Process and Environmental Variation Impacts on ASIC Timing

Paul S. Zuchowski, Peter A. Habitz, Jerry D. Hayes, Jeffery H. Oppold

IBM Microelectronics Division
Essex Junction, Vermont 05452, USA

Introduction

With each semiconductor process node, the impacts on performance of environmental and semiconductor process variations become a larger portion of the cycle time of the product. Simple guard-banding for these effects leads to increased product development times and uncompetitive products. In addition, traditional static timing methodologies are unable to cope with the large number of permutations of process, voltage, and temperature corners created by these independent sources of variation. In this paper we will discuss the sources of variation; by introducing the concepts of systematic inter-die variation, systematic intra-die variation, and intra-die random variation. We will show that by treating these forms of variations differently, we can achieve design closure with less guard-banding than traditional methods.

ASIC providers are typically responsible for the performance and yield of the devices they deliver. It is therefore common in the ASIC industry to require timing closure, as measured by a static timing analysis tool, at fast process and slow process timing corners. These corners are supposed to represent the maximum variation that is possible between any two die due to normal manufacturing tolerances. The definition of these fast and slow corners is usually done by moving all of the relevant process parameters (eg. channel length, threshold voltage, etc.) to some statistical limit and developing timing models with these process assumptions. It is also now common for ASIC providers to require timing sign-off assuming some amount of on-chip variation. This additional conservatism is added to account for the intra-die variations which can result in missed timings due to differential process variation (and therefore delays) on the clock and data paths. Environmental condition variations, such as end user voltage and temperature, are also accounted for by running additional static timing corners. In the recent past, these methods were sufficient to guarantee timing, and therefore yield, across the range of the normal manufacturing process window. With continued scaling of CMOS technology however, the numbers of relevant sources of variation and their magnitude have increased. In an attempt to account for this, additional static timing corners are being added to ASIC design flows to account for sources of variation that were previously ignored, such as mismatch between PFET on-current and NFET on-current due to threshold voltage variation.

As additional sources of variation become important, either the total guard-band applied during static timing is increased, or the risk of impacting yield is increased. This comes about due to the different delay sensitivities of each path on a design and the inability of the currently available design automation tools to handle the unique sensitivities of each path on the chip without running 2^n timing corners, where n is the number of independent variables of interest. Some paths are predominately sensitive to metal delay while others are predominately sensitive to silicon (device) delay. Assuming that these paths will track the same from manufacturing lot to lot can lead to silicon that fails its timing requirements. Clearly a better method for dealing with variations is required.

Background

Continued scaling of CMOS technologies is increasing the magnitude of intra-die variation [1] [2]. The magnitude of intra-die channel length variations is estimated to grow from 35% of the total variation for a 0.13um technology to almost 60% in a 0.07um technology [3]. Intra-die variation on wire width, height, and thickness is also expected to grow from 25% to about 35% for these same technologies. This variability can be expressed in terms of both a random or un-correlated variation and a systematic or correlated variation. Furthermore, the systematic variation can further be sub-divided into inter-die systematic variation and intra-die systematic variation.

Inter-die systematic variation is due to normal manufacturing tolerances that affect the mean value of a parameter from lot to lot, wafer to wafer and die to die. Inter-die variations between parameters within a single die can lead to either performance degradation or failing hardware when not properly accounted for. Examples of inter-die systematic variation include channel length variation due to the length

0-7803-8702-3/04/$20.00 ©2004 IEEE.

of exposure, and variations between individual metal layers used for routing. Each metal layer represents an independent processing step in manufacturing, thereby insuring a high degree of miscorrelation between one layer and the next. Traditional timing assumes correlation between metal layers by timing at a fast corner where all metal layers are represented by minimum capacitance and maximum resistance, and a slow corner where all metal layers are represented by maximum capacitance and minimum resistance. When two paths with dissimilar metal layer content are racing towards a latch where setup and hold requirements must be met, the variability in delay due to differing metal content can cause timing exposure not seen with traditional timing methods. Another important source of systematic inter-die process variation is threshold voltage. At 90 nm it is important to consider the relative difference between PFET and NFET threshold voltages, for example, a fast NFET and a slow PFET. It is also possible to have a fast high-vt NFET and a slow low-vt NFET. Both of these cases represent timing exposure when not properly accounted for.

Intra-die systematic variation comes about because of layout specific variations. These variations can be the result of semiconductor process methods or environmental differences that are seen across the design based on layout. Examples of process induced systematic intra-die variation include 1) optical proximity effects that causes polysilicon features sizes such as Leff to vary as a function of local layout, 2) local wire densities that influence the inter-layer dielectric thickness achieved during chemical-mechanical polishing (CMP), and 3) spatial variation of Leff due to lens aberration across the die [4] [5] [6]. Many techniques have been used in manufacturing to reduce systematic intra-die variation. One of the most widely adopted is the use of optical proximity correction (OPC) for modifying features sizes across the mask in order to compensate for local proximity effects [7]. It has also been suggested that OPC could be used to compensate for spatial variation due to lens aberrations [5]. Another common technique to reduce systematic intra-die variation is phase-shift masking (PSM) that improves depth-of-field and resolution in lithography [8]. Systematic variation due to CMP has been addressed by improving metal uniformity across the die. While these techniques have been successful at reducing systematic intra-die variation, the ability to continue to improve manufacturing systematic intra-die tolerances is limited, particularly as feature sizes continue to shrink [2].

In addition to process induced intra-die variation, there are also environmentally induced sources of intra-die variation. Examples of environmentally induced variation include voltage and temperature values that vary across the die. These parameters are influenced by power grid design, the placement of circuits, and vector set.

Random variation can be caused by any number of things including lithography, etching, polishing and doping effects. An example of random intra-die variation is the variation in device threshold voltage due to quantization effects of doping atoms within increasingly smaller silicon structures [9][10]. These quantization effects represent a continued increase in the fundamental randomness of silicon behavior as device dimensions continue to decrease.

For many technology generations, the intra-die variation could be safely ignored when compared to the dominant inter-die variation. In older technologies, chip clock frequencies (FMAX) varied around an average value defined by the mean of the inter-die variation while the inter-die variance directly related to the variance of FMAX. As the mean of the inter-die variation changed, either by line tailoring, process learning, or technology migration, a corresponding shift in the average FMAX would result. As intra-die variation becomes more significant, the average FMAX will begin to decrease even when the mean of the inter-chip variation remains constant. Intra-die variation will make some paths faster and other paths slower as a function of their gate and wire composition, local layout attributes, and spatial location. Since the maximum clock frequency is defined by the maximum path delay, intra-chip variation results in an overall degradation of performance [11].

In addition to process variations, non-process related parameters such voltage islands with timing interactions between islands, synchronous phase lock loop (PLL) domains, and FET degradation due to negative bias temperature instability (NBTI) and hot-electrons that result in a difference in performance across the life time of the product can cause timing exposure when not properly accounted for in static timing. This exposure is expected to increase as the numbers of independent parameters, both process and non-process related, increase in technology offerings. There are two ways to reduce this exposure. The first is to perform additional timing runs to cover the entire parameter space, quickly becoming prohibitive with traditional timing due to the excessive number of runs required. A second approach is to add enough uncertainty between racing paths to remove this exposure. For the design team this uncertainty can lead to unnecessary work on non-critical paths resulting in lower design performance and an increase in time-to-market.

The amount of degradation seen during chip timing will depend on the magnitude of each variation type, i.e. systematic and random, the characterization and extraction techniques used in the timing models to represent these variations, and the algorithms used by the timing tool to apply these variations during chip timing. Historically the technique used by ASIC vendors to account for intra-die

variability has been to apply an uncertainty to path delay. The amount of uncertainty applied is proportional to the magnitude of the intra-die variation and is typically a fixed percentage of the overall path delay. Early paths will get faster due to this uncertainty while late paths will get slower. For rapidly increasing intra-die variations, using an uncertainty approach that attempts to bound intra-die variation on chip timing will become excessively conservative, significantly degrading the predicted circuit performance while forcing the design team to work on non-critical paths. On the other hand, not providing enough uncertainty to cover the intra-die variations will increase exposure to non-working hardware. Given that intra-die variability is on the rise, three areas of modeling improvement will be needed in order for ASIC vendors to remain competitive. The first area is a better understanding of the sources of variability. It is not sufficient to know only the magnitude of each source of variation, one also needs to know if the variation is 1) locally systematic that can be handled through extraction techniques, 2) spatially systematic that can be handled using spatial proximity techniques, 3) random that can treated statistically across the die and 4) unknown that would require continued use of an uncertainty approach. It may be that some sources of variations will have elements of all of these variations. As timing methodologies continue to evolve towards true statistical timing, an understanding of the distributions of each source of variation will also be required. The second area of improved modeling requirements is the ability of the timing models used during chip level timing to capture this variability in each of its forms. The standard timing models used in industry today do not have this capability. The third area of required improvement is in the algorithms used by the timing tool during chip level timing in order to exploit the new variability information in the timing models.

As we describe below, it is necessary to make the distinction between inter-die and intra-die systematic variation and random variation in order to properly account for the effects of these variations. A timing methodology that addresses the impact of both systematic and random variations on design performance will be required as we move deeper into the sub-micron range.

Modeling Delay Variation in a Digital Library

Figure 1 shows some important sources of variation in the 90 nm CMOS technology node. For each component of variation we have also listed our classification of that source of variation as inter-die systematic, intra-die systematic, random or some combination of the three.

Hardware measurements were used to characterize the magnitude of systematic and random delay variation due to various sources. Figure 2 shows hardware data taken from a 90 nm test chip. One of the experiments on the test chip was specifically designed to look at variation. It consisted of four copies of a ring oscillator, one copy placed at each corner of the chip. The layout for each instance was identical. The oscillator design is almost insensitive to metal variation (it has almost no metal) and when tested, no other circuits on the die are active so as to minimize other systematic effects (voltage and temperature differences). Each series in the graph represents data from one of the four oscillators (upper right, upper left, lower right, lower left). The frequency of oscillation for each oscillator circuit is plotted as a function of the mean frequency of all four oscillators on the same die. The hardware represents many die over multiple wafers and multiple lithography exposures. A clear intra-die systematic variation can be seen based on location of the oscillator within the die as shown on the y-axis. The inter-die systematic variation is represented by the x-axis. The magnitude of this delay variation is approximately ±7.5% of the mean oscillator delay.

Component	Form of Variation
Channel length	Inter-die systematic, intra-die systematic, intra-die random
Mean threshold voltage difference between device types	Inter-die systematic
Threshold voltage	Inter-die systematic, intra-die random
Mean metal R and C differences between metal levels	Inter-die systematic
Voltage and Temperature	Intra-die systematic
NBTI, hot-e	Intra-die systematic

Figure 1. Some 90 nm Components of Variation

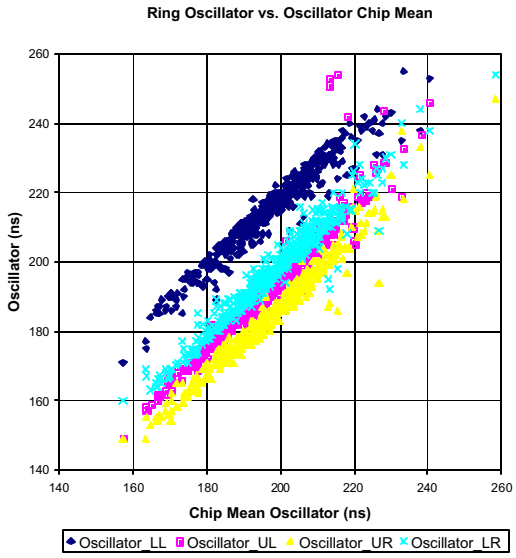

Figure 2. Systematic Oscillator Variation

The "width" of each series represents the random variation. Figure 3 is a plot of the same data as in Figure 2, however it has been normalized to show the magnitude of the random delay variation. First, the systematic component of delay variation was removed by subtracting the mean frequency of the fastest oscillator series from the mean frequency of each of the other series. Then, the difference in oscillator frequency between each oscillator and every other oscillator on an individual die was plotted against the mean frequency of all oscillators for that same die. The 3-sigma magnitude of the random delay variation shown in Figure 3 is approximately ±5.5%.

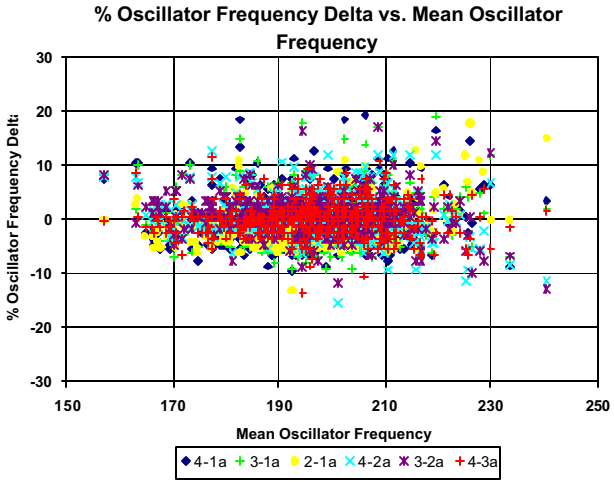

Figure 3. Random Oscillator Variation (normalized)

Special Spice models were created which allowed us to uniquely characterize the sensitivity of each circuit for inter-die systematic variation, intra-die systematic variation and random variation. Figure 4 is a schematic representation of our circuit characterization philosophy. First, a set of chip mean circuit delays are calculated for each timing arc. The set consists of multiple pairs of delays (slow and fast), each pair representing a possible process corner (eg. NFET slow, PFET slow; NFET fast, PFET slow). Full statistical chip timing is not being implemented at the 90nm node, therefore only the delay endpoints are of interest and not the shape or the density of the delay distributions. This set of mean delays captures the variation that is possible due to inter-die systematic variation. These delays are calculated using simulation, by setting the appropriate process parameters in the spice model to the sigma values of choice. In practice, the individual process end points are determined using statistical techniques. During chip level timing, inter-die systematic variation is accounted for using separate "corners" where the process corner is the same for every circuit on the chip.

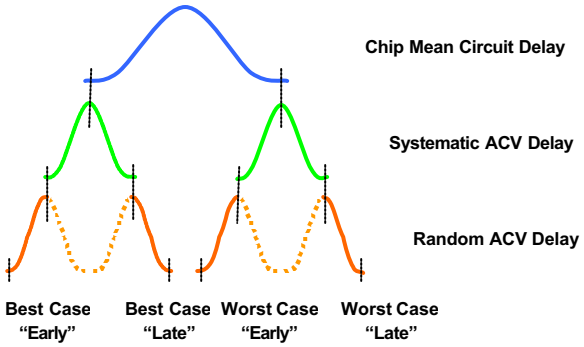

Figure 4. Modeling of Variations

Intra-die systematic variation, or across-chip variation (ACV), sensitivities are obtained next. This is done by applying different process parameter assumptions to the spice model and re-simulating each arc. The systematic components of variation need to be broken into distinct terms for those effects that can be modeled (eg. Voltage) and those that currently cannot be modeled (eg. Local poly density which depends on circuit placement at the chip level). These distinct "buckets" will be treated independently during chip level timing. In general, systematic variations that are dependent on a variable that can be modeled (such as voltage, NBTI, metal level content) can be optimized for. Those systematic variations that are dependent on unknown variables or variables that cannot be efficiently modeled at the chip level must be accounted for as a split between the "early" delay and the "late" delay of the circuit which manifests itself as uncertainty in the delay of the circuit as shown in Figure 4. At the chip level, the early delay for each circuit and wire are summed to calculate the early arrival time and the late delays for each circuit and wire are summed to calculate the late arrival time.

Capturing the random sensitivity vector involves approximating a Monte Carlo simulation for each arc. This is computationally intensive, however, truly random variation can be root sum squared (RSS'd) such that the penalty for this type of variation is smaller when a large number of variables or circuits are involved (the Central Limit Theorem). Therefore, the random vector will tend to be smaller for large collections of transistors and larger for a small collection of transistors. If random variation were treated as systematic, it would need to be linearly added to the early and late timings. By characterizing this component separately, we are able to RSS the random variation at the chip level to reduce pessimism in paths with large numbers of components. The benefit of this approach is shown in Figure 5. A 1000 case Monte Carlo simulation was performed on both a single stage inverter and the same circuit, cascaded 31 times. The delay values for the single stage inverter were multiplied by 31 to normalize the figure. The figure clearly shows the benefit of treating random variation differently than systematic variation.

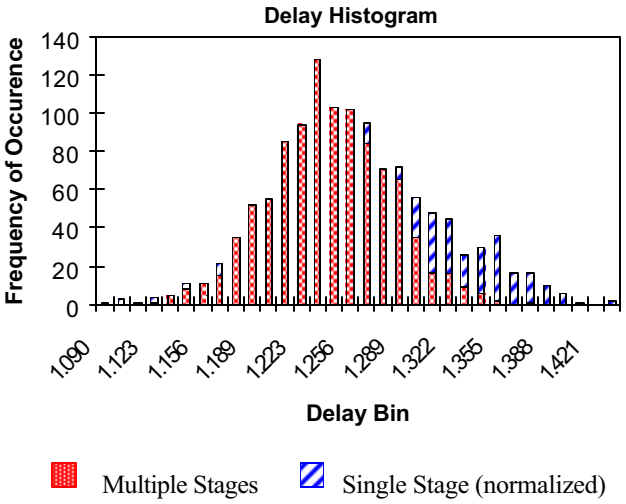

Figure 5. Random Variation for Single Stage vs. Multi-Stage

The three types of sensitivities, inter-die systematic, intra-die systematic and random, are captured for each arc for each circuit and are compiled into a DCL timing model [12]. These DCL models are used by the newly developed variation-aware static timing analysis tool which will be discussed in more detail below.

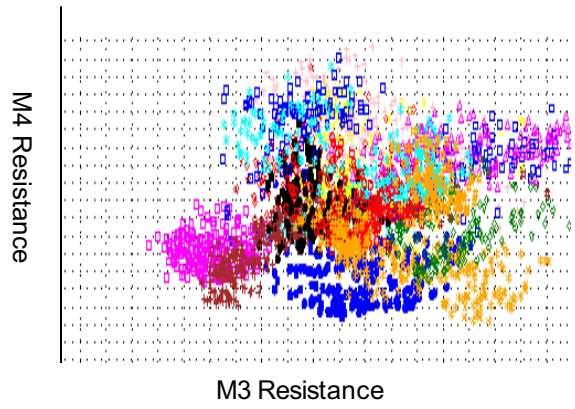

Figure 6. Normalized Metal Resistance Correlation

Modeling Inter-die Variation at the Chip Level

Figure 6 illustrates metal resistance relationships between two geometrically identical metal layers. Each point on the graph represents one die measurement comparing the resistance of metal layer 3 (M3) on the horizontal axis to the resistance of metal layer 4 (M4) on the vertical axis. Each color grouping of die measurements represents hardware from a single lot. Systematic shifts between M3 and M4 can be seen from lot to lot. The figure highlights a fundamental inter-die variation that needs to be accounted for during chip level timing. We will use this example to describe how inter-die systematic variation can be accounted for at the chip level. The concepts below can be expanded to cover other sources of systematic variation.

Two techniques can be used to insure adequate timing margins to systematic variations. The first technique is applying a parameter skew between any two racing paths in the design. We call the faster of the paths the early path and the slower path the late path. The skew would need to cover the complete range of systematic variations that are possible (or some percentage of this range). An example for metal would be to use minimum capacitance and maximum resistance (resistance and capacitance track inversely) for the early path and use maximum capacitance and minimum resistance for the late path. An advantage of this guard-banding approach is that it requires no additional timing runs beyond what is done in traditional timing since it assumes a worst case metal scenario for timing analysis. The disadvantage of this technique is that it increases timing conservatism that may unnecessarily reduce design performance and increase design turn around time (TAT). The resulting timing is also not "physical". M3 cannot simultaneously be at minimum capacitance for early paths and maximum capacitance on late paths. This is not physically realizable. This guard-banding technique must be used for systematic effects that cannot be modeled efficiently at the chip level.

A second technique for modeling systematic variation is to add parameter awareness to the timing methodology. This technique is preferred for systematic variation that is a function of a variable that can be modeled during chip level timing. For metal, a straight forward approach would be to perform multiple static timing runs where each run used a parasitic extraction with a unique combination of fast and slow metal for each wiring level on the chip. This would result in 2^N possible permutations of extraction and static timing runs, where N is the number of metal levels. The advantage of running all 2^N metal permutation corners is that each corner would be physical, i.e. both the early and late paths would see the same inter-die process space, thereby removing the conservatism associated with the guard-banding technique. The disadvantage of running 2^N timing runs, where N can be six or more in today's processes should be obvious to anyone familiar with analyzing timing on ASIC designs. A new timing approach has recently been introduced that can cover this set of permutations using a small number of timing runs [13] [14]. This method basically searches all of the critical timing tests to find the most pessimistic parameter combination for each test. A salient feature of this new timing approach is the method used to represent metal parasitics. Instead of multiple parasitic extractions where each extraction represents a metal corner permutation, only one extraction is used. Each parasitic element is referenced to a metal level and its value is

dynamically assigned during run time as a function of metal corner and temperature settings. The timing methodology then exploits the independence between metal levels by performing a metal layer sensitivity analysis on each critical test of the design. Leveraging incremental timing capabilities, each test is then separately checked at its own unique bounding process corner determined from the sensitivity analysis. Using metal layer sensitivities to guide the timing tool converts 2^N timing runs required for analysis to N+1 runs. In addition to reducing the number of required timing runs to N+1, these runs are performed seamlessly within the inner loop of the timing tool, thereby greatly increasing run time efficiency. The net effect is complete inter-die metal variation coverage at a small fraction of the run time associated with traditional timing methodologies achieving the same corner coverage.

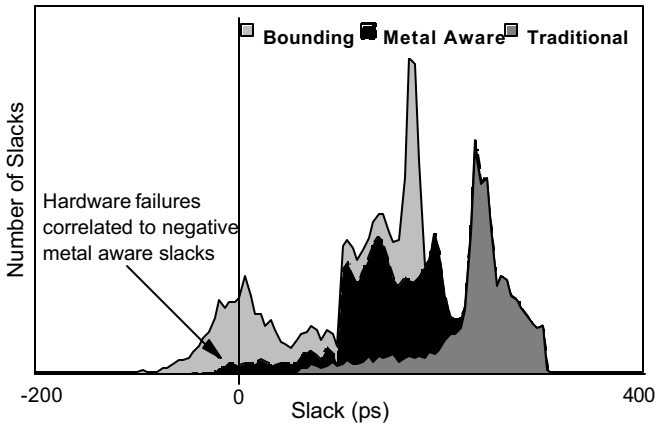

Figure 7. Slack Histograms

The impact on chip level timing due to metal layer variations can be illustrated using actual chip slack histogram reports as shown in Figure 7 where the horizontal axis denotes quantized slack ranges and the vertical axis records the number of slacks per slack range. Slacks in this chart are a measure of timing margin between clock and data arrival times at a latch, where positive slacks indicates adequate timing margin, and negative slacks indicate potential timing failures. In Figure 7 there are three slack histograms. The first histogram, denoting all positive slacks, was obtained using traditional timing methodologies that provides no awareness or guard-banding to metal layer variations. A second histogram was obtained from a metal awareness timing methodology, and a third histogram was obtained using a guard-banded (bounded) timing methodology that utilizes non-physically skewed metal parasitics between clock and data paths just enough to cover the inter-die metal variations. The histogram obtained from the traditional timing methodology seems to indicate that adequate timing margin was obtained on all latches in the design and the hardware for this part should be free from timing problems. In fact, hardware measurements on this design show multiple latches that fail hold time requirements when metal layer RC parameters on one layer are skewed from metal layer RC parameters on another layer. The parameters were still within the allowable manufacturing tolerances. The source of these failures was found to be a significant difference in the RC of the metal layers used to route the clock and data paths. The magnitude of the delay variation was not accounted for in traditional timing methodologies. The slack histogram that matches closely with the hardware was obtained using a metal variation aware timing methodology [13] [14]. Note the significant shift towards negative slacks this histogram has in comparison to the original slack histogram obtained from traditional timing. This shift represents potential timing exposure when using traditional timing methodologies as confirmed when correlating the failing latches in hardware back to negative slacks in the metal variation aware slack histogram. Guard-banding techniques that non-physically skew metal parasitics between racing paths can also remove this timing exposure, however the impact on slack distributions can be extreme as seen by the shift in the guard-banded slack histogram as compared with the metal aware slack histogram in Figure 7. This shift represents unnecessary timing conservatism that can greatly increase turn-around-time for the design and potentially lead to significant chip performance degradation. A metal variation aware timing methodology offers protection against hardware failures while eliminating unnecessary conservatism.

Conclusions

We have briefly reviewed the concepts of inter-die systematic variation, intra-die systematic, and intra-die random variation. We have shown techniques for modeling circuit delay as a function of these sources of variation using the DCL modeling language. We have shown that each form of variation requires unique modeling techniques and that by employing these techniques in both the library characterization and chip static timing methodology, designs that are tolerant of variation can be created with less guard-banding than with traditional methods. To achieve this benefit requires the modeling of systematic sources of variation as a function of their dependent variables in both the timing models and in the static timing environment. It also requires modeling the random component of variation separately from the systematic component to allow for RSS'ing during static timing analysis. Lastly, we have introduced a method of using early and late delays to model the impacts of variation on delay.

Acknowledgements

The authors would like to acknowledge the work of Eric Foreman and Toshi Saitoh for helping with hardware data collection and full-chip timing analysis. Without their help this work would not have been possible.

References

[1] Boning, D., and Nassif, S., "Models of Process Variations in Device and Interconnect," *Design of High-Performance Microprocessor Circuits*, A. Chandrakasam (ed.), 2000.

[2] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, 2001.

[3] Nassif, S., "Within-chip variability analysis," *IEDM Technical Digest*, p283, 1998.

[4] Chang, E., et al., "Using a Statistical Metrology Framework to Identify Systematic and Random Sources of Die-and Wafer-level ILD Thickness Variation in CMP Processes," *Proc. of IEDM*, 1995.

[5] Orshansky, M.,et al., "Characterization of spatial CD variability, spatial mask-level correction, and improvement of circuit performance," *Proceedings of the International Society for Optical Engineering*, vol.4000, pt. 1-2, 2000, 602-611.

[6] Stine, B., et al., "Inter- and intra die polysilicon critical dimension variation," *Microelectronic Manufacturing Yield, Reliability, and Failure Analysis II*, SPIE 1996, Oct. 1996, Austin TX.

[7] Burggraaf, P., "Optical lithography to 2000 and beyond", *Solid State Technology*, Feb. 1999.

[8] Liu, H., et al., "The application of alternating phase-shifting masks to 140nm gate patterning: line width control improvements and design optimization," *Proc. of SPIE 17th annual BACUS Symposium on Photomask Technologies*, volume 3236 of SPIE, 1998.

[9] Burnett, D., et al., "Implications of Fundamental Threshold Voltage Variations for High Density SRAM and Logic Circuits," *Symp. VLSI Tech*., pp.15-16, June 1994.

[10] Takeuchi, K., et al., "Channel Engineering for the Reduction of Random-Dopant-Placement-Induced Threshold Voltage Fluctuations." *IEDM Tech. Dig.,* Dec. 1997.

[11] Orshansky, M, "Increasing Circuit Performance through Statistical Design Techniques," *Closing the Gap Between ASIC & Custom*, Kluwer Academic Publishers, 2003.

[12] 1481-1999 IEEE Standard for Integrated Circuit (IC) Delay and Power Calculation System, copyright 1999 by IEEE.

[13] Visweswariah C., et al., "First-order incremental block-based statistical timing analysis," Design Automation Conference, pp 331-336, June 2004.

[14] Jess J. A. G., et al., "Statistical timing for parametric yield prediction fo digital integrated circuits," Design Automation Conference, pp 932-937, June 2003.