7-2-2011

Characterizing within-die and die-to-die delay variations introduced By process variations and SOI history effect

James Aarestad

James C. Aarestad
Candidate

Electrical and Computer Engineering
Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

,Chairperson

CHARACTERIZING WITHIN-DIE AND DIE-TO-DIE DELAY VARIATIONS INTRODUCED BY PROCESS VARIATIONS AND SOI HISTORY EFFECT

by

JAMES C. AARESTAD

PREVIOUS DEGREES
ASSOCIATE OF APPLIED SCIENCE, ELECTRONICS,
NORTH DAKOTA STATE SCHOOL OF SCIENCE
BACHELOR OF SCIENCE, COMPUTER ENGINEERING,
UNIVERSITY OF NEW MEXICO

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Engineering

The University of New Mexico
Albuquerque, New Mexico

May, 2011

DEDICATION

This thesis is dedicated to my parents, Wilfred (W. C.) and Thelma Aarestad of Jamestown, North Dakota. Their love and care for their children, and the examples they have shown the world of grace and virtue through adversity, continue to inspire me and my work.

ACKNOWLEDGMENTS

It is with heartfelt gratitude that I offer my thanks to Dr. Jim Plusquellic, my advisor and committee chair, for providing me with the privilege of graduate school through his generous financial support, educational advisement and guidance. I am grateful to have been exposed to opportunities to learn and experience many facets of hardware research. Above all this, I am glad to have seen the potential there is in such a focused, dedicated, fertile mind.

I'd further like to thank my committee members, Dr. Payman Zarkesh-Ha and Dr. Marios Pattichis, for their willingness to share with me their knowledge and for their part in this important step in my professional development.

I also extend thanks to my closest teammates, Dr. Ryan Helinski, Greg Feucht, and Charles Lamech. Their friendship and encouragement has inspired me. So, too, have my friends and fellow students, Mitch Martin, Patrick Donnelly, Jeffrey Rohrbacher, and Amir Shirkhorshidian.

Thank you, Craig Kief and Dr. Steve Suddarth, for allowing me to experience life as a practicing design engineer. This was an invaluable part of my education.

To Dr. Howard Pollard, I have appreciated all that I have learned, and for working with you at Cosmiac. I wish to also acknowledge the National Science Foundation for their part in providing funding for my graduate education.

And, finally, I wish to thank my son, Scott Aarestad, for his example of strength, boldness, courage, and confidence. *Credo in te.*

CHARACTERIZING WITHIN-DIE AND DIE-TO-DIE DELAY VARIATIONS INTRODUCED BY PROCESS VARIATIONS AND SOI HISTORY EFFECT

by

JAMES C. AARESTAD

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Computer Engineering

The University of New Mexico
Albuquerque, New Mexico

May, 2011

CHARACTERIZING WITHIN-DIE AND DIE-TO-DIE DELAY VARIATION INTRODUCED BY PROCESS VARIATIONS AND SOI HISTORY EFFECT

by

James C. Aarestad

A.A.S., Electronics, North Dakota State College of Science, 1982
B.S., Computer Engineering, University of New Mexico, 2009

ABSTRACT

Variations in delay caused by within-die and die-to-die process variations and SOI history effect increase timing margins and reduce performance. In order to develop mitigation techniques to reduce the detrimental effects of delay variations, particularly those that occur within-die, new methods of measuring delay variations within actual products are needed. The data provided by such techniques can also be used for validating models, i.e., can assist with model-to-hardware correlation. In this research work, a method is proposed for a flush delay technique to measure both regional delay variations and SOI history effect. The method is then validated using a test structure fabricated in a 65 nm SOI process.

TABLE OF CONTENTS

LIST OF FIGURES viii

CHAPTER 1. INTRODUCTION 1

CHAPTER2. TEST STRUCTURE DESIGN / EXPERIMENTAL TECHNIQUES 3

2.1 SOI History Effect 3

2.2 Regional Delay 5

3. SOI HISTORY EFFECT (HE) MODELING 7

4. EXPERIMENTAL RESULTS 11

4.1 SOI History Effect 11

4.2 Regional Delay Variations 17

4.3 Within-Die Path Delay Variation Analysis 21

5. CONCLUSIONS 27

6. REFERENCES 29

LIST OF FIGURES

Figure 1. Block diagram of test structure's scan path.....3

Figure 2. Example waveforms and timing characteristics for POS pulse and NEG pulse experiments 4

Figure 3. Waveforms from regional delay experiments showing launch/capture cycle for 50 ns experiment 5

Figure 4. (a) Dominant sources of charge transfer in floating body of an SOI device, (b) est. of body potential swings for a switching inverter, (c)(d) impact on inverter delays 7

Figure 5. Percentage change in delay of the LEADING edge of a positive pulse emerging from the scan chain output plotted against ln(pulse delay) (x-axis) and pulse number (y-axis) showing SOI history effect 11

Figure 6. Percentage change in delay of the TRAILING edge of a positive pulse emerging from the scan chain output plotted against ln(pulse delay) (x-axis) and pulse number (y-axis) showing SOI history effect 12

Figure 7. Delay variation between 1^{st} and 5^{th} pulses (y-axis) for 35 chips (x-axis) in positive and negative pulse exps. with pulse width of 500 ns and pulse delay of 600 ns .13

Figure 8. Delay variation between 1^{st} and 5^{th} pulses (y-axis) for 35 chips (x-axis) in positive and negative pulse exps. with pulse width of 1000 ns and pulse delay of 2000 ns 14

Figure 9. Average delay variation in 5^{th} pulse (x-axis) using 1^{st} pulse as reference for pulse width 250 ns across pulse delay experiments for positive (top) and negative (bottom) pulse exps. 16

Figure 10. Average delay variation in 5^{th} pulse (x-axis) using 1^{st} pulse as reference for pulse width (a) 500 ns and (b) 1000 ns, across pulse delay experiments for positive (top) and negative (bottom) pulse exps. 17

Figure 11. Regional delay exps showing average number of FFs (AFF) traversed during each 5ns time interval 18

Figure 12. Blowup of Chip #2 regional delay experiments 18

Figure 13. Regional I_{on} variations across the array 19

Figure 14. Regional delay variations across 36 chips (y-axis) with the array partitioned into (a) 8 regions, (b) 16 regions. 20

Figure 15. Average variance (y-axis) and mean (x-axis) in the number of FFs traversed under each LC test computed across all chips. Includes both within-die and chip-to-chip variation. 22

Figure 16. Scaled average variance (y-axis) and mean (x-axis) of the number of FFs traversed under each LC test computed across all chips. Include within-die (and noise) variation. 23

Figure 17. Scaled (for chip-to-chip process variations and noise) average variance (y-axis) and mean (x-axis) delays under each LC test computed across all chips. Includes only within-die variation. 24

Figure 18. Positive edge experiments: Within-die delay variations expressed as % change against mean path delay on the x-axis for a set of 65 nm test chips, (a) 500 point view, (b) 30 point view. 25

Figure 19. Negative edge experiments: Within-die delay variations expressed as % change against mean path delay on the x-axis for a set of 65 nm chips, (a) 500 point view, (b) 30 point view. 26

1. INTRODUCTION

It is well established that the voltage on the isolated body of an SOI device varies as a function of its switching history and this voltage variation affects the threshold voltage of the device [1-5]. The variation in threshold voltage impacts the magnitude of the drain current and switching speed. Therefore, the switching speed of an SOI logic path depends on how often the path is exercised, with more frequent excitations resulting in faster switching speeds. The magnitudes of the delay variations vary widely (up to 15% per stage delay according to [1]) and depend on several process parameters, including well implantation, gate oxide thickness and halo implantation [4].

Previous works propose a variety of test structures for measuring delay variations introduced by SOI history effect (**HE**) [1][3] but many are fabricated and measured in dedicated, stand-alone test chip or scribe line contexts. In this work, a novel, minimally invasive technique is proposed that leverages the LSSD-style scan chain in actual products to allow HE-induced delay variations to be measured and analyzed.

A similar strategy is proposed for measuring die-to-die and within-die delay variations. The global nature of the scan chain allows delay variations in different regions of the chip to measured. By configuring the scan chain into flush delay mode (which effectively turns it into a long delay chain) and using a timed sequence of launch/capture edges on the scan input and scan clocks, regional, within-die variations in delay can be captured as a digital thermometer code and scanned out for analysis.

Previous work on the characterization of within-die and die-to-die delay variations focus on the use of ring oscillators (RO) as the basic test structure [6-7], with the exception of [8], which uses a custom test structure based on a 64-bit Kogge-Stone adder. The

authors report in [8] that within-die variation is spatially un-correlated but die-to-die variation is strongly correlated. In [9], the analysis of within-die and die-to-die delay variations show that die-to-die and layout-induced variations are significant.

Flush delay techniques have also been proposed, but in the context of fault detection [10], and speed-binning [11]. This is, to the author's knowledge, the first time that flush delay is proposed and used for measuring and analyzing HE and regional delay variations.

The remainder of the paper is organized as follows. Chapter 2. describes the test structure used in the hardware experiments, as well as the proposed techniques. Chapter 3 presents a model for SOI history effect, whose behavior is validated in the experimental results presented in Chapter 4 using a set of 65 nm SOI test chips. Conclusions are presented in Chapter 5.

2. TEST STRUCTURE DESIGN / EXPERIMENTAL TECHNIQUES

A block diagram of the test structure on the 65 nm chips is shown in Figure 1(a). The test-chip consists of an 80x50 array of test circuits (TCs) connected together through a scan chain. Each of the 4,000 TCs contains three master-slave FFs for a total of 12,000 FFs. The master-slave FFs are designed in an LSSD fashion, with separate clocks driving the master and slave latches as shown in Figure 1(b). The dual clock configuration allows a long delay chain to be created by setting both clocks high. Since each FF has two pass-gates and two inverters in series, the delay chain is effectively 48,000 gates long (12,000 FFs x 4 gates/FF).

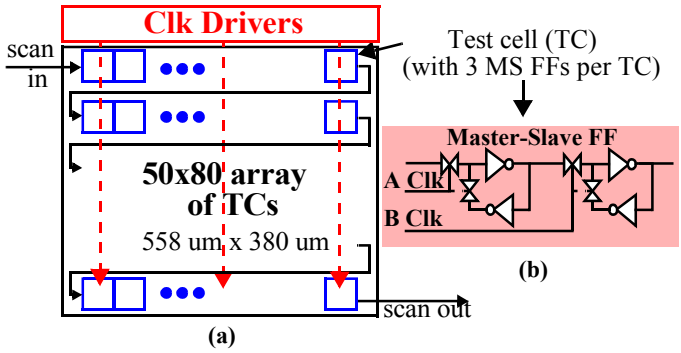

Figure 1. (a) Block diagram of test structure's scan path, and (b) FF with A/B Clocks to allow flush delay tests.

2.1 SOI History Effect (HE)

For the SOI HE experiments, a series of five positive pulses are driven into the *scan in* pin as shown by the waveform labeled as 'Input Signal' in the top plot of Figure . With both A and B clocks high, the pulses propagate through the entire scan chain and emerge at the scan chain output as shown by the waveform labeled as 'Output Signal' in Figure 2. The 'pulse delay' and 'pulse width' labels identify two additional parameters in the experiments. Pulse delay represents the time period between consecutive pulses, which

varies over the range from 300 ns to 100,000 ns in 15 experiments. Three different pulse widths of 250 ns, 500 ns and 1000 ns are also investigated. The bottom plot in Figure 2 shows the input and output waveforms from a second set of negative pulse experiments.

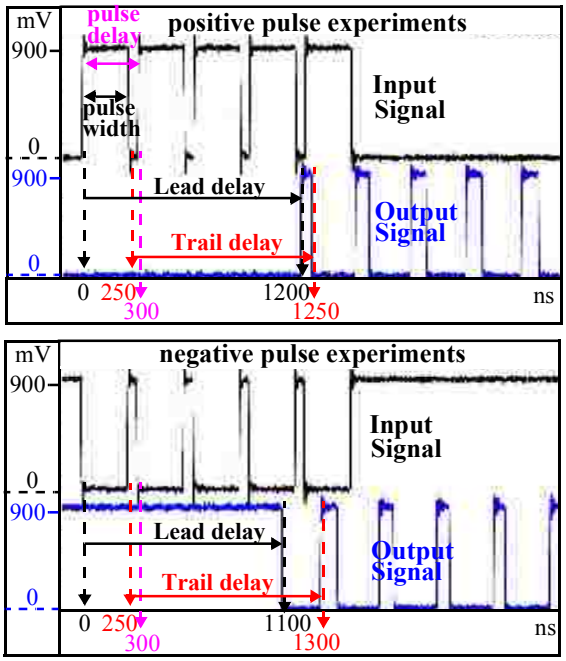

Figure 2. Example waveforms and timing characteristics for POS pulse (top) and NEG pulse (bottom) experiments.

The two timing parameters that lay at the center of the analysis in this work are labeled as 'Lead delay' and 'Trail delay' in the plots of Figure 2. These delays are measured across the input and output waveforms between the rising and falling edges, respectively, as shown by the arrows in figure Figure 2. It is clear from the figure that the width of the pulse that emerges from the scan chain changes across the five output pulses. These variations in pulse widths are caused by HE. In particular, the width of the emerging positive and negative pulses grows larger for consecutive pulses in both experiments. The change in the width of the consecutive, emerging pulses reflects the charging/discharging time constants associated with the floating channels of the pass gates and inverters, and is explained by the model in the Chapter 3. The results presented in Chapter 4 demonstrate

that the rate and magnitude of change in the pulse widths is a function of the input pulse width and pulse delay (switching frequency).

2.2 Regional Delay

A similar setup is used for the regional delay experiments, i.e., the A and B clocks are held high. For these experiments, however, only a single rising edge is introduced into the scan chain input. In order to obtain the delay in different regions of the test structure, the A clock is used to stop the propagating edge at specific time intervals after the edge is launched into the scan chain input. For example, Figure 3 shows the scan in and A clock waveforms for the 50 ns experiment. A rising edge is launched at time 0 and the A clock is driven low 50 ns later. With the scan chain initialized to all 0s, the number of 1's captured

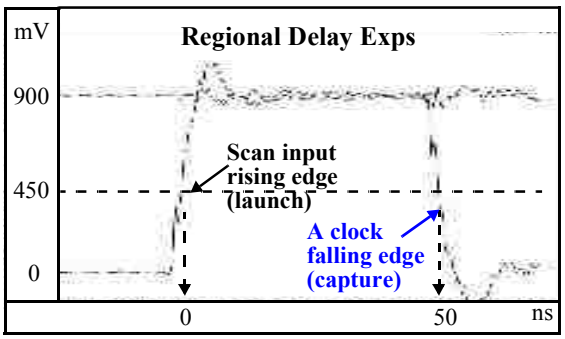

Figure 3. Waveforms from Regional Delay experiments showing launch/capture cycle for 50 ns experiment

in the scan chain indicates how far the edge propagated over the 50 ns time interval. Given the serpentine configuration of the scan chain as shown in Figure 1, longer timing delays between the launch/capture (**LC**) events measure the delay characteristics in larger portions of the array. A sequence of experiments was conducted on each of the chips in which the launch/capture delay was varied from 0 ns to approx. 1200 ns in 5 ns intervals, i.e., approx. 240 experiments were carried out per chip[1].

The number of FFs that the propagating edge traversed during any given 5 ns interval can be computed by subtracting the number of '1's measured under the previous test, e.g., the 45 ns LC test, from the number measured under the current test, e.g., the 50 ns LC test. In practice, measurement noise impacts the accuracy of the results but, fortunately, it can reduced by repeating the LC tests and computing an average. In these experiments, each experiment was repeated 12 times. Equation 1 gives the expression for computing the average number of additional FFs (*AFF*) traversed between any two tests, *k* and *k-n*. *N1*

$$\mathrm{AFF}_k = \frac{1}{12}\sum_{i=1}^{12}(\mathrm{N1}_i - \mathrm{AFF}_{k-n}) \qquad \textbf{Eq.1.}$$

indicates the number of '1's read from the scan chain. This expression allows regional variations in delay to be analyzed at various levels of granularity by choosing appropriate values for *n*. The average delay per FF in each region can also be derived by dividing the difference in the LC time intervals of two tests *k* and *k-n* (each of which is a multiple of 5 ns) by the value obtained from Eq. 1.

1. The actual delay along the chain in each chip determined the number of experiments, which varied because of process variations. For each of the chips, a terminal LC interval was reached beyond which the propagating edge was able to pass through all 48,000 gates in the scan chain and emerge from the SO pin.

3. SOI HISTORY EFFECT (HE) MODELING

Figure 4(a) shows the dominant charge transfer paths to and from the floating body of an SOI NFET device. The body is capacitively coupled to the FET terminals and hence a switching event at the drain, source, or gate of the device can dynamically inject charge in and out of the body. The floating body can also lose or gain charge due to static leakages associated with the PN junction diodes formed between the body and the source/drain nodes, and gate leakage currents. The dynamic capacitive coupling effect and the static leakage mechanism combine to create the HE, which manifests itself in the form of delay dependence on the switching history.

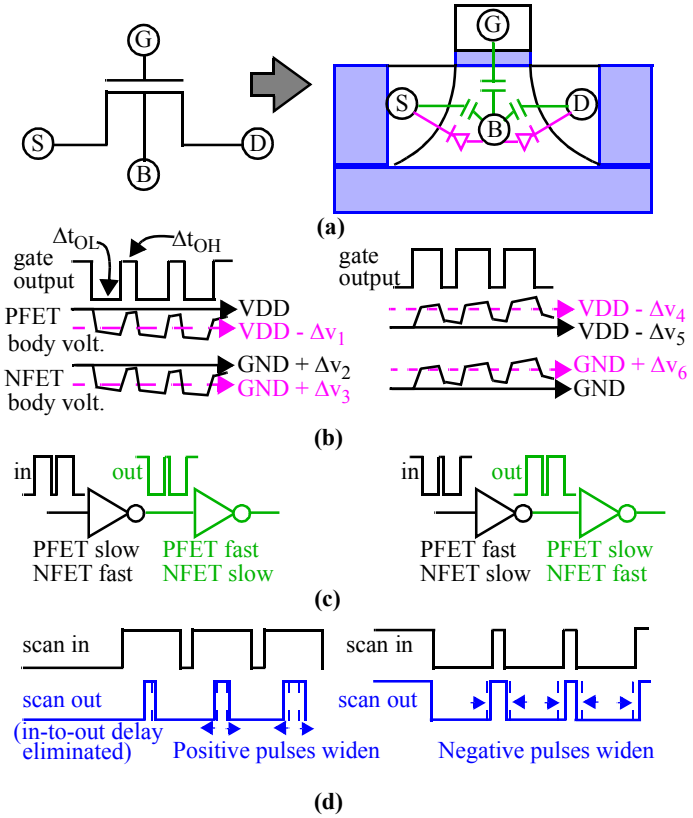

Figure 4. (a) Dominant sources of charge transfer in floating body of an SOI device, (b) est. of body potential swings for a switching inverter, and (c)(d) impact on inverter delays.

Since the actual body voltage swings in these experiments cannot be directly observed or measured, the analysis presented here is intuited from the observed behavior, which is

presented in the next chapter. References [1-5] may be consulted for further details on the various parameters that impact body voltage behavior.

Figure 4(b) shows two scenarios for the NFET and PFET body voltage initial values and swings for an inverter configuration. The scenario on the left depicts the situation where the gate input remains at logic 0 for a long period before the rising transition occurs while the scenario on the right shows the opposite situation. Note the initial values of the body voltages of the NFETs and PFETs are indicated by the solid lines in both scenarios. With the gate at DC for a long period of time, the leakage mechanisms in the NFET and PFET define the magnitudes of these long-term DC body voltages. From simulations, these long-term DC values are below/(above) VDD/(GND) by approx. 25% of the supply rail range for PFETs (NFETs).

Consider the arrival of a rising transition on the input of the inverter in the left scenario. The rising input injects charge into the body of both devices due to gate-to-body coupling (Figure 4(a)). A rising input also results in a falling transition at the drain nodes which pulls the body potential lower (in both devices) due to drain-to-body capacitive coupling. The drain-to-body coupling effect dominates the gate coupling effect, which is easily justified considering the device is conducting. Following the rising input transition, the output of the inverter is low for a time period Δt_{OL} as shown in the figure. During this period, the leakage mechanisms work to move the body voltages toward their long-term DC values under the new output state. When the falling transition arrives on the gate input, the opposite occurs, i.e., both body voltages rise due to the coupling effect and subsequently, leakage works to pull them to the alternate long-term DC values. Note that the relative magnitudes of the coupling and leakage voltage variations shown in the figure depict one

scenario of several that are possible.

The scenarios shown in Figure 4(b) purposely depict the duty cycles of the positive pulses on the inverter input as unequal. In particular, the duty cycle is greater than 50% for the scenario on the left, i.e., $\Delta t_{OL} > \Delta t_{OH}$, while it is less than 50% for the scenario on the right. The overall effect of the asymmetry in the duty cycle works to exacerbate the rate and magnitude of the "drift" of the body voltages to specific 'average' DC values (shown by the dotted lines). The average DC values occur between the maximums defined for the two long-term DC values. From the diagram, it is also apparent that these final 'average' body voltages are not attained instantaneously, but rather several cycles are necessary to reach them.

The variations in the average body voltages affect the threshold voltages of the PFET and NFET devices in each inverter and, correspondingly, the delay. Figure 4(c) labels the impact on the transistors in a two-inverter chain, for each of the two scenarios shown in Figure 4(b). The downward movement of the body voltages for the scenario on the left degrades the responsiveness of the PFET in the first inverter increasingly over time, i.e., as more transitions occur on its input. At the same time, the responsiveness of the NFET improves. The overall effect on the inverter's delay is that falling output transitions occur sooner in time than they would under the initial condition of the NFET body voltage, while rising output transitions occur later in time. Unfortunately, these variations in delay culminate at gates downstream as shown for the second inverter in the left scenario of Figure 4(c). Here, the opposite conditions exist, resulting in body voltage behavior as depicted in the right scenario of Figure 4(b).

Figure 4(d) shows the history effect on the delay of a chain of inverters (a path) under

each of the scenarios. The stimulus applied to the path input is labeled as ‘scan in’ (to relate it to the test structure shown in Figure 1), while the response waveform on the path output is labeled ‘scan out’. The ‘scan out’ waveform has been skewed to the left in time, i.e., the overall path delay is eliminated, to allow comparisons of pulse widths between the input and output waveforms. Differences in the capacitive loads along the path and HE both act to change width of the output pulses in both scenarios (also shown in Figure 2). However, delays introduced by HE additionally change over time, which is reflected across the sequence of output pulse widths. In both scenarios, consecutive pulses become increasingly wider (dotted lines show the expected result without HE) because the first edge is sped up while the second edge is slowed down.

4. EXPERIMENTAL RESULTS

4.1 SOI History Effect

The experiments described in Chapter 2. allow delay variations introduced by HE to be measured as a function of several parameters. As indicated in the previous chapter, the main contributor to delay variation introduced by HE is the frequency of excitation of the logic path. A second contributor is the relative amount of time the logic path is maintained in one of the two possible states. To investigate these parameters, a set of five pulses are applied to the *scan in* input as shown in Figure 1 using a variety of *pulse delay* and *pulse width* values. Most of the previous work focuses on 'first-switch, second-switch'. In this work, it is shown that, as alluded to in the previous chapter, HE plays a role in delay variations beyond the second switch, and the variation in delay over the sequence of pulses behaves like a RC time constant.

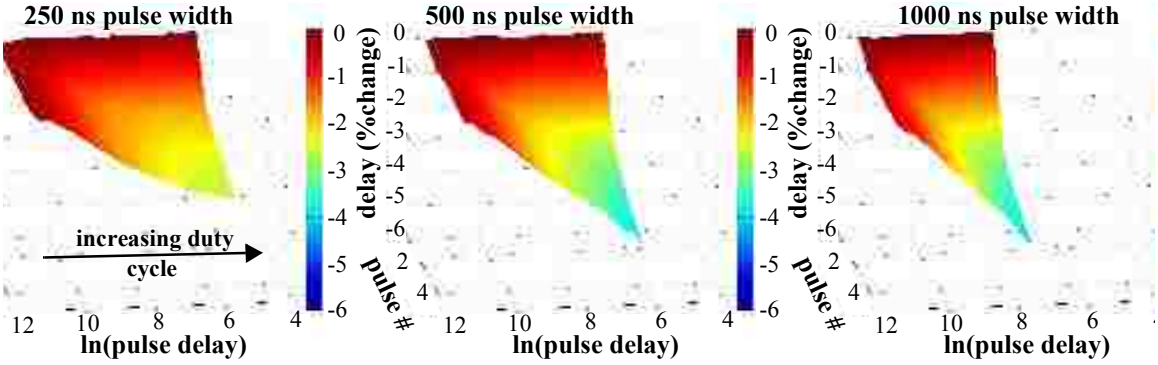

Figure 5. Percentage change in delay of the LEADING edge of a positive pulse emerging from the scan chain output plotted against ln(pulse delay) (x-axis) and pulse number (y-axis) showing SOI history effect.

As an illustration of the impact of HE on delay, Figures 5 and 6 show the results obtained from a sequence of positive pulse experiments carried out on one of the chips. In each of these 3D plots, the x-axis plots *ln* (natural log) of the pulse delay in ns, with range of 4 to 12 from right to left. The actual pulse delays used were 300, 400, 500, 600, 800, 1,000, 2,000, 4,000, 8,000, 16,000, 20,000, 40,000, 60,000, 80,000 and 100,000 ns. The y-

axis represents the pulse number from 5 (front) to 1 (back) while the z-axis plots the percentage change in delay under each of these 75 experiments (15 pulse delays * 5 pulses per train). The percentage change ($\overline{pch}$) is computed with respect to the upper left-most data point, i.e., the *reference* delay is the delay of the first pulse under pulse delay experiment of 100,000 ns. Eq. 2 gives the expression for *pch* where d_t represents the measured delay in experiment *t* (measured as shown in Figure 2) and d_{ref} is the reference delay. Each

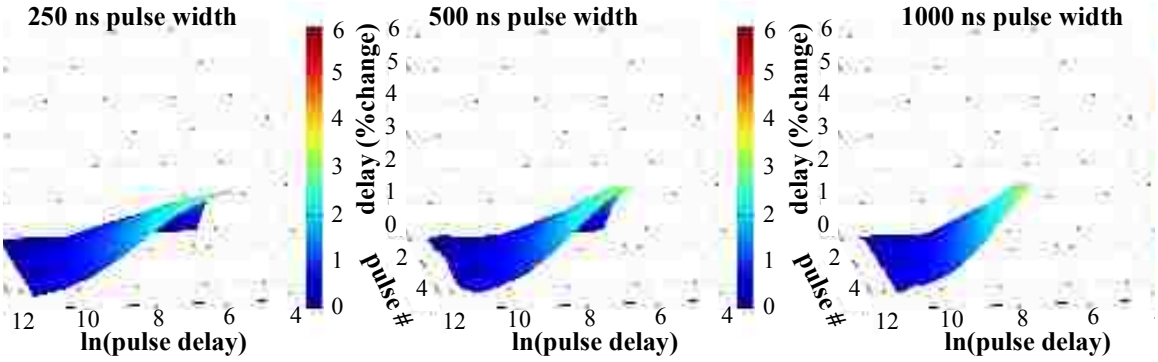

Figure 6. Percentage change in delay of the TRAILING edge of a positive pulse emerging from the scan chain output plotted against ln(pulse delay) (x-axis) and pulse number (y-axis) showing SOI history effect.

$$\text{pch} = \frac{(d_t - d_{\text{ref}})}{d_{\text{ref}}} 100 \qquad \textbf{Eq. 2.}$$

of the three plots in Figure 5 show the delays of the leading edge in experiments with the pulse width set to 250 ns, 500 ns or 1,000 ns (note that the smaller number of pulse delays represented on the x-axis, e.g., 300 ns, are not relevant for the 500 ns and 1000 ns pulse width experiments). Figure 6 gives the results for the trailing edge under the same conditions.

The negative values on the z-axis from Figure 5 indicate that the reference experiment generates the longest leading edge delay. For the 100,000 ns pulse delay experiment, the long Δt between each pulse allows the long-term DC body voltage conditions to be re-established. Therefore, the leading edge delays portrayed on the left side of the surface plots are constant and nearly equal to the reference delay. As the frequency of the pulses is

increased, the delay of the first edge remains constant (back edge of plots). However, subsequent pulses begin to experience HE, resulting in a smaller delay of approx. 4-5% for this chip. HE takes hold earlier and more significantly for wider, i.e. 500 ns and 1,000 ns, pulse widths, as shown by the center and right-most plots. The trailing edge analysis shown in Figure 6 depicts opposite behavior, where the delays of the trailing edges from later pulses under higher pulse frequencies increase.

In order to add perspective with regard to the magnitude of delay variation introduced by HE, that component of the delay is compared with chip-to-chip process variations in Figure 7. The graph plots the delays of the 1st and 5th leading and trailing edges obtained

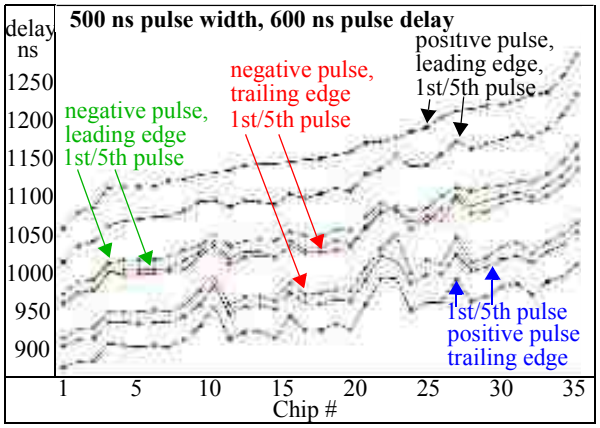

Figure 7. Delay variation between 1st and 5th pulses (y-axis) for 35 chips (x-axis) in positive and negative pulse exps. with pulse width of 500 ns and pulse delay at 600 ns.

from the set of 500 ns pulse width, 600 ns pulse delay experiments along the y-axis for each of the 35 chips (x-axis). The chips are sorted by the delays measured under the positive pulse, leading edge experiments (top-most waveform). Each of the four pairs of waveforms corresponds to the delays of one of the edges (leading or trailing) in either the positive or negative pulse experiments. Figure 8 gives the results under similar conditions, for the 1000 ns pulse width, 2000 ns pulse delay experiments.

Given this plot, it is straightforward to compute the relative variations introduced by

global process variations and HE. The % change in delay, measured using the left-most and right-most data points in the top-most waveform is approx. 21%, while the % change due to HE is approx. 4.5%, given by the left-most data points of the top two waveforms. The % change in the negative pulse, trailing edge waveforms is similar but opposite in polarity, and is smaller for the other two waveform pairings. Interestingly, HE delay variations are relatively constant across the chips, i.e., the vertical spacing between waveform pairs, and therefore appear to be relatively insensitive to process variations.

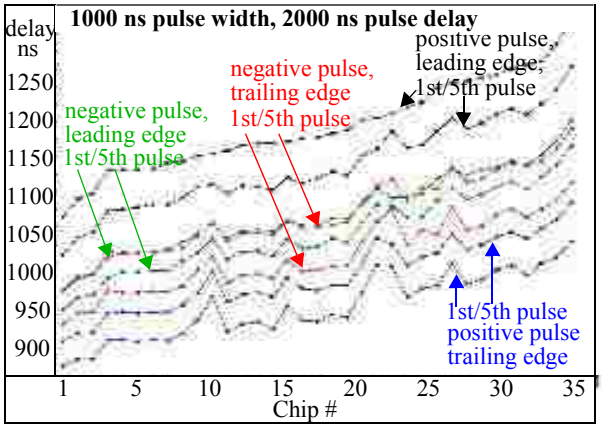

Figure 8. Delay variation between 1st and 5th pulses (y-axis) for 35 chips (x-axis) in positive and negative pulse exps. with pulse width of 1000 ns and pulse delay at 2000 ns.

The waveforms in Figure 8 are similar in shape to those shown in Figure 7, as expected given that the same chips are used in the experiments. However, the vertical displacements of the waveforms corresponding to the 5th pulse are slightly larger with respect to their corresponding reference pulse (1st pulse), suggesting that the average body voltages described in Chapter 3 are slightly exacerbated under this particular pulse width and pulse delay. This is predicted by the model derived herein, which indicates that changes in the duty cycle will correspondingly impact the magnitude of HE delay variations.

A third interesting observation is the difference in the vertical spacings of the waveform pairs associated with the positive and negative pulse experiments. The positive pulse

waveforms occupy the vertical extremes, which is counter-intuitive given the symmetry of a scan chain architecture. However, the test chip's scan chain is not completely symmetrical as illustrated in Figure 1. In particular, the right-most TCs on each row drive long wires that connect the last element of the row with the first element of the next row. As indicated in Chapter 2, each row has an even number of FFs (and inverters). Therefore, the positive pulse experiment requires the inverters on the right edge of the array to drive a rising edge onto each of these long wires (80 total). Given that PFETs tend to be weaker drivers than NFETs, this contributes to the asymmetrical offsets associated with the waveform pairs in Figure 7. A second, more important, asymmetry that exists in the chain is the difference in the capacitive loads driven by the outputs of the FFs (not shown). In particular, the slave components of the FFs drive an additional fanout load (in addition to the input of the next FF). These asymmetries add approx. 8% variation in delay. However, unlike HE and process variations, this source of variation can be eliminated by sizing the transistors appropriately.

The vertical ordering of waveform pairs shown in Figures 7 and 8 indicate that pulse expansion occurs for both the positive and negative pulses experiments with a pulse width of 500 ns and pulse delay of 600 ns. This type of pulse expansion behavior occurs in all of the experiments, but the magnitude is dependent on the pulse delay, and to a smaller degree, on the pulse width. Figure 9 shows the *average* pulse expansion that occurs in each of the experiments (y-axis), expressed as percentage change (x-axis). The average is computed across all 35 chips for pulse width experiments of 250 ns. The results for the 500 ns and 1000 ns pulse width experiments are shown in Figure 10(a) and (b), resp. The horizontal bars give the mean and 3 σ limits as indicated in Figure 9. The percentage

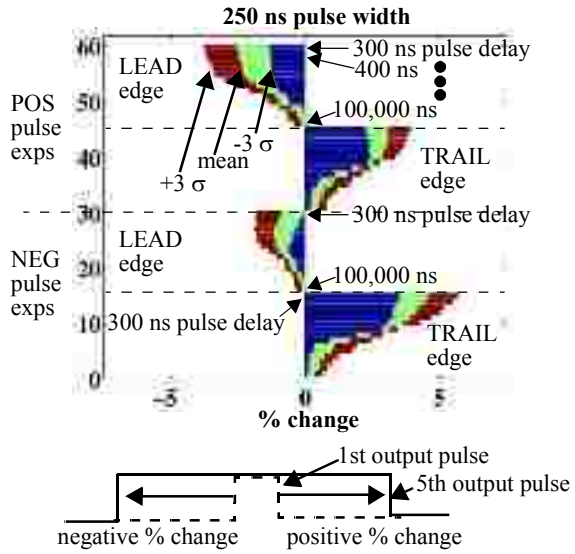

Figure 9. Average delay variation in 5th pulse (x-axis) using 1st pulse as reference for pulse width 250 ns across pulse delay experiments for positive (top) and negative (bottom) pulse exps.

change value is computed using the LEAD (or TRAIL) delays, measured as shown in Figure 2, of the 1st pulse (reference) and 5th pulse, as given by Eq 3. Bars that have a wider

$$\%\ \text{change} = \frac{(d_{5\text{th}} - d_{1\text{st}})}{d_{1\text{st}}} 100 \qquad \textbf{Eq3.}$$

excursion in the negative or positive x-dimension indicate that the edges of the 5th output pulse have moved more dramatically under the experiment. As shown by the graphic under the horizontal bar graph in Figure 9, edges from the 5th pulse that move to the left of the corresponding reference (1st) pulse, generate negative % change values, while those moving to the right generate positive values. The pulse delay increases downward within each group of bars, which causes a corresponding decreases in the difference between the edges of the 1st and 5th pulses. In the worst case, e.g., for the TRAIL edge delay at 600 ns in the NEG pulse experiment in Figure 10(a), HE introduces approx. a 6% variation in delay. As expected, HE variations approach 0 for large Δt's, e.g., 100,000 ns, betweeen consecutive pulses in all cases.

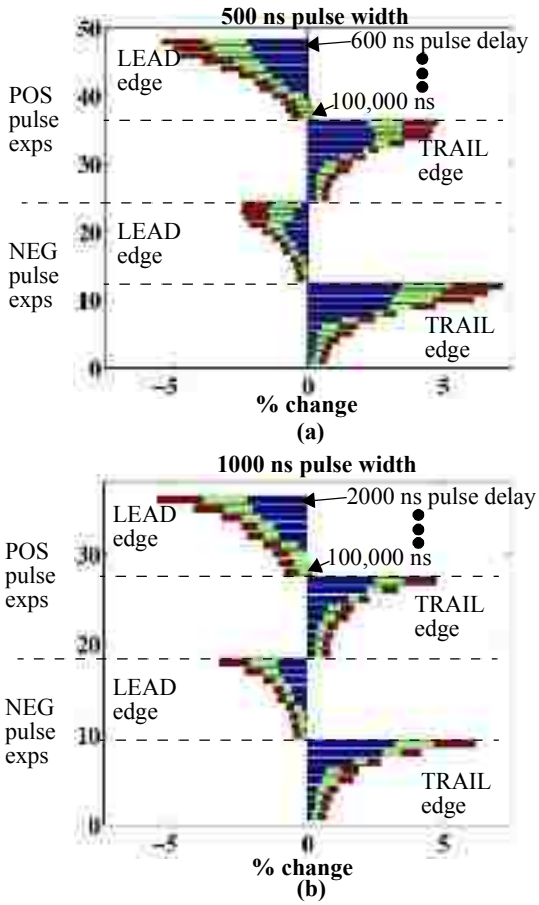

Figure 10. Average delay variation in 5th pulse (x-axis) using 1st pulse as reference for pulse width (a) 500 ns and (b) 1000 ns, across pulse delay experiments for positive (top) and negative (bottom) pulse exps.

4.2 Regional Delay Variations

The use of the launch/capture scheme described in Chapter 2.2 enabled the measurement of regional delay variations, and comparison of these with the chip-to-chip variations reported in the previous chapter. As described in Chapter 2.2, a sequence of experiments was carried out on each of the chips in the population, in which the capture event (the act of driving the A clock low) was increased from 0 ns to approx. 1200 ns in 5 ns increments. After each launch/capture experiment, both the A and B clocks were used to scan out the sequence of 12,000 bits to determine how far the propagating edge advanced through the scan chain.

Figure 11 shows the results for two of the chips from the population of available devices. The x-axis gives the LC time interval while the y-axis plots the average number of FFs (AFF) traversed during each consecutive 5 ns time interval, computed using Equation 1. The center waveform represents the AFF values while the bounding waveforms (top and bottom) represent the upper and lower 3 σ limits (computed using the 12 samples taken for each LC test). The AFF values vary around 52 and 48 for Chip #1 and #2, resp. which indicates the number of FFs that the edge propagates through during each 5 ns time interval. The limits increase from left to right for both chips, indicating that the level of uncertainty increases for regions at the bottom of the array (see Figure 1).

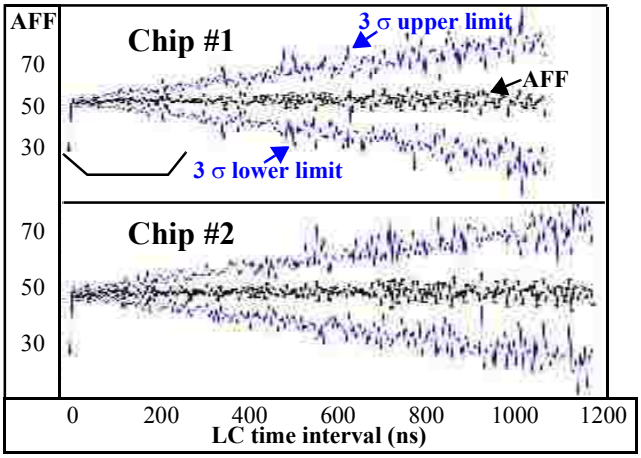

Figure 11. Regional delay exps showing average number of FFs (AFF) traversed during each 5 ns time interval.

A blow-up of the left-most region of Chip #1 from Figure 11 is shown in Figure 12. The

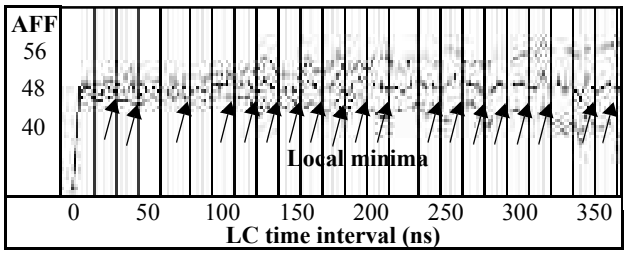

Figure 12. Blow-up of Chip #2 regional delay experiments.

vertical bars in the plot identify time intervals in which the propagating edge moves onto a

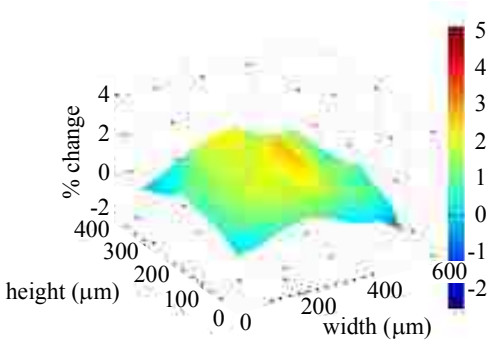

Figure 13. Regional I_{on} variations across the array.

new row in the array. The long wires connecting consecutive rows add wire delay and should therefore reduce the AFFs traversed in these LC tests. The arrows shown in Figure 12 illustrate where local minima occur in the measured AFF values. The absence of some local minima illustrates that although the wire delay component is measurable, the uncertainty in the measurements makes it difficult to measure them accurately. This is particularly evident in lower regions of the array (not shown in Figure 12) where the level of uncertainty increases.

Although the analysis of regional delay variations at the 5 ns scale is interesting, it is dominated by local (and random) process variation effects. The large magnitude of these random variations make it difficult to observe within-die and across-macro systematic delay variations. As indicated in Chapter 2.2, the parameters to Equation 1 allow other levels of granularity of delay variations to be measured and analyzed.

Of particular interest for this research has been to determine if the variation in I_{on} current, as shown in Figure 13, is correlated to regional delay variations. The (x,y) plane of the figure represents the spatial domain of the array while the z-axis plots the percentage change in the magnitude of I_{on} currents in different regions of the array. From the figure, it is clear that I_{on} is larger in the central region of the array, and decreases toward the edges.

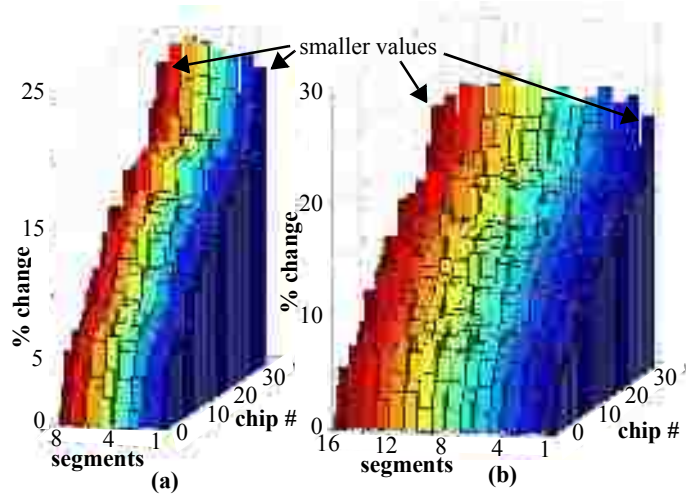

Figure 14. Regional delay variations across 36 chips (y-axis) with the array partitioned into (a) 8 regions, (b) 16 regions.

Systematic, within-die variations in delay are more easily observed by partitioning the array shown in Figure 1 into larger regions. Equation 1 is used to derive regional delay variations at granularities of 8 and 16 by dividing the total number of LC tests by these numbers and computing a 'normalized' AFF value for each segment. This is accomplished by computing the AFF value for each segment (using Equation 1) and dividing by the number of LC tests in each segment. Intuitively, the larger regions also reduce the level of uncertainty in the reported values.

Figure 14(a) and (b) give the results for the 8 and 16 segment analyses, respectively, for 36 chips. The segments are plotted along the x-axis (in reverse order), chip numbers along the y-axis, and the normalized, regional delays as percentage change along the z-axis. The reference component is segment #1 from chip #1 (lower, right-most element). Although die-to-die variations are clearly visible, the within-die delay variations are also observable. In particular, the values on the edges of the graphs, i.e., those corresponding to segments 1 and 8 from Figure 14(a) and to 1 and 16 from Figure 14(b), are slightly smaller for each chip by approx. 1-2% under the 8 segment analysis, and 2-3% for the 16 segment

analysis. This correlates well with the I_{on} analysis presented in Figure 13, that shows about a 5% reduction along the edges when compared with the central portion of the array. Given that the regions analyzed under the delay analysis are 1-D in nature (in the y dimension of the array only), it is reasonable to expect only the top (segment 1) and bottom (segment 8 or 16) components to be smaller than the other regions. This is true because increases in delay for gates along the left and right edge of the array appear in every segment and therefore cancel out. This, in turn, reduces the magnitude of the measured delay variations to about half of that measured for I_{on}.

4.3 Within-Die Path Delay Variation Analysis

The regional delay experiments are also used to determine the intrinsic delay variations in this technology as a function of path length. Intuitively, for any technology, within-die variance should increase as the number of inverters that the edge propagates through decreases. The focus in this case is to determine the impact of within-die variations on delay in this advanced technology node. In order to extract within-die variations from the chip data, additional experiments and a sequence of post processing steps are needed.

The additional experiments are designed to measure edge propagation for small path segments at the beginning of the scan chain. The graph in Figure 15 illustrates that the shortest testable path segment in the initial experiments is between 25 and 30 FFs, which constitutes a path longer than 100 gates in length (4 gates per FF). This was true because the A Clk path from the pattern generator to the actual test structure is longer than the path for the scan in signal. Even with the LC time interval set to 0 on the pattern generator for

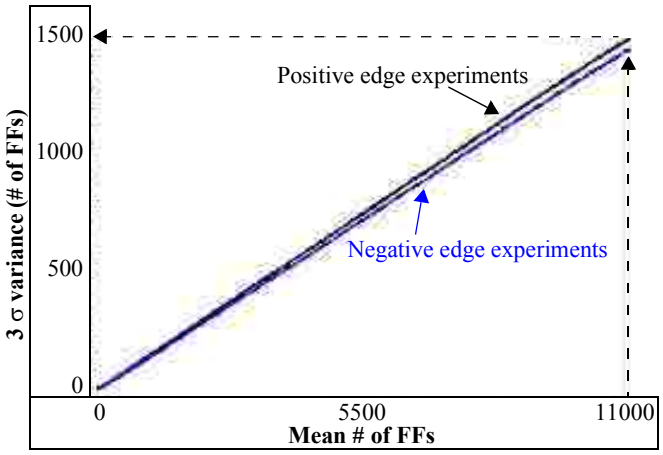

Figure 15. Average variance (y-axis) and mean (x-axis) in the number of FFs traversed under each LC test computed across all chips. Includes both within-die and chip-to-chip variation.

the first LC test, which launches the scan in signal and the A Clk simultaneously, the difference in these on- and off-chip path delays allowed the edge to propagate along a sufficiently long segment of the scan chain, e.g., 25 FFs. Although it was possible to configure the pattern generator to drive the A Clk low before launching the edge on the scan in pin, the resolution of adjusting this time interval was limited to 5 ns increments. Instead, an alternative, higher resolution scheme was used, wherein the relative lengths of the co-axial cables were changed for these two signals between the pattern generator and their board connections. Three new cabling configurations were used that enabled actual on-chip LC time intervals as low as 250 ps, i.e., where 'actual' is defined as the time interval between the arrival of the scan in signal on the input of the first FF in the scan chain and the arrival of the capturing A clk on the clock terminals of the FFs in the first row.

The curves in Figure 15 depict the unprocessed data from the original and additional experiments. The two curves are derived from the average of data collected from all 36 chips under the positive edge and negative edge experiments, respectively. The x-axis plots the mean number of FFs traversed under each of the LC tests. Since some chips are slower than others, the number of LC tests that could be applied to each chip varied. In

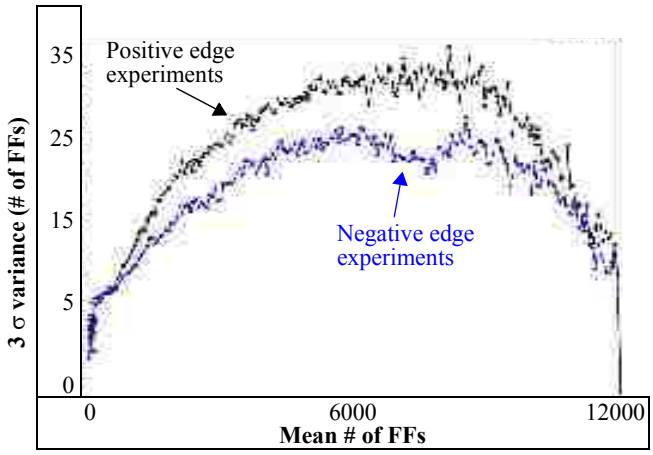

Figure 16. Scaled average variance (y-axis) and mean (x-axis) of the number of FFs traversed under each LC test computed across all chips. Includes within-die (and noise) variation.

order to include data from all chips in each mean value computed, the maximum number of LC tests that could be applied to **all** chips is used as the last LC test (which is dictated by the fastest chip in the population of 36 chips used in this research). The LC time interval for this test, subsequently referred to as the *terminal LC test*, is 1055 ns under the positive edge experiments and is 965 ns under the negative edge experiments. From the plot, the mean number of FFs traversed is slightly less than 11,000 and the 3 σ variance, plotted along the y-axis, is approx. 1,500 FFs. The curves are nearly linear, indicating that the variation scales with the average length of the path.

Unfortunately, chip-to-chip variations dominate the behavior in the unprocessed data as shown in Figure 15. However, chip-to-chip variations can be removed from the raw data by 'scaling' the data from each chip by a ratio. The ratio is computed by dividing the 'number of FFs' measured from each chip under the terminal LC test by that measured under the fastest chip for this LC test (fastest chip values are 11,935 and 11,920 for the positive and negative experiments respectively). This effectively eliminates chip-to-chip variations and leaves only within-chip and noise variations in the data. Figure 16 plots the

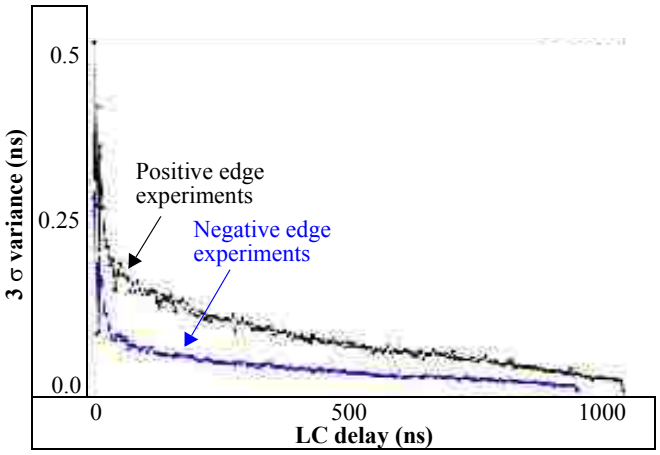

Figure 17. Scaled (for chip-to-chip process variations and noise) average variance (y-axis) and mean (x-axis) delays under each LC test computed across all chips. Includes only within-die variation.

results in the same format as given for Figure 15, with the mean number of FFs traversed across all chips along the x-axis, against the 3 σ variance in these values on the y-axis. Although the chip-to-chip variance is eliminated, e.g., the maximum variance is reduced from approx. 1500 to 35 FFs, the variance is now dominated by the noise in the measurements. The 'hump' like structure of the curves, in particular, the decreasing magnitude of variance for LC tests on the right, is an artifact of the scaling operation, which uses the right-most LC data values as the reference. In other words, after scaling, the right-most data points are all equal and therefore have 0 variance.

From Figure 11, it is clear that noise is proportional to the LC time interval, i.e., larger LC time intervals are correspondingly noisier. The apparent linear trend in noise as a function of path length suggests that noise, like chip-to-chip process variations, can also be removed from the raw data using, in this case, a set of scaling factors. First, the variance introduced by noise for each LC test is computed using the 12 samples collected for each chip. These are the values shown in Figure 11. The average variance is then computed across all chips for each LC test. Note these calculations use the unscaled data from Figure

15. A set of ratios are then computed for each LC test by dividing the reference noise variance (from the first LC test) by the noise variance computed at each LC test. These ratios are then used to multiply the values shown in Figure 16. Given the noise levels at the first LC test are close to 0, this strategy effectively removes the noise from the entire data set. Figure 17 gives the results in a similar format to the previous figures except the '# of FFs' is converted to actual delays. Here, for the first time, the intuitive result can be shown that the variance in delay for shorter segments of the scan chain, as shown on the left in the figure, is larger than it is for longer paths. It is also clear that this within-die variation is non-linearly related to the length of the path. In particular, the left-most data point represents a segment of the scan chain that has approximately a 500 ps delay in the positive edge experiments and a 250 ps delay in the negative edge experiments (it was indicated earlier

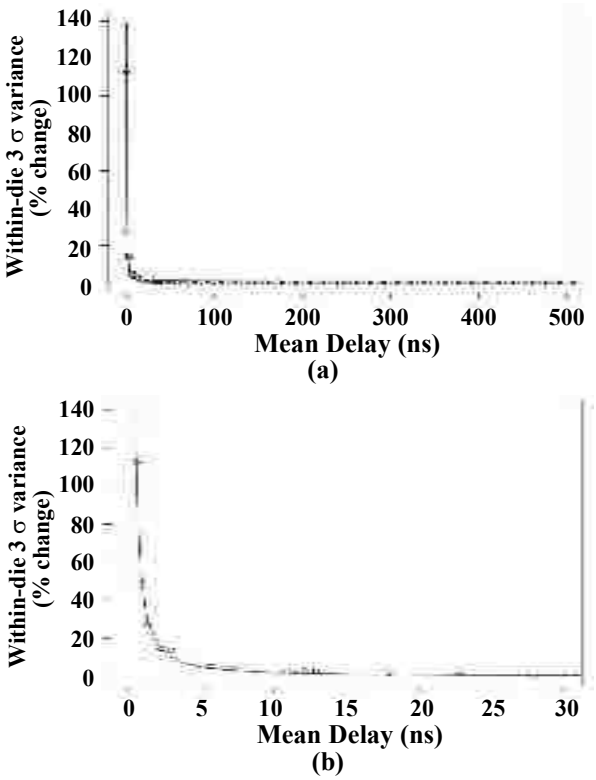

Figure 18. Positive edge experiments: Within-die delay variations expressed as % change against mean path delay on the x-axis for a set of 65 nm test chips, (a) 500 point view, (b) 30 point view.

that additional LC tests were carried out to measure these small path segments at the beginning of the scan chain). From the graph, the variance in these paths is also approx. 500 ps and 250 ps, respectively. Therefore, variations in short paths, e.g., 10 to 20 gates in this technology, approaches 100%.

This is illustrated in Figures 18 and 19 for the positive and negative edge experiments, resp., which plot the variance as percentage change on the y-axis for two 'zoomed-in' views, one with 500 points (a) and one with 30 points (b) on the x-axis. The % change curves are well fit with exponential curves, shown plotted on top of the data points. 95% confidence interval curves are also plotted but are difficult to see because the limits are close to the mean curves. The upward trajectory of the left-most portion of the curves confirm that delay variations in short paths are significant.

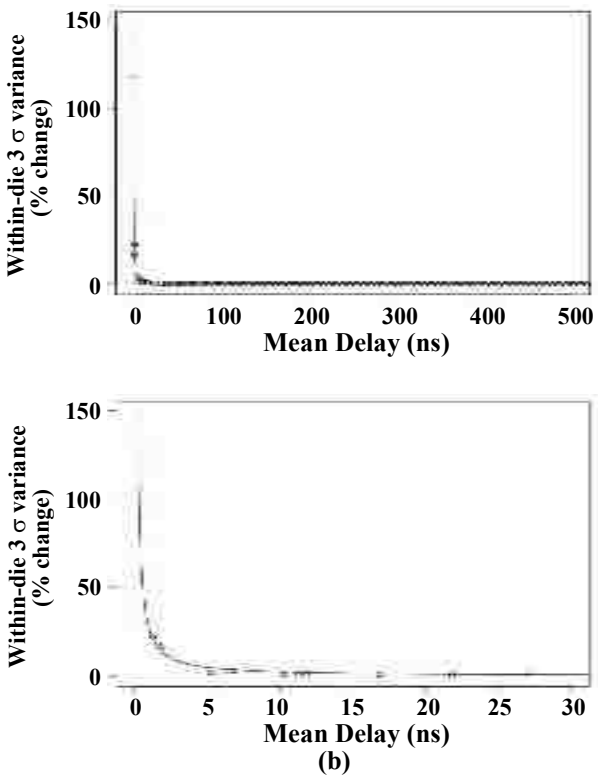

Figure 19. Negative edge experiments: Within-die delay variations expressed as % change against mean path delay on the x-axis for a set of 65 nm test chips, (a) 500 point view, (b) 30 point view.

5. CONCLUSIONS

In this work, there has been proposed two FLUSH delay based techniques for measuring delay variations introduced by SOI history effect and regional process variations. A model to explain the behavior of SOI devices under differing dynamic conditions was derived and validated through experimentation. The analysis was carried out on a test structure fabricated in IBM's 65 nm SOI technology. The results show that worst-case delay variations introduced by HE are approximately 4.5% while those caused by chip-to-chip process variations can be as large as 21%. Worst-case systematic, within-die process variations introduce delay variations of approximately 1-3% while those caused by random, within-die process variations can be as large as 10%.

The second proposed technique used a pair of launch-capture (LC) signals, at 5 ns intervals, to propagate an edge at full device speed through the FLUSH-enabled scan chain of the test chip. This testing provided insight into regional variations in the delay across the individual test chip.

A number of techniques were used to normalize the collected data and to eliminate chip-to-chip variation and measurement noise, which provided a clearer analysis of the actual within-die variation. This analysis permitted a more meaningful comparison between within-die variation and the effects of spatially correlated I_{on} variation. From the analysis of the conditioned data, it became readily apparent that the variations in the regional delay grew exponential greater as the LC intervals approached the shortest testable paths.

Additionally, a comparative evaluation was made of the differing effects of process variation on average propagation delay and the pulse shrinking/expanding caused by SOI HE. The experimental data suggests that, while the total delay through the chips varied,

from the fastest to the slowest, by approximately 21%, the change in delay due to history effect is approximately 4.5% and shows no correlation to PV at all.

6. REFERENCES

[1] K. A. Jenkins, S. Kim, S. P. Kowalczyk, D. Friedman, “Impact of SOI History Effect on Random Data Signals,” in Proc. of *Integrated Circuit Design and Technology*, 2007, pp. 1-4.

[2] S. Narendra, J. Tschanz, A. Keshavarzi, S. Borkar, V. De, ”Comparative Performance, Leakage Power and Switching Power of Circuits in 150 nm PD-SOI and Bulk Technologies including Impact of SOI History Effect,” in Proc. *VLSI Circuits*, 2001, pp. 217-218.

[3] O. Faynot, T. Poiroux, J. Cluzel, M. Belleville, J. de Pontcharra, “A New Structure for In-Depth History Effect Characterization on Partially Depleted SOI Transistors,” in Proc. of *SOI Conference*, 2002, pp. 35-36.

[4] Q. Liang *et al*, “Optimizing History Effects in 65nm PD-SOI CMOS,” in Proc. of *SOI Conference*, 2006, pp. 95-96.

[5] S.K.H. Fung *et al*, “Controlling floating-body effects for 0.13 um and 0.10 um SOI CMOS”, In Proc. of Electron Devices Meeting, 2000, pp. 231

[6] B.P. Das, B. Amrutur, H.S. Jamadagni, N.V. Arvind, V. Visvanathan, "Within-Die Gate Delay Variability Measurement using Re-configurable Ring Oscillator," in Proc. of *Custom Integrated Circuits Conference*, 2008, pp. 133-136.

[7] H. Onodera, H. Terada, "Characterization of WID Delay Variability using RO-array Test Structures," in Proc. of *International Conference on ASIC*, 2009, pp. 658-661.

[8] N. Drego, A. Chandrakasan, D. Boning, "All-Digital Circuits for Measurement of Spatial Variation in Digital Circuits," *Solid-State Circuits*, vol.45, no.3, 2010, pp. 640-651.

[9] P. Liang-Teck, B. Nikolic, "Measurements and Analysis of Process Variability in 90 CMOS," *Solid-State Circuits*, vol.44, no.5, May 2009, pp. 1655-1663.

[10] F. Yang, S. Chakravarty, N. Devta-Prasanna, S.M. Reddy, I. Pomeranz, "On the Detectability of Scan Chain Internal Faults An Industrial Case Study," in Proc. of *VLSI Test Symposium*, 2008, pp. 79-84.

[11] C. Thibeault, "On the Potential of Flush Delay for Characterization and Test Optimization," in Proc. of *Current and Defect Based Testing Workshop*, 2004, pp. 55- 60.