase coding for on-chip interconnects, with the goal of more effectively utilizing the bandwidth of the interconnect to encode data bits along the line. The phase coded interconnect is then observed in the presence of process variations, to determine the impact due to the additional circuitry. Finally, Chapter 5 will draw conclusions on the observed behavior and offer insight into future work.

CHAPTER 2

SUPPLY VOLTAGE ASSIGNMENT FOR REPEATED INTERCONNECTS

Global on-chip interconnects exhibit increased delay as CMOS technology scales down. One widely used circuit technique to cope with this problem is repeater insertion, which reduces delay dependence on wirelength from quadratic to linear. As repeater size increases, power increases and delay decreases. Designers are continuously exploring optimization techniques to reduce repeater delay and power. In addition, other repeater optimizations can be explored to reduce the effects of process variations. Designing for process variations is an important step towards meeting timing budgets within the allowable power constraints. This chapter focuses on exploring the effects of intra-die, channel length (L_{eff}) variation on interconnect performance, and establishes a relationship between assignment of supply voltages and reduction of the delay distribution.

2.1 Effect of L_{eff} Variation on Interconnect Performance

Effective channel length (L_{eff}) variation in MOSFETs can significantly alter the output current of the device [7], and hence, cause delay uncertainty. With delay uncertainty there is unnecessary waste of power due to the size of repeaters that are designed to meet the delay requirements. The International Technology Roadmap for Semiconductors (ITRS) [1] provides designers with parameter variation values for current technology nodes and predictions for future technology nodes, based on real-time

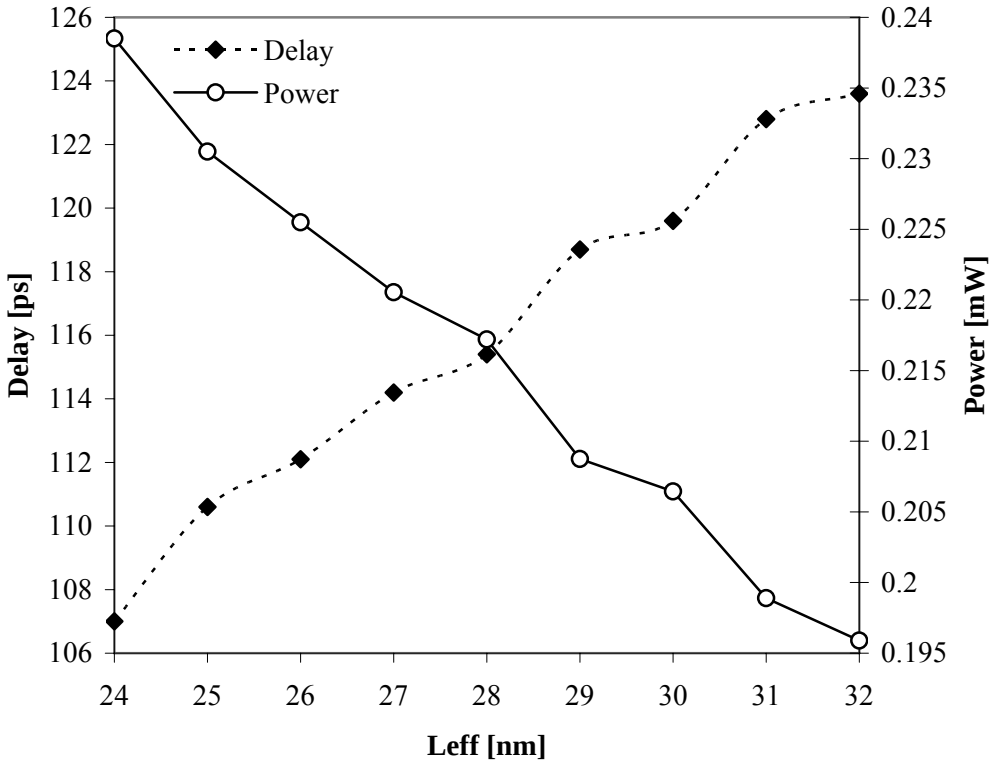


Figure 7. Impact of effective channel length (Leff) variation on interconnect delay and power on a 2mm repeated interconnect in 70nm technology node.

measurements. Using data provided by ITRS and other industry sources for the 70nm technology node as in [13], the impact of intradie, process-induced L_{eff} variation on the delay and power of a repeated interconnect structure was investigated. In response, a suitable technique is proposed to mitigate the effect of L_{eff} variation on a delay distribution consisting of 1,000 random samples.

The L_{eff} for the NMOS and PMOS transistors is obtained randomly from a Gaussian pattern within the given 3σ threshold of variation provided by ITRS. In theory, L_{eff} for an NMOS and a PMOS device is given by Equations 2.1 and 2.2, respectively, as a function of the gate-drain and gate-source overlap region (L_{int}), and the drawn length which is usually the feature size (L_{drawn}).

$$L_{effN} = L_{drawn} - 2 * L_{\text{int } N} \quad (2.1)$$

$$L_{effP} = L_{drawn} - 2 * L_{\text{int } P} \quad (2.2)$$

The trend of L_{eff} on delay and power of a repeated interconnect is shown in Figure 1. L_{eff} was swept from 24nm to 32nm for a 2mm repeated interconnect, and delay and power were observed. As expected, delay of the repeated wire varied proportionally with respect to L_{eff}. Power decreases as L_{eff} increases due to decrease in current through the repeaters. For a 14% variation from the nominal L_{eff} of 28nm, a worst case delay variation of 7% was observed, while the worst case power variation was 10%. Such variations may threaten timing and/or power budgets, and implementing solutions at design time may compensate for such variations when the device is put to use.

The Predictive Technology Model (PTM) provides accurate model files for current and future technology nodes, which are compatible with circuit simulators such as SPICE [14]. In this work, PTM models are used for the devices and typical interconnect dimensions for the global interconnect layer are also obtained from PTM to determine the R, L and C for a given wirelength. It is assumed that NMOS and PMOS transistors have different resulting L_{eff} values due to the inherent uncertainty in manufacturability, therefore different random values are assigned to each device for simulation purposes. The L_{eff} values that were used in this work, from [13], are a 28nm nominal with a 3σ process tolerance of ±16.7%. Since PTM models explicitly make use of L_{int} rather than L_{eff} (and assuming L_{drawn} is constant), this process tolerance translates into a nominal L_{int} value of 21nm with a variation of ±11.1%. A 2mm repeated interconnect was simulated to analyze the impact of these variations by using a Monte Carlo analysis to obtain random values of L_{int} from a Gaussian distribution according to the 3σ values.

2.2 Supply Voltage Assignment Methodology

To correct the effects of L_{eff} variation shown in the previous section, the relationship between supply voltage and interconnect delay has been used to reduce the delay distribution. In this work, the delay distribution consists of 1,000 samples of interconnect delays of a 2mm global interconnect in the presence of random patterns of L_{eff} variation. A Monte Carlo methodology is used to draw random samples of L_{int} from a Gaussian distribution, and these in turn are assigned to the corresponding device in the interconnect. After all devices have a randomly assigned L_{int} value, the interconnect is simulated in Hspice [14] to determine the output delay. This output delay is then compared to what the output delay of the interconnect should be if all devices were ideally drawn (interconnect in the presence of no process variations yields the nominal delay).

The relationship of supply voltage (V_{dd}) and delay is the following: delay decreases as V*dd* increases, thereby increasing the speed of the repeated interconnect. Conversely, delay increases as V_{dd} decreases. Power also increases as V_{dd} increases, since there is a quadratic dependence of power on $V_{dd,}$ as supported by the theoretical definition of power shown in Equation 2.3.

$$P_{total} = \left(CLV_{dd}^2 + V_{dd} I_{peak} t_s\right) f_{0\to1} + V_{dd} I_{leak} \tag{2.3}$$

It was shown in the previous section that due to process-induced L_{eff} variation there is delay uncertainty in repeated interconnects. The V_{dd} assignment technique is based on the idea that delay can be controlled by the given supply voltage at initialization phase, regardless of the variation on the devices resulting from the manufacturing phase. By analyzing the delay distribution obtained after applying the variations to 1,000 interconnect samples, an appropriate V_{dd} is assigned to each interconnect as needed, to bring the delay

closer to the nominal delay value. The allowed proximity of the new delay to the nominal delay is determined by a tolerance value. The tolerance is assigned such that all new delays that fall within this number from the nominal will be accepted as close enough to the nominal value. The appropriate V_{dd} for each interconnect is obtained by binning delay values. The faster bins are assigned a lower V_{dd} and the slower bins are assigned a higher V_{dd}. After the assignments, the new delay obtained is within the tolerance range around the nominal delay value. The granularity of the assigned V_{dd}s will determine how accurate this assignment will be. Certain delays may never be brought close enough to fall within a set tolerance, due to granularity constraints. For example, for the nominal delay of 114.8ps, with a 1ps tolerance, only values that fall within the range of 113.8ps – 115.8ps will be accepted as part of the new, reduced delay distribution. Given a granularity of 100mV, when the interconnect delay is less than 113.8ps, the V_{dd} will be set to 100mV less than its previous value in an attempt to bring it within the 1ps tolerance constraint. However, when the new V_{dd} is applied, the interconnect is now slower than what the tolerance will allow (surpassing the 115.8ps maximum accepted value). This problem is only solved by increasing the granularity at which V_{dd}s are assigned, i.e. assigning a granularity of less than 100mV, such that more precise delay values can be obtained. This indicates that no matter what the effect of L_{eff} variation is on the delay, we can always assign a new V_{dd} value that can correct this uncertainty. However, this is an unrealistic solution, since it requires a lot of computation effort, especially in large and complicated systems, in which many long interconnects could undergo this technique. In addition, such fine granularity requires the availability of fine-tuned voltages that can accomplish the task of reducing the variation.

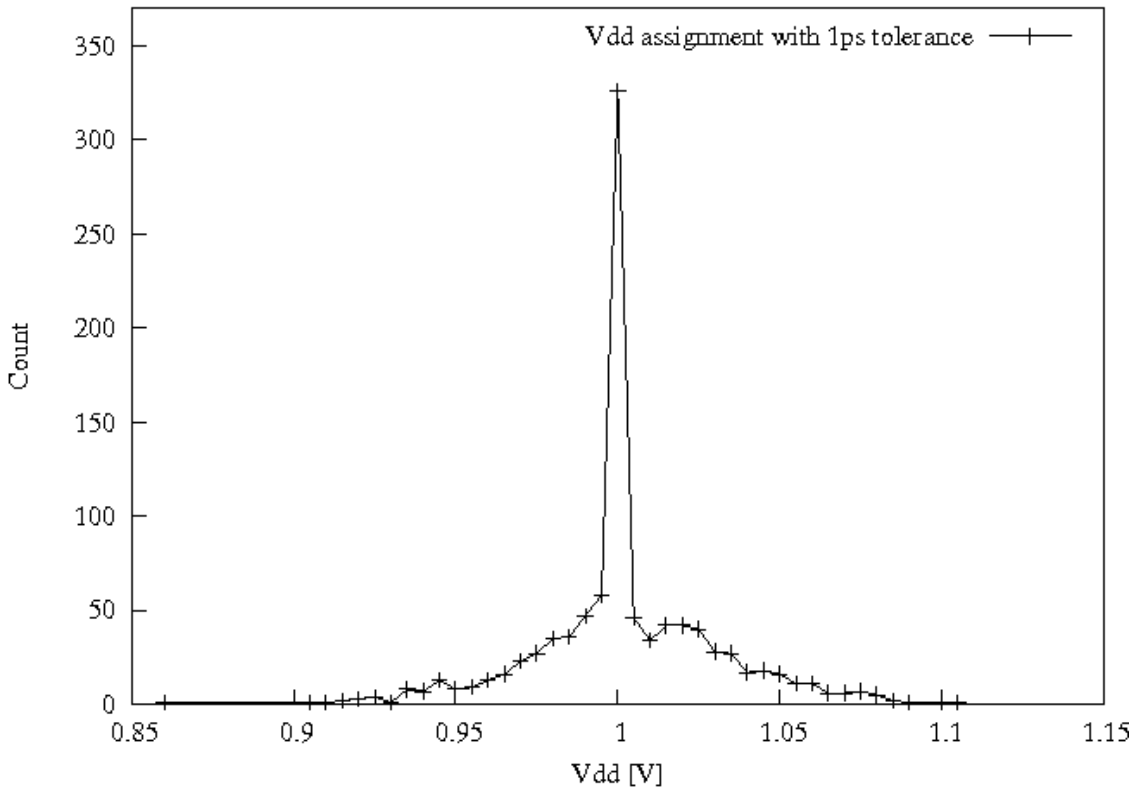


Figure 8. Supply voltage distribution as assigned by the proposed technique for 1,000 interconnect samples with different L_{eff}'s for each device.

Adaptive V_{dd} techniques have been proposed in the past by [15] for low swing interconnects to reduce power consumption. This work focuses on reducing the delay spread due to process-induced parameter variations with specific focus on L_{eff} with negligible power overhead. Therefore, the power distribution after V_{dd} assignment will directly depend on the shape of the V_{dd} distribution.

For this work, 70nm PTM models were used with nominal V_{dd} being 1V. The P/N skew was determined as 2.57, using an inverter's voltage transfer characteristic curve. To mimic a random variation in L_{eff} , Hspice Monte Carlo analysis was used based on a Gaussian distribution. A total of 1,000 Monte Carlo runs were used to analyze the impact of L_{eff} variation. During each iteration, a random value of L_{eff} was generated for NMOS and

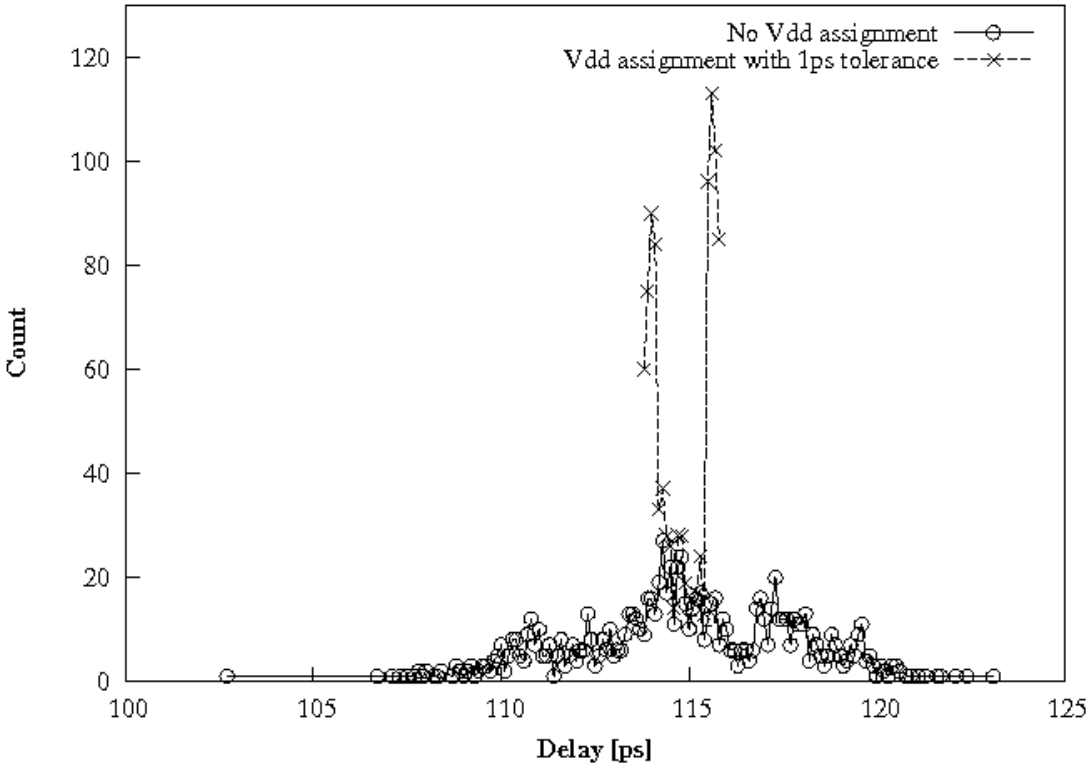


Figure 9. Delay distribution before and after supply voltage assignment for a 2mm repeated interconnect.

PMOS transistors. The delay and power due to L_{effN} and L_{effP} variation was then recorded for each value. Each of the 1,000 runs was then analyzed separately to determine the V_{dd} value that would bring the delay into the desired tolerance of ±1ps from the nominal (114.8ps for a 2mm line with 4 repeaters). The V_{dd} distribution after applying the V_{dd} assignment technique to a repeated line is shown in Figure 7. The bulk of the distribution falls at or within the 1V mark, since most of the interconnect delays already fall within the set tolerance range. The V_{dd}s that were assigned stay within a range of 0.2V from the nominal 1V, implying no overwhelming difference from the nominal to achieve delay within the established tolerance. The impact on delay can be observed in Figure 9 where it is evident after assigning V_{dd}s, the delay distribution is significantly narrower. Due to the nature of the implementation, most

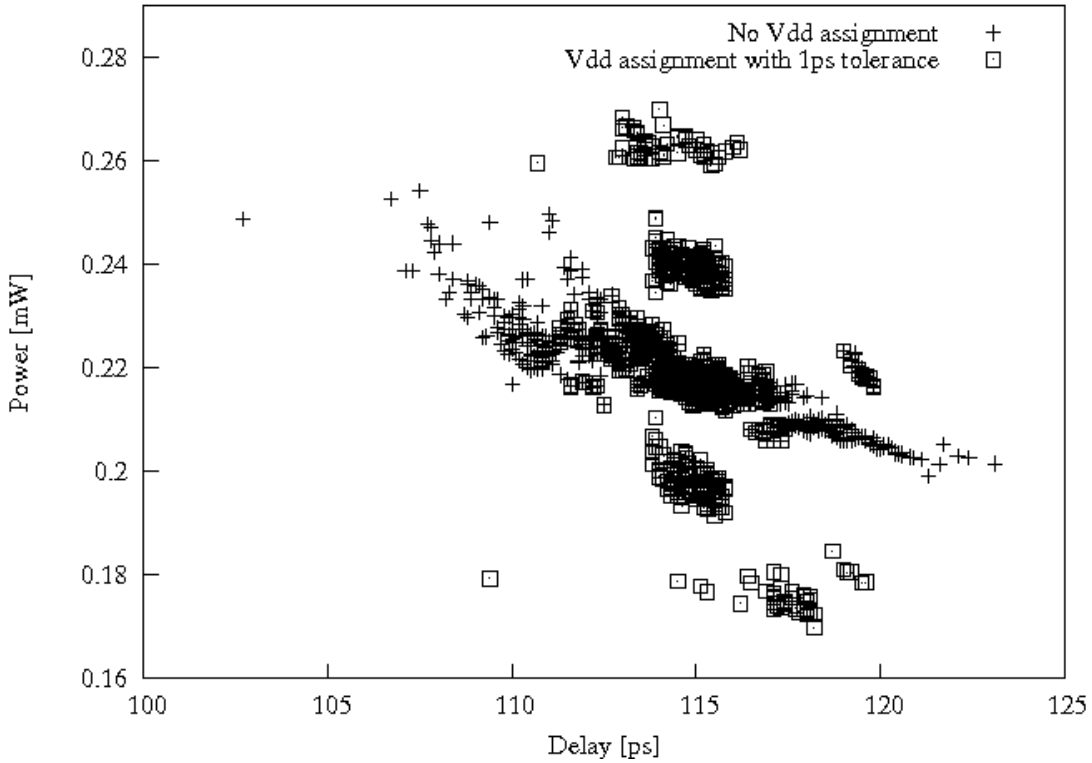


Figure 10. Power vs. delay tradeoff for a 2mm interconnect before and after supply voltage assignment.

delay occurrences occupy the outer corners of the new, reduced distribution. Once the algorithm finds a solution that falls within the required tolerance, it will not try to optimize further or look for a previous solution that may accommodate the current interconnect (a voltage which is required for a previous solution may yield better performance for the current interconnect than the first solution the algorithm has found). This has not been implemented to avoid computation complexity.

The new distribution shows a reduction of 90% in the delay spread. Only 0.74% penalty in power is observed, which implies that the delay spread optimization can be achieved with negligible power penalty. The power vs. delay tradeoff can be observed in Figure 10. Again, the narrower shape of the new delay distribution is noticeable, but the power range has increased. Even so, the technique does not cause a significant additional

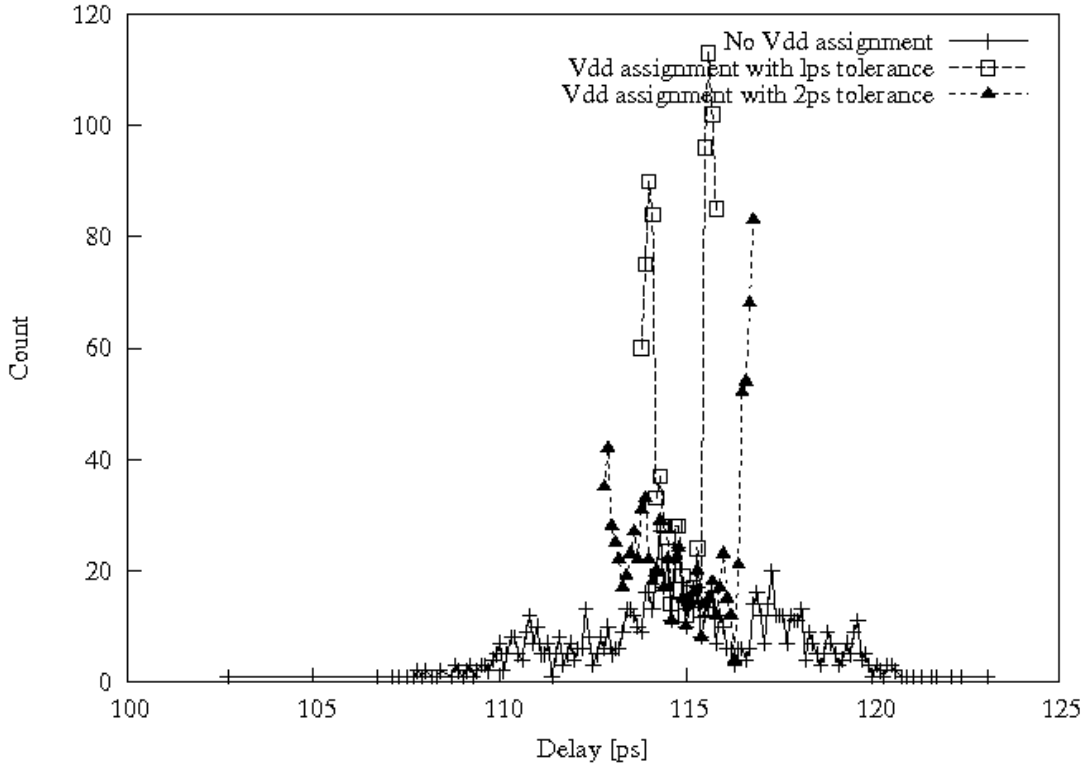


Figure 11. Delay distribution comparison for supply voltage assignment with tolerances of 1ps and 2ps.

power overhead as shown before.

The effect of changing the delay tolerance from ±1ps to ±2ps when assigning V_{dd}s was analyzed. V_{dd}s were assigned over a range of 225mV, a slight increase over the 200mV range when setting a ±1ps tolerance constraint. The resulting delay distribution is narrower, although the reduction with respect to the 1ps tolerance distribution is not very significant.. Nonetheless, a delay spread reduction of 80% is obtained with a power overhead of 0.32%. The comparison of the delay distributions without V_{dd} assignment, with assignment of 1ps tolerance, and with assignment of 2ps tolerance can be observed in Figure 11. Similarly, the power vs. delay tradeoff for 1ps and 2ps tolerances is observed in Figure 12. The power overhead decrease when using a 2ps tolerance (as opposed to using 1ps tolerance) is not

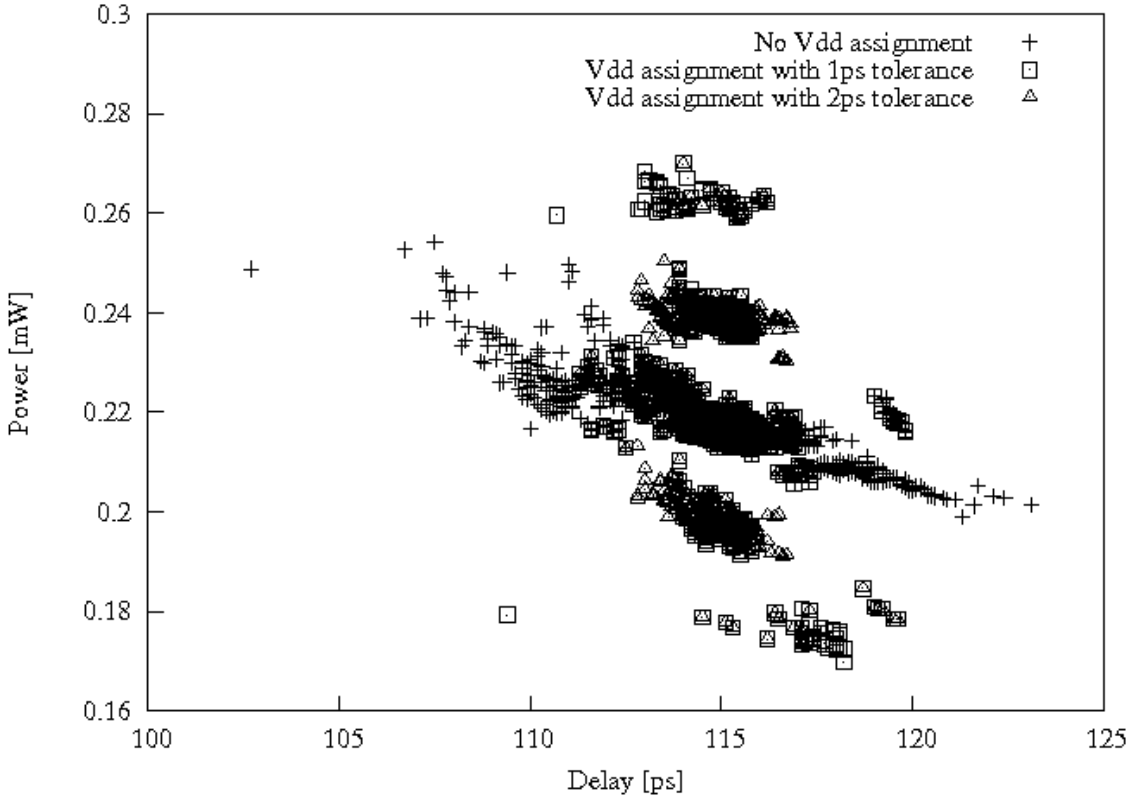


Figure 12. Power vs. delay tradeoff comparison for supply voltage assignment with tolerances of 1ps and 2ps.

readily observed from the figure, since it is such a small difference in percentage. The designer now has the option of choosing a more flexible delay tolerance, while still obtaining a very significant benefit in terms of delay distribution reduction with a minimal power overhead penalty.

2.3 Advantages and Tradeoffs of Supply Voltage Assignment

The V_{dd} assignment method depends on the availability of multiple supply voltages and in this work, the assumption is that it is possible to provide those supplies. Jeong et al. [15], for instance, proposed several methodologies which can be used to supply the multiple V_{dd}s. In addition, since multiple V_{dd}s are used in this work, level converters are required at

the end of each interconnect to restore the signal levels back to nominal voltages. Another assumption is that each repeated interconnect is shielded by a supply line running parallel to it and will be used by the repeaters to draw the supply.

The proposed technique reduces the overall delay spread that results from process-induced L_{eff} parameter variation with a negligible power overhead. Multiple V_{dd}s are assigned at initialization time only, in order to reduce the delay spread. Since the V_{dd} assignments are carried out at initialization time, there are no power overheads as incurred by many adaptive techniques. By using a V_{dd} assignment technique with 1ps tolerance, the delay spread is reduced by 90%, with a penalty in power of only 0.74%. For more flexibility, a 2ps tolerance constraint can be set, which will still deliver 80% reduction in the delay distribution with a negligible power overhead of 0.32%.

CHAPTER 3

THERMAL VS MANUFACTURING EFFECTS IN ON-CHIP INTERCONNECT TIMING

Static and dynamic variations are becoming a bigger concern as transistors scale. Due to manufacturing inaccuracies, transistor dimensions may not be as precise as accounted for during the design stage. Uncertainty in performance is expected if these factors are not modeled in the design. Given a performance budget, meeting timing constraints is a very important part of obtaining optimal design performance. Furthermore, dynamic variations are an unpredictable source of timing failures and need to be addressed accordingly. This chapter discussed both spatial and thermal variation effects on global interconnect design.

3.1 Variation Effects on Global Interconnect Performance

As transistor and wire dimensions shrink with each new technology node, designers face challenges that compromise system performance. One of the major challenges that are widely explored today is process variation. Smaller transistors inherently imply shorter channel lengths, and with shorter lengths there will be more leakage and less control of the device threshold voltage. In addition, inaccuracies during the lithographic process can cause uncertainty in device dimensions, while interconnect construction can introduce uncertainty in wire dimensions which will directly affect the resistance, capacitance and inductance of the line. All these factors contribute to delay uncertainty that can potentially cause timing failures if not accounted for at design time.

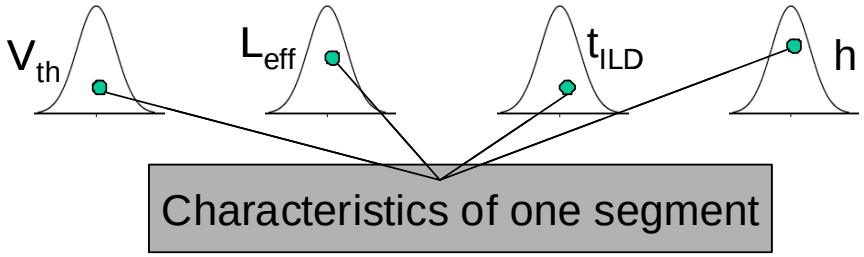


Figure 13. Sources of static variation considered in this thesis.

Static variations correspond to those physical parameters which are not expected to change at run-time. The effective channel length is a parameter of concern because of its direct impact on output current. Similarly, the threshold voltage is also a parameter of interest. Since a change in effective channel length will affect threshold voltage, they are said to be correlated parameters. As mentioned before, inaccuracies in wire dimensions are also present. Two dimensions of interest are the interlayer (ILD) thickness and the height of the metal (h). During chemical-mechanical polishing (CMP), sparse regions of a chip polish faster than dense regions, causing an uneven ILD thickness across the chip.

In order to study the impact of static variations on repeated on-chip interconnects, the sources of variation of interest must be simultaneously affecting the wire. Assuming threshold voltage (V_{th}), effective channel length (L_{eff}), interlayer dielectric thickness (t_{ILD}) and metal height (h) variations, each of them is modeled independently as a Gaussian (normal) distribution statistical source as shown in Figure 13. A wire segment of a repeated interconnect consists of a repeater device and a wire section, typically represented with a 5-pi model, shown in Figure 14. In this figure, the driving repeater and the 5-pi wire constitute a segment. In this work, the Gaussian distributions are obtained by generating a Monte Carlo simulation of 1,000 samples in Hspice. The source distributions are based on data from ITRS and industry sources shown in Table 3.

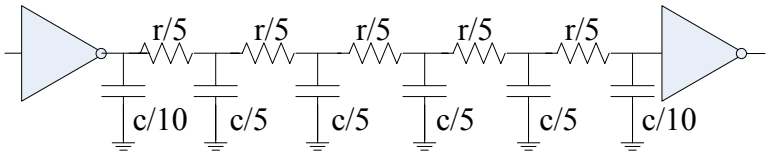


Figure 14. 5-pi wire model between two repeaters.

Table 3. Device and wire variation assumptions (from ITRS and industry sources).

	65nm	45nm	32nm
Threshold Voltage (V_t)	33%	42%	58%
Effective Channel Length (L_{eff})	12%	12%	12%
Interlayer Dielectric Thickness (t_{ILD})	10%	10%	10%
Metal height (h)	10%	10%	10%

This table shows percentages corresponding to the whole distribution; therefore, in terms of statistical 3σ notation, the 3σ variation for each of these parameters is half of the percentage shown. Since the effective channel length has an effect on threshold voltage, and interlayer dielectric thickness and metal height must meet a certain aspect ratio, these two pairs of variables are correlated within each pair. To satisfy the V_t-L_{eff} correlation and for simplicity of modeling, this work assumes that the percentage of variation shown for V_t encompasses the effect seen by L_{eff}, so the latter is not modeled in the work presented in this chapter. To satisfy the t_{ILD}-h correlation, both parameters must vary in the same direction, i.e. if t_{ILD} is larger than the nominal for a random case, metal height must also vary positively and take on a larger value than its own nominal.

When wire dimensions are affected by process variation, the resistance, capacitance and inductance are different from their expected values. Since this work considers t_{ILD} and metal height variations, these are incorporated into the design by dynamically recalculating R, C and L. This work assumes an RC wire model; therefore inductance will be ignored for

now. Equations 3.1 through 3.6 are used for recalculation of R, C and L [12] and Table 4 lists the parameters used in the equations and their value, when applicable.

$$R = \frac{\rho \cdot l}{w \cdot h} \tag{3.1}$$

$$L_{self} = \frac{\mu_0 \cdot l}{2\pi}\left[\ln\left(\frac{2 \cdot l}{w+h}\right) + \frac{1}{2} + \frac{0.22(w+h)}{l}\right] \tag{3.2}$$

$$M = \frac{\mu_0 \cdot l}{2\pi}\left[\ln\left(\frac{2l}{d}\right) - 1 + \frac{d}{l}\right] \tag{3.3}$$

$$C_g = \varepsilon\left[\frac{w}{t_{ILD}} + 2.22\left(\frac{s}{s+0.70 \cdot t_{ILD}}\right)^{3.19} + 1.17\left(\frac{s}{s+1.51 \cdot t_{ILD}}\right)^{0.76} \cdot \left(\frac{h}{h+4.53 \cdot t_{ILD}}\right)^{0.12}\right] \tag{3.4}$$

$$C_c = \varepsilon\left[1.14\frac{h}{s}\left(\frac{t_{ILD}}{t_{ILD}+2.06 \cdot s}\right)^{0.09} + 0.74\left(\frac{w}{w+1.59 \cdot s}\right)^{1.14} + 1.16\left(\frac{w}{w+1.87 \cdot s}\right)^{0.16} \cdot \left(\frac{t_{ILD}}{t_{ILD}+0.98 \cdot s}\right)^{1.18}\right] \tag{3.5}$$

$$C_{total} = C_g + 2 \cdot C_c \tag{3.6}$$

Table 4. Wire equation parameters and description.

l	Wire length.
w	Wire width.
h	Metal height.
t_{ILD}	Interlayer dielectric thickness.
s	Spacing between wires.
d	Center-to-center distance between two wires.
ρ	Resistivity. For copper, it is 2.2 uohm-cm.
μ_0	Absolute permeability of air = 4 Pi x 10^{-7} H/m.
R	Wire resistance.
L_{self}	Self inductance.
M	Mutual inductance.
C_g	Ground capacitance.
C_c	Coupling capacitance.

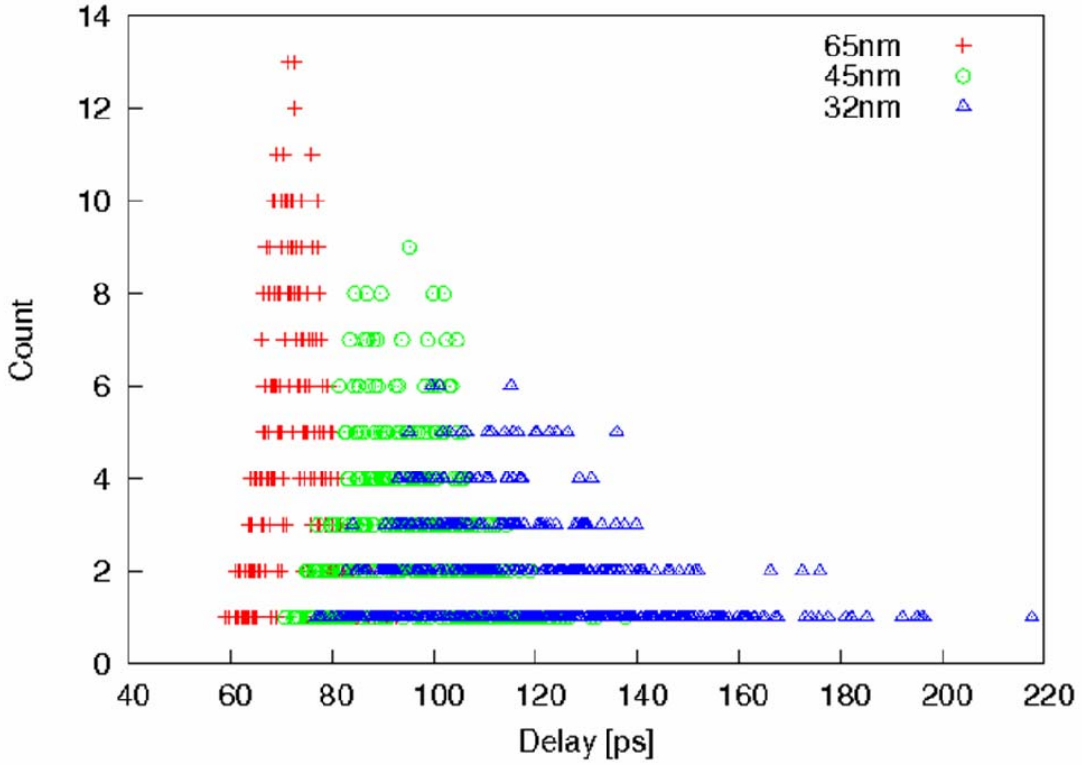


Figure 15. Delay distribution of a 1mm repeated interconnect in the presence of random process variations for 65nm, 45nm and 32nm technologies.

The delay distribution of a 1mm repeated interconnect in the presence of 1,000 random scenarios of process variation is shown in Figure 15. The wire is simulated at a cold temperature of 70ºC. Delay values are binned accordingly, with a granularity of 0.1ps. The 32nm curve is farthest to the right because the dimensions of the wire in that technology are the smallest of the three, and resistance is the largest (see Equation 3.1). The distribution for 32nm is also the widest, since the smaller the technology, the more variation the wire is likely to experience. For each smaller technology, the distribution widens by approximately 50% with respect to the distribution of the previous (larger) technology node. The distributions flatten out with increasing technology node due to the wider delay spread. They also follow a Gaussian shape, since the sources of variation are assumed to be Gaussian. We

can conclude that static process variations become more significant as we scale down transistor and wire dimensions, and delay spread becomes a critical concern

3.2 Thermal Effects on Global Interconnects

Even though static variations have a significant impact on performance, as the industry moves to future technologies and the power density of chips increases, temperature is a factor that can greatly compromise the performance of a chip. Thermal budgets are rapidly decreasing in order to accommodate the demand for larger integration and higher-frequency chips [16]. High activity areas can experience harsh gradients that can degrade performance significantly. IBM's Power4 microprocessor has shown thermal gradients as large as ~14ºC for a 1mm distance [17]. Therefore, it is important to explore the performance effects of harsh temperature gradients on interconnects, and aggressively design to compensate for such effects. [11] showed that nonuniform heating (or spatial temperature variation) can considerably degrade interconnect performance.

Two typical classifiers for thermal variation are spatial and temporal variation. Temporal variation indicates that temperature changes with time, but is uniform along the interconnect at a given time instant. Spatial (or nonuniform) temperature variation indicates a thermal gradient along the interconnect at a given time instant. Since gradients of ~14ºC are possible at a 1mm distance, such a harsh thermal change in a given region can significantly harm performance.

Temperature variations for this work are modeled in both the device and the wire. Device temperature is specified through the DTEMP transistor parameter in Hspice. Wire temperature is modeled by using DTEMP for each component of the wire (each R and C) and

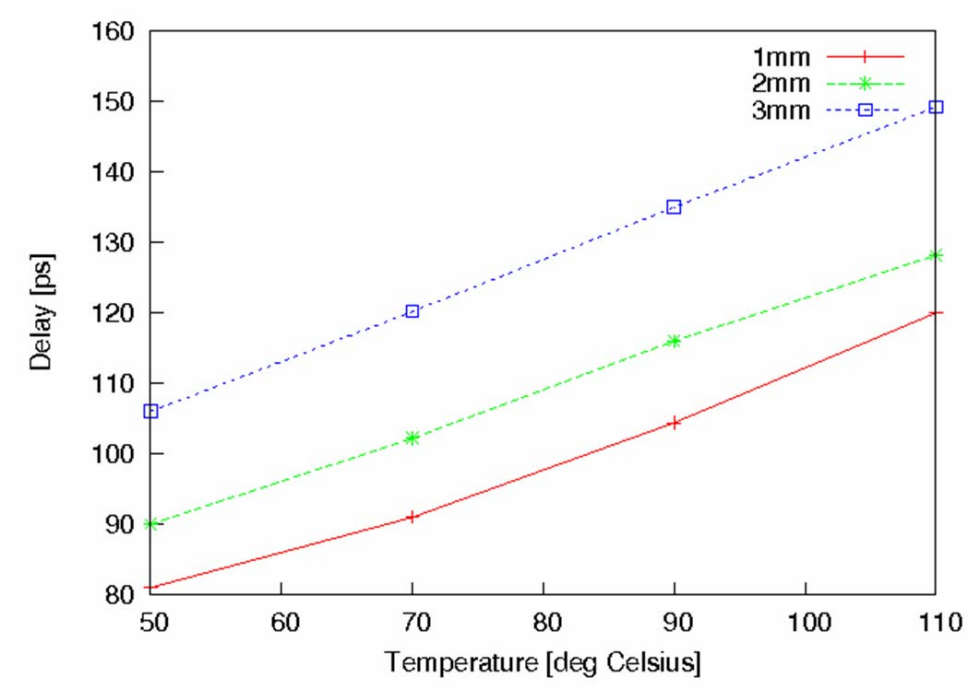


Figure 16. Delay dependence on temperature for a 45nm repeated interconnect of 1mm, 2mm, and 3mm wirelengths.

an additional temperature coefficient TC1 is given. TC1 depends on the metal assumption; for copper, TC1 is 3.9e-3/ºC. Using these assumptions, the linear dependency of delay on temperature for a 45nm repeated interconnect of various lengths is shown in Figure 16. Longer wires have longer delays because of the larger resistance component, but longer wires do not necessarily show a bigger delay dependency on performance. Figure 17 shows the delay dependency on temperature for a 1mm repeated interconnect in 65nm, 45nm, and 32nm. From Figure 16, for a 20ºC temperature gradient on a 1mm repeated interconnect in 45nm technology, there is a 22% degradation in performance. Clearly, this is an unacceptable performance change that if not accounted for, can lead to unwanted timing failures.

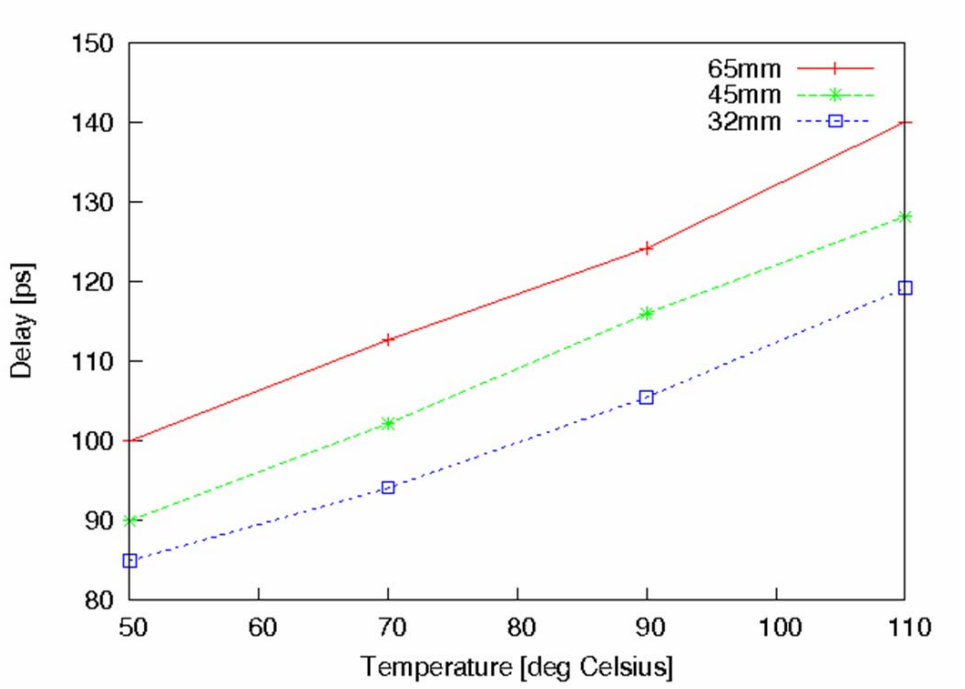


Figure 17. Delay dependence on temperature for a 1mm repeated interconnect in 65nm, 45nm, and 32nm technologies.

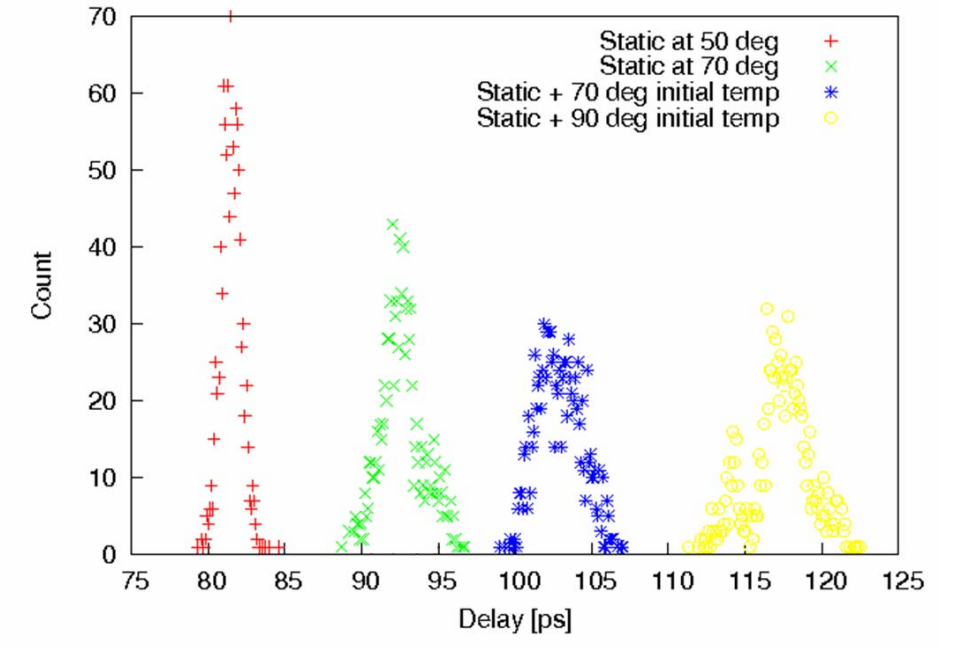


Figure 18. Delay distributions for different thermal profiles on a 1mm repeated interconnect in 45nm technology. The delay distributions also reflect the same static variations for all cases.

Figure 18 shows the delay distributions for a 1mm repeated interconnect in 45nm technology, operating under the effects of static variations, and for both static and thermal variations simultaneously. The red plot shows the delay distribution of an interconnect operating at a uniform 50ºC temperature and only experiencing static variations. The green plot shows the same situation for a 70ºC uniform operating temperature. The blue plot shows the combined effects of static and thermal variation on the interconnect. The thermal variation for this case is modeled by an increasing temperature gradient along the interconnect. Since a 1mm interconnect has 2 repeaters according to the repeater insertion optimization guidelines followed in this work, the thermal gradient will be discretized into three regions. The thermal gradient assumption is 20ºC along a 1mm interconnect. The cascaded drivers to the interconnect will operate at a nominal temperature of 70ºC. The first segment of the interconnect (device and wire section) will operate 10ºC higher, i.e. at 80ºC. The second segment of the interconnect will operate 10ºC higher than that, i.e. at 90ºC. When compared to the scenario of just having static variation, the delay distributions spans about the same delay range (~8ps). However, when increasing the temperature profile along the interconnect to a 90ºC starting temperature, the delay uncertainty increases by ~40% with respect to the 70ºC distributions. This implies that temperature has a higher impact on the delay distribution of an interconnect than only having static variations.

3.3 Variation Effects on On-Chip Interconnect Timing

Critical lines, such as clock paths are very vulnerable to the effects of process variation, since variation can induce negative clock skew which will adversely impact performance. Numerous algorithmic solutions, such as [18,19] aim to re-optimize the clock tree network topology in the presence of temperature variations to compensate for

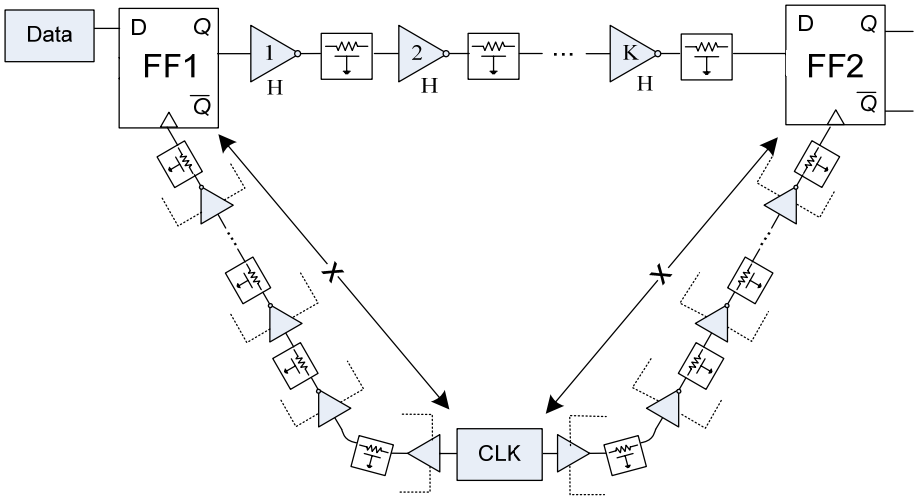


Figure 19. Data path clocked by two leaves of an H-tree.

performance. [20] uses a set of benchmarks to show that spatial thermal variations do in fact cause dynamic delay variations that can cause timing violations on a wire.

With temperature and process firmly established as significant sources of clock skew, many researchers have developed techniques to compensate for skew in the presence of thermal variations: [21,22,23]. To study the effect of thermal and process variations on clock skew, a data path clocked by two H-tree leaves has been modeled (Figure 19).

The clock source is assumed to be equally spaced from both the driving and the receiving master-slave registers (FF1 and FF2, respectively). This ensures a zero-skew nominal condition under an ideal environment. Data is sent along a 45nm, 1mm repeated interconnect and is captured by FF2 at the end of the line. For the purposes of this work, both clock paths and the repeated interconnect are 1mm long, global interconnects. The power-optimal number *K* and size *H* of repeaters per mm was obtained experimentally with HSPICE. BSIM4 Predictive Technology Models (PTM) [13] are used for all simulations.

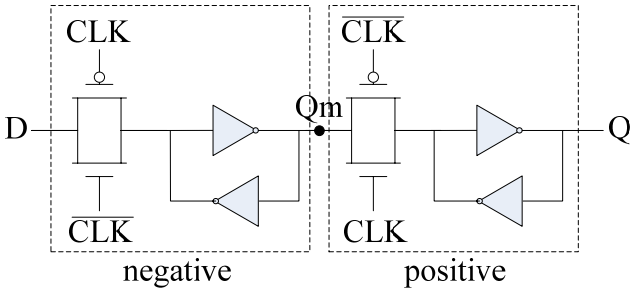


Figure 20. Implementation of each flip flop using a transmission gate configuration.

The minimum clock period (T) for this design is given by Equation 3.7,

$$T \geq t_{c-q1} + t_{line} + t_{su2} \tag{3.7}$$

where t_{c-q1} is the data propagation delay of FF1, t_{line} is the propagation delay of the repeated interconnect and t_{su2} is the setup time of FF2. This constraint is set for an ideal case, where no variation is present.

The master-slave register used for this work is illustrated in Figure 20. It consists of a negative latch followed by a positive latch and has input D and output Q. Since two master-slave flip flops are used in our design, we must analyze the timing of the system as a whole, as shown in Figure 21. CLK1 corresponds to the signal clocking FF1, while CLK2 corresponds to the signal clocking FF2. The ideal timing scenario shown in Figure 21(a), corresponds to the case when there is no variation present in the interconnect or the clock paths. The clock inputs to the flip flops, CLK1 and CLK2, respectively, are perfectly synchronized since the source clock is at an equal distance from each flip flop clock signal. The setup time for D with respect to CLK1 is met, therefore D propagates to Q1, where Q1 is the output of the first flip flop. After a time t_{line} during which the data propagates through the interconnect, the next rising edge of the clock is encountered and in this case we look at CLK2 since the objective now is to propagate the data to the Q2 absolute output. Since the

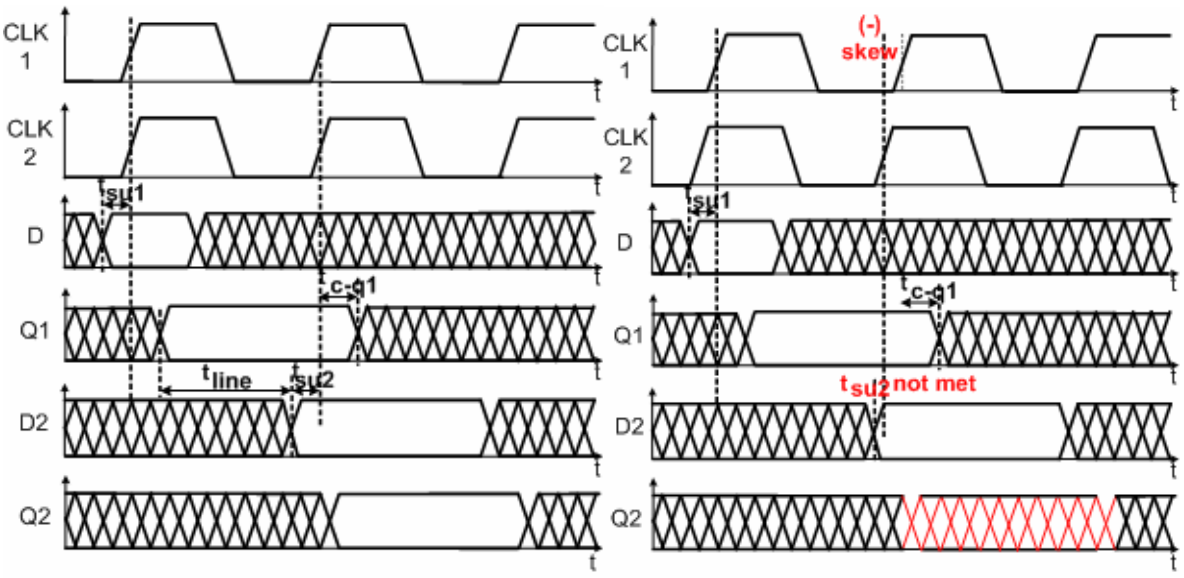


Figure 21. Timing analysis in (a) an “ideal” scenario, and (b) a "negative skew" scenario.

data D2 was available to FF2 at *least* t_{su2} before the rising edge of the clock, D2 propagates to the output Q2. In this case, perfect timing is achieved, because no sources of variation are present. However, as we look at static and thermal variations, it is evident that a timing failure is a possibility and the design becomes vulnerable.

Figure 21(b) shows the case when negative clock skew is encountered. Due to some variation, CLK1 is slower than CLK2, and CLK2 triggers first. This is no problem in the first cycle, since D propagates to Q1 with respect to the time when CLK1 triggers. As D1 propagates through the interconnect, it is obvious that CLK2 has triggered before it was expected, therefore it does not have time to meet t_{su2} and see the correct value of D2, so captures the wrong value. Since the system is made of edge-triggered devices, even though D2 is available, the CLK2 edge failed in capturing the correct value of D2. The effect propagates out to Q2, and the system has a timing failure due to negative skew. Negative skew can be represented by Equation 3.8, where δ is the symbol representing clock skew.

Since it is *added* into the inequality, it essentially means the clock period must be larger than the ideal, because it must take into account the time that CLK2 triggers before it was supposed to, to still obtain the correct value at the output in the presence of this phenomenon.

$$T \geq t_{c-q1} + t_{line} + t_{su2} + \delta \tag{3.8}$$

$$T \geq t_{c-q1} + t_{line} + t_{su2} - \delta \tag{3.9}$$

Equation 3.9 shows the opposite case, where CLK2 arrives later than CLK1. Ideally, this should be a case that helps timing, because it means the data has more time to get to the input of CLK2 before it tries to read the value. However, if the CLK2 path is very slow (or alternatively, the interconnect distance is very short), D2 may have already switched to a different value which may be the one FF2 actually captures, given the timing difference. This is known as a race condition and it is also an undesirable situation since it can lead the circuit to capture the wrong value.

3.3.1 Variation Assumptions

Assuming intra-die device and wire variations as given in Table 3, the absolute percentages of variation for the threshold voltage (V_t) and the effective channel length (L_{eff}) were obtained from the International Roadmap of Semiconductors (ITRS) [1]. Interlayer dielectric thickness (t_{ILD}) and metal height (h) were reasonable assumptions from industry. To take into account the correlation between V_t and L_{eff}, only V_t variation was modeled and assumed to include all sources of variation on which it depends.

Hspice Monte Carlo analysis was performed on the data path and clock path transistors, obtaining 1,000 random scenarios. At each of these random instances, each

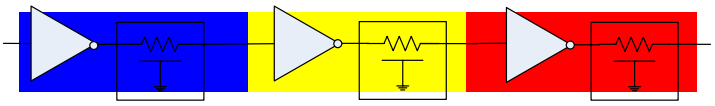


Figure 22. Increasing, discrete thermal profile on a repeated interconnect.

transistor has a different V_t, while dielectric thickness and metal height are fixed for that instance. Nominal (or cold) temperature is fixed at 70ºC. Delay for this system was defined as t_{c-q2}, i.e. the clock-to-q delay of FF2.

In this work, only increasing temperature profiles are considered, i.e. temperature goes from cold to hot along the interconnect as shown in Figure 22. A 20ºC thermal gradient is assumed across a 1mm interconnect experiencing a "hot spot". Temperature is assigned based on the number of buffer-segment pairs along the interconnect. Figure 22 illustrates the segmentation. In this case, the 20ºC would be divided into three smaller gradients, with each one being assigned to a buffer-wire pair.

Three thermal scenarios of interest are explored in this work. . Figures 23 – 25 show sketches of these scenarios: negative skew, and two race condition scenarios. In the negative skew scenario, the CLK1 path experiences a high-temperature environment, and therefore executes slowly. The CLK2 path and the interconnect are assumed to run at the nominal temperature of 70ºC. This is the timing failure scenario discussed in the timing analysis of Figure 21(b). The CLK2 rising edge triggers quicker than expected and t_{su2} is not met.

Figure 24 shows the case of a positive skew which should be beneficial in theory but is susceptible to race conditions. In other words, since the CLK2 path is slow, D2 may change before CLK2 triggers and it may capture the wrong value. Figure 25 shows the worst case negative skew condition, this time with the interconnect at a hot temperature as well as the CLK1 path, which can cause CLK2 to trigger so quickly (with respect to the other clock)

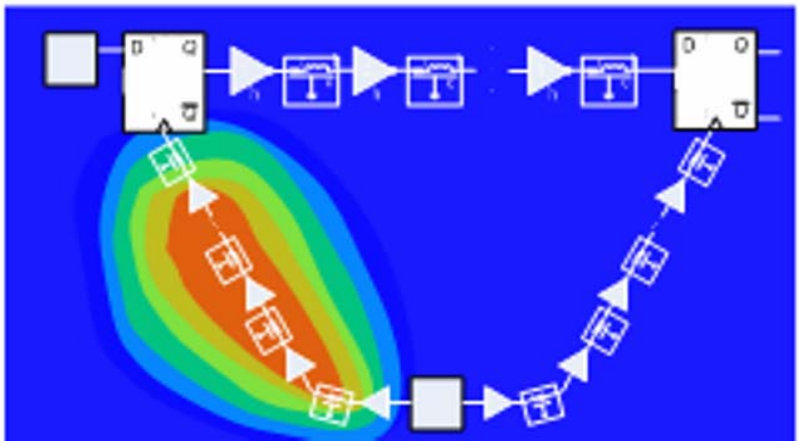


Figure 23. Negative clock skew.

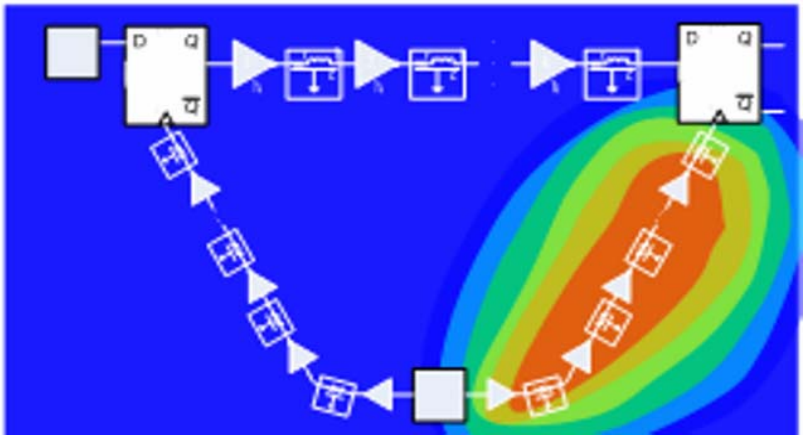


Figure 24. Race condition due to CLK2 being slow.

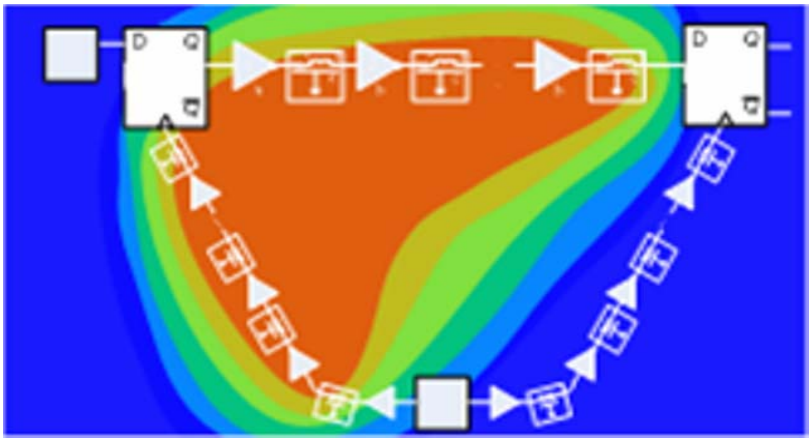
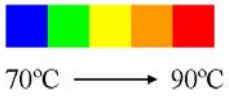


Figure 25. Worst case negative clock skew condition.



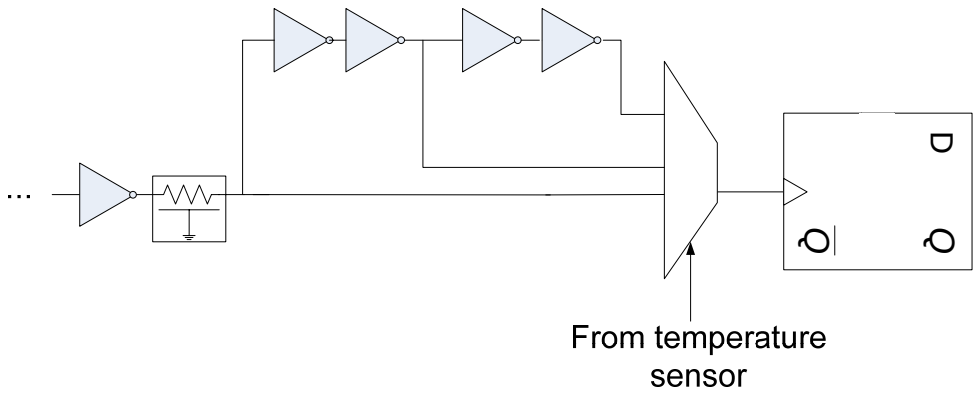


Figure 26. Skew compensation using additional delay provided by buffers.

that it will capture the old value of D2, since the data may still be traveling along the interconnect, or may not have been clocked into FF1 yet due to the slow CLK1 path.

3.3.2 Skew Compensation Method

One of the most common ways to compensate for the delay variation seen in the presence of process variations is adding extra delay to paths which need skew compensation. Tunable Delay Buffers (TDB) are a widely used method of skew compensation, and have been used by [23] in a partially off-line scheme where the degree of tuning and number of tuning elements have been pre-computed for all given scenarios. Their TDBs were implemented by hanging capacitances that are tapped into as needed to provide the necessary delay.

To compensate for the negative clock skew induced by the variations, a variable delay compensation mechanism similar to [24] is proposed in this work for critical paths. As seen in Figure 26, the idea is to have small buffers as an "optional" line that can be used when additional delay is needed for skew compensation. The amount of delay desired can be chosen by a multiplexer which is controlled by a signal coming from the nearest temperature

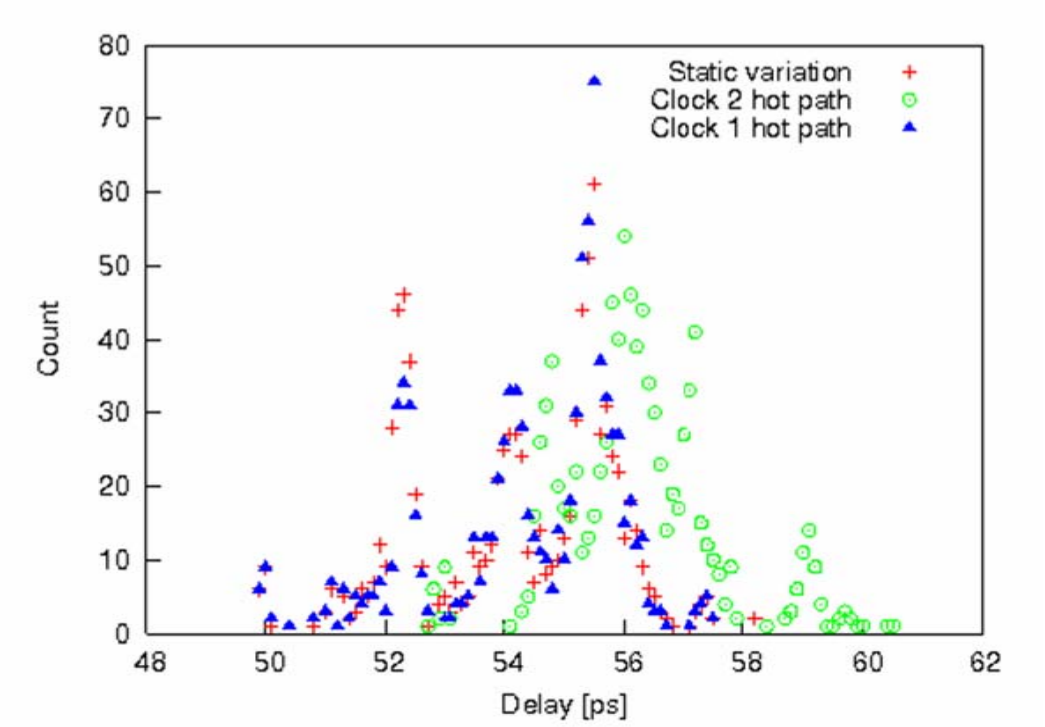


Figure 27. Delay distribution in the presence of static variation and for several thermal scenarios of interest.

sensor or the one assigned to that region of logic. With this mechanism present in critical paths, clock skew compensation in the presence of static and thermal variations is possible. This delay compensation is very similar to what was proposed for the Itanium processor in [24] for clock resynchronization.

The combined effects of static and temperature variations were modeled on the data path and clock paths and 1,000 random Monte Carlo samples were generated, to obtain a delay distribution in the presence of device and wire uncertainty. Figure 27 shows three distributions: the delay distribution of the whole system while encountering only static variations and running at a nominal temperature of 70ºC; a scenario where CLK2 experiences an increasing, high temperature gradient of 20ºC, potentially causing a race condition; the scenario where CLK1 is running at a hot temperature and causes a negative clock skew effect.

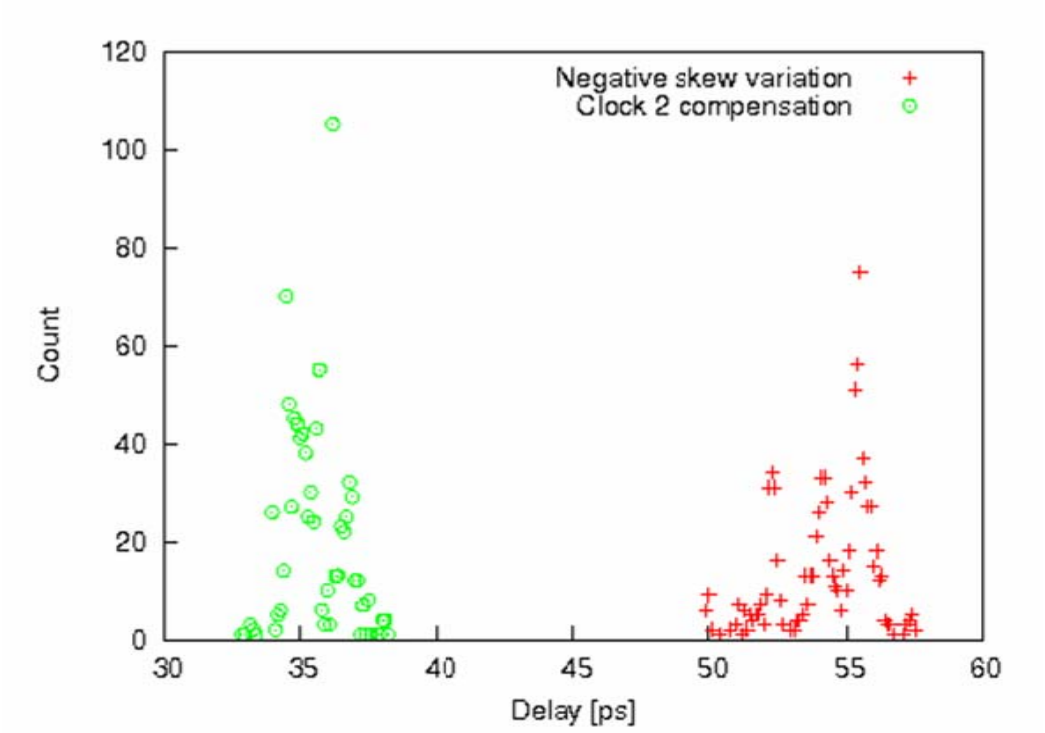


Figure 28. Shifted distribution after skew compensation through buffer delay insertion in CLK2 to avoid timing failures due to negative clock skew.

The skew compensation method introduced above was applied for the negative clock skew case where the CLK1 path is hot. Figure 28 shows the static variation delay distribution with the shifted distribution for the former race condition scenario. With just two minimum buffer delays applied for this particular case, the delay distribution is faster by about 35% with a buffer power penalty of 10% and the negative skew scenario has been averted.

CHAPTER 4

PHASE CODED INTERCONNECTS

Phase coding for interconnects is an idea proposed by [25] which allows the efficient use of the interconnect bandwidth by encoding data bits onto a single interconnect system and decoding the data at the end. This novel idea evolved from the concept of pulse modulation used in analog systems, where the pulse width represents the data being transmitted.

Having characterized interconnects under the effects of process variation for current and future technology nodes, one can further utilize the bandwidth of a well-characterized interconnect to transmit data through the line. Since phase coding requires the use of a decoder and an encoder, additional uncertainty could be introduced in the form of variations in these components of the system. This chapter will look at the possible sources of variation, given a characterized interconnect, and look at the overall effect of having this additional circuitry on a well characterized interconnect.

4.1 Phase Coding Technique

Phase coding consists of encoding data bits onto a single, repeated interconnect by using delayed versions of the signal to create a unique 2^n-bit output for each possible data input of n bits. Figure 29 shows the overall open-loop system structure (note that [25] has also proposed a closed-loop configuration but this has been left for future work, as it implies

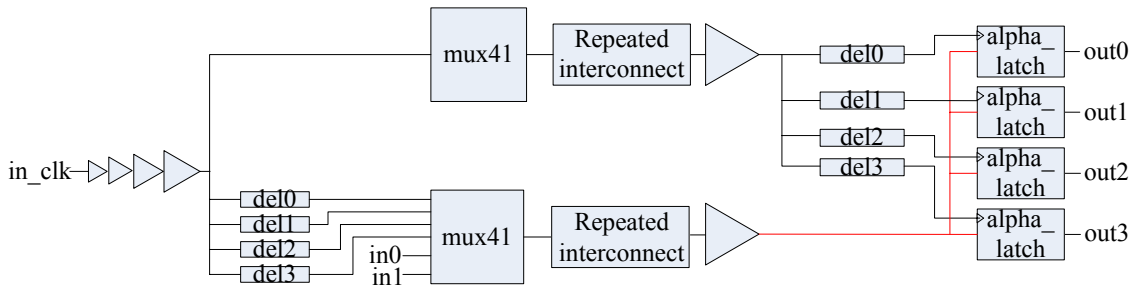


Figure 29. Phase-coded signaling for two data bits on a repeated interconnect, as proposed by [25].

a much more complex configuration). An input signal in_clk travels through two paths. The upper path simply serves as a reference signal and bypasses the multiplexer directly. The reason for having a multiplexer at the top is that we want to have the same loading and similar variation going through both wires, since this is a timing-sensitive system. Through the bottom path, in_clk is delayed by four different delay amounts: del0, del1, del2 and del3. The input data consisting of two bits, in0 and in1, serve as selecting signals for the multiplexer, and they choose which of the delayed versions of in_clk to allow through. Once the respective signals have passed through the multiplexer stage, they travel through their corresponding repeated interconnects. In the top (reference) path, the signal is now delayed by del0, del1, del2 and del3 which are *identical* delays to the ones used at the beginning of the bottom path. The output of each of these delays in the top path clocks a corresponding latch. The input data for the latch comes from the bottom path. To accurately capture the correct signal at the input of the latch, its clocking must be very accurate and delays adequately spread apart. One design constraint however, is not making the devices unnecessarily big, to avoid additional power penalties. The design of the delay elements becomes a tradeoff between power and performance.

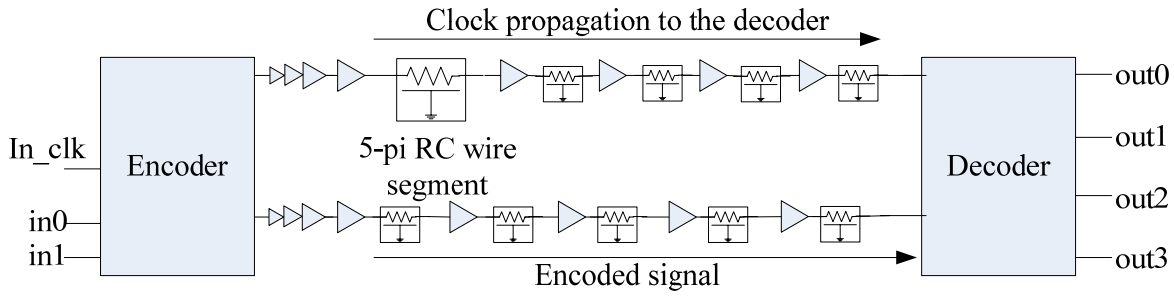


Figure 30. Reference and encoded signal travel across two different wires to be encoded into a unique 4-bit output, depending on the input data bits.

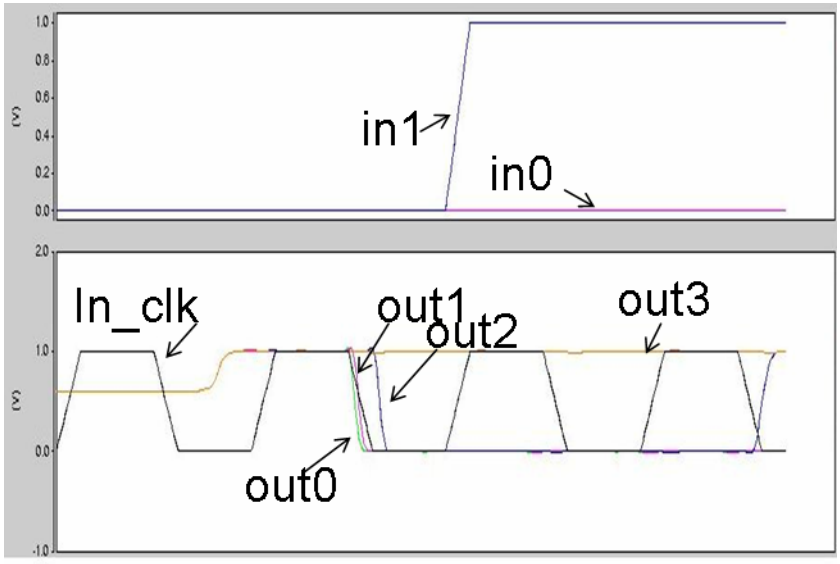


Figure 31. Hspice output waveforms showing encoding of two bits in0 and in1 into four output bits out0, out1, out2 and out3.

Figure 30 shows the whole system, abstracting away from the encoder and decoder, and focusing on the interconnect. The interconnects in both reference and data path are the same length and have the same optimal-repeater insertion. Again, it is important that both paths are identical so they represent the same loading and ideally the output latches capture the data at the correct time, as shown in Figure 31.

4.2 Variation Impacts on Phased Coded Interconnects

Phase coded interconnects are sensitive to timing variations because the latches in the decoding stage must capture the correct data coming in from the reference signal. Given an n-bit input data, the decoder circuit will choose which delayed version of the data will go through to the output, and this generates a 2^n-bit output pattern that will uniquely identify the n-bit input. In order to have correct values at the output, the timing of the latches must be accurate. As discussed in Chapter 3, this may not always happen, due to the process and thermal variations the interconnect will experience.

Since phase coded interconnects require additional circuitry for the encoder and decoder parts of the system, this may hinder the overall performance and the timing may fail. For example, if the encoder is operating in a hot environment the delay elements will provide slower signals than originally designed for at the nominal temperature case. If in this scenario, the decoder is running in a cold environment, the delay elements will have the delay that was designed for, but when decoding, the output bit sequence will be different than expected, and the input bits will be identified erroneously. Similarly since all transistors can experience different process variations, this will also compromise correct transmission of the bits. From Chapter 3, the delay variation for a 45nm repeated interconnect in the presence of process variations can span a range of ~67ps, which translates into a variation of 32%. This variation will slightly increase when taking into account the variation in the encoder and decoder parts of the phase coding system. If 32% variation were encountered overall, the possibility of getting an incorrect bit sequence at the output is very high. Such uncertainties can severely compromise performance and correctness of the design and must be addressed,

taking into account that worst-case design is not a feasible option due to power and area overheads, among others. Passive solutions, such as repeater number and sizing, can be explored in the presence of thermal or static uncertainties, to ensure accurate performance. In addition, active solutions, such as the closed-loop approach proposed by [25], can provide a means of correcting the effects of variation dynamically by compensating at run time. Passive and active solutions will be explored as future work.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

Variability is a growing issue as interconnects and devices scale further in the deep submicron regime. Static variations due to manufacturing inaccuracies or limitations can cause a very significant impact on performance, but are not variables that change at run time. Dynamic variations such as temperature, however, are much harder to predict due to their uncertain nature. With increasing uncertainty in smaller devices, design for compensation becomes a viable option to pursue.

This thesis gave an insight into the effects of both sources of variability in on-chip interconnect performance and explored a wide range of variation cases. Based on the observed trends, smaller design margins can be set, thus avoiding the disadvantages of conventional worst-case design.

5.2 Contributions

The impact of effective channel length variation on interconnect performance was reduced by proposing a supply voltage assignment technique to reduce the delay distribution. A 90% reduction of delay distribution with negligible power overhead was achieved for a 1ps tolerance in distribution width, however, it required a fine granularity of supply voltages. For a less constrained scenario of 1ps tolerance in distribution width,

a 80% reduction in the delay distribution was achieved with a power overhead of less than 1%.

Temperature and manufacturing variations were explored on a 45nm data path clocked by 2 H-tree leaves. Several thermal scenarios were shown, and their impact on the delay distribution while in the presence of static variations was observed. A clock skew compensation technique was suggested, for preventing timing failures in the presence of variability. When applied to a negative clock skew scenario, the delay distribution was 35% faster with a 10% power overhead and timing failures were averted.

To increase the bandwidth usage of the interconnect, once it has been optimized for variation-tolerance, phase coded interconnects were proposed. An introduction of variation concerns on the phase coding system was given, and timing vulnerabilities were discussed.

5.3 Future Work

Power supply variations are a significant source of timing and delay uncertainty since a large number of buffers will be subjected to the variation at the clock grid level. Therefore, it is important to consider power supply variation in conjunction with the other sources considered in this thesis (process and temperature) as an extension to this work.

Interconnect timing in the presence of supply voltage variation will be incorporated in the data path study and more methods for skew compensation in the presence of process-voltage-temperature (PVT) variations will be explored for current and future technologies.

Phase coded interconnect performance in the presence of process, voltage and temperature (PVT) variations for 45nm and 32nm technologies will be addressed in future work, and techniques for performance compensation will be explored. Since the effectiveness of phase coding is highly dependent on timing accuracy, these issues must be addressed. Passive and active methods of compensation to reduce the effects of variation on phase coding performance will be applied.

BIBLIOGRAPHY

[1] International Technology Roadmap for Semiconductors, *http://public.itrs.net*.

[2] "45nm Hi-k Next Generation Intel® Core™ Microarchitecture", *http://www.intel.com/technology/architecture-silicon/45nm-core2/description.htm.*

[3] V. Adler, E.G. Friedman, "Repeater Design to Reduce Delay and Power in Resistive Interconnect", *IEEE Transactions on Circuits and Systems II: Analog and Digital.*

[4] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs", *IEEE Transactions on Electron Devices, Vol. 49, Issue 11, pp. 2001-2007, 2002.*

[5] Ja Chun Ku and Y. Ismail, "Thermal-Aware Methodology for Repeater Insertion in Low-Power VLSI Circuits", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 15, Issue 8, pp. 963-970, 2007.*

[6] J. Xiong, K. Tam, L. He, "Buffer Insertion Considering Process Variation", *Design, Automation and Test in Europe Conference and Exhibition, 2005, pp. 970-975.*

[7] A. Chandrakasan, W. Bowhill, F. Fox, "Design of High-Performance Microprocessor Circuits", *IEEE Press, 1999, pp.98-114.*

[8] K. Bernstein, et al. "High-performance CMOS variability in the 65-nm regime and beyond", *IBM Journal of Research and Development, Vol. 50, Number 4/5, 2006*.

[9] M. Pedram, S. Nazarian, "Thermal Modeling, Analysis, and Management in VLSI Circuits: Principles and Methods", *Proceedings of IEEE, special issue on Thermal Analysis of ULSI, Vol. 94, No. 8, pp. 1487-1501, 2006*.

[10] S. Im, N. Srivastava, K. Banerjee, K. Goodson, "Scaling Analysis of Multilevel Interconnect Temperatures for High-Performance ICs", *IEEE Transactions on Electron Devices, Vol. 52, No. 12, pp. 2710-2719, 2005*.

[11] A.H. Ajami, K. Banerjee, M. Pedram, "Modeling and Analysis of Nonuniform Substrate Temperature Effects on Global ULSI Interconnects", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 24, No. 6, pp. 849-861, 2005*.

[12] Predictive Technology Model., *http://www.eas.asu.edu/ ptm*

[13] Y. Cao et al., "Design Sensitivities to Variability: Extrapolations and Assessments in Nanometer VLSI", *IEEE ASIC/SOC Conference, 2002*.

[14] Star-Hspice Manual, Release 2006., *Synopsys Corporation , Dec. 2006*.

[15] W. Jeong, B.C. Paul, K. Roy, "Adaptive Supply Voltage Technique for Low Swing Interconnects", *Proceedings of the Asia and South Pacific Design Automation Conference: Electronic Design and Solution Fair, 2004, pp. 284-287.*

[16] R. Viswanath et al., "Thermal Performance Challenges from Silicon to Systems", *Intel Technology Journal, Vol. 4, Issue 3, August 2000*.

[17] J.D. Warnock et al., "The circuit and physical design of the POWER4 microprocessor", *IBM Journal of Research and Development, Vol. 46, No. 1, pp. 27-51, January 2002*.

[18] A. Chakraborty et al., "Thermal Resilient Bounded-Skew Clock Tree Optimization Methodology", *Design, Automation and Test in Europe, 2006*.

[19] M. Cho, S. Ahmed, D.Z. Pan, "TACO: Temperature Aware Clock-tree Optimization", *IEEE/ACM International Conference on Computer-Aided Design, 2005*.

[20] K. Sundaresan and N. Mahapatra, "An Analysis of Timing Violations Due to Spatially Distributed Thermal Effects in Global Wires", *IEEE/ACM Design Automation Conference, 2007*.

[21] D. Chen et al., "Interconnect Thermal Modeling for Accurate Simulation of Circuit Timing and Reliability", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 19, No. 2, pp. 197-205, February 2000*.

[22] K. Duraisami et al., "Design Exploration of a Thermal Management Unit for Dynamic Control of Temperature-Induced Clock Skew", *IEEE International Symposium on Circuits and Systems, 2007*.

[23] A. Chakraborty et al., "Dynamic Thermal Clock Skew Compensation using Tunable Delay Buffers", *International Symposium on Low Power Electronics and Design, 2006*.

[24] S. Naffziger et al., "The Implementation of a 2-Core, Multi-Threaded Itanium Family Processor", *IEEE Journal of Solid-State Circuits, Vol. 41, No. 1, pp. 197-209, January 2006*.

[25] A. Maheshwari, “Circuit and Signaling Methods for On-Chip Interconnects”, *PhD Thesis*, University of Massachusetts Amherst, 2004.

[26] G.M. Link, N. Vijaykrishnan, “Thermal Trends in Emerging Technologies”, *International Symposium on Quality Electronic Design, 2006*.

[27] J. Cong, “Challenges and opportunities for design innovations in nanometer technologies”, *SRC Design Sciences Concept Paper, 1997*.

[28] J.M. Rabaey, A. Chandrakasan, B. Nikolić, “Digital Integrated Circuits: A Design Perspective”, *Prentice Hall, 2003, pp.120-121.*

[29] I.M. Filanovsky, A. Allam, “Mutual Compensation of Mobility and Threshold Voltage Temperature Effects with Applications in CMOS Circuits”, *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, Vol. 48, No. 7, pp. 876-884, 2001*.

[30] S..A. Bota et al., “Impact of Thermal Gradients on Clock Skew and Testing”, *IEEE Design & Test of Computers, Vol. 23, Issue 5, pp. 414-424, September 2006*.

[31] S. Sauter et al., “Effect of Parameter Variations at Chip and Wafer Level on Clock Skews”, *IEEE Transactions on Semiconductor Manufacturing, Vol. 13, No. 4, November 2000*.

[32] S. Borkar et al., “Parameter Variations and Impact on Circuits and Microarchitecture”, *Design Automation Conference, pp. 338-342, 2003*.

[33] R. Rao et al., “Statistical Estimation of Leakage Current Considering Inter- and Intra-Die Process Variation”, *International Symposium on Low Power Electronics and Design, 2003*.

[34] S. Boyd et al., “A Heuristic Method for Statistical Digital Circuit Sizing”, *SPIE International Symposium on Microlithography, Vol. 6156, pp. 58-66, February 2006*.

[35] M. Guthaus et al., “Optimization Objectives and Models of Variation for Statistical Gate Sizing”, *ACM Great Lakes Symposium on VLSI, pp. 313-316, 2005*.

[36] R. Chen and H. Zhou, “Fast Min-Cost Buffer Insertion under Process Variations”, *IEEE/ACM Design Automation Conference, 2007*.