Pitfalls and Tradeoffs in Simultaneous, On-Chip FPGA Delay Measurement

Timothy A. Linscott
Seattle University
901 12th Avenue
Seattle, WA 98122
timothy.a.linscott@gmail.com

Benjamin Gojman*
University of Pennsylvania
3330 Walnut St.
Philadelphia, PA 19104
bgojman@acm.org

Raphael Rubin
University of Pennsylvania
3330 Walnut St.
Philadelphia, PA 19104
rafi@seas.upenn.edu

André DeHon
University of Pennsylvania
200 S. 33rd St.
Philadelphia, PA 19104
andre@acm.org

ABSTRACT

Recent work shows how to use on-chip structures to measure the fabricated delays of fine-grained resources on modern FPGAs. We show that simultaneous measurement of multiple, disjoint paths will result in different measured delays from isolated configurations that measure a single path. On the Cyclone III, we show differences as large as ±33 ps on 2 ns-long paths, even if the simultaneously configured logic is not active. This is over 20× the measurement precision used on these devices and over 50% of the observed delay spread in prior work. We characterize the magnitude of the impact of simultaneous measurements and identify strategies and cases that can reduce the difference. Furthermore, we provide a potential explanation for our observations in terms of self-heating and the configurable clock network architecture. These experiments point to phenomena that must be characterized to better formulate on-chip FPGA delay measurements and to properly interpret their results.

Keywords

FPGA; Timing; Self Measurement; Component-Specific Map

1. ON-CHIP DELAY MEASUREMENT

Recent work [10, 11, 9, 12, 7] shows how to perform on-chip, self measurement of the delay of FPGA resources. In one strategy [10, 7] registers are placed around a path to be measured composed of LUTs and wires (Circuit Under Test, or CUT, Fig. 1). The self test programs the on-chip PLLs to vary the clock period for the registers and identifies when the path fails because the path delay exceeds the clock period.

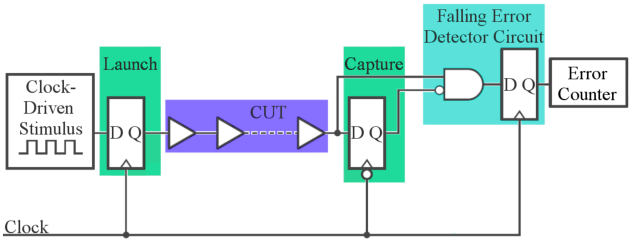

Figure 1: Path-Delay Circuit Under Test (CUT)

Gojman showed that a Cyclone III LAB has over 2,400 independently varying delay components [6], meaning even a small FPGA with 963 blocks can have over two million component delays and demand measurement of five to six million paths. Performed serially, each measurement requires both reconfiguration time, T_{reconf}, to define the path and testing time to exercise a configured path.

$$T_{char} \approx N_{path} \times (T_{freq} \times N_{freq} \times N_{samples} + T_{reconf}) \quad (1)$$

T_{reconf} can be 5–200 ms for the Cyclone III [2]. As a result, characterization time, T_{char}, for even a small FPGA could extend to days. Both Wong [10] and Gojman [7] suggest that it might be valuable to perform delay experiments in parallel. Parallel testing could divide the characterization time by the number of simultaneously placed and activated CUTs. Parallelism can scale with the size of the chip such that characterization time need not increase with chip capacity. However, Gojman does not perform measurements in parallel citing the possibility that the experiments could affect each other. Nonetheless, both Wong and Gojman build configurations with multiple CUTs instantiated in each configuration in order to reduce the total number of configurations they must generate and, consequently, the number of times the FPGA must be reconfigured; they do this even when they only enable one CUT at a time.

This prior work left open two important questions:

1. Can we run simultaneous delay measurements without significantly corrupting the measured results? That

*Now affiliated with Google, Inc.

FPGA'16 February 21-23, 2016, Monterey, CA, USA
ACM ISBN 978-1-4503-3856-1/16/02.
DOI: http://dx.doi.org/10.1145/2847263.2847334

is, how much does simultaneous delay measurement impact the measured results?
2. Does placing multiple CUTs onto an FPGA in a single configuration have an impact on the measured results?

This paper provides a direct answer to these questions, quantifying the impact of simultaneous measurement and placement of multiple measurement circuits on an FPGA. We show which effects are present and characterize their magnitude for the 65 nm Cyclone III FPGA used by both Gojman and Wong. The paper also identifies potential sources for the effects and a strategy for minimizing the impact of simultaneous measurements.

2. METHODOLOGY

A simplified version of the measurement circuit is shown in Fig. 1. The measured path—labeled the CUT—contains six LUTs configured as buffers and using only the C and D inputs linked with LAB Local Tracks. Surrounding each CUT are registers that latch on alternate edges of the clock: the launch register at the front latches on the positive edge, and the capture register latches at the falling edge. The input signal is an oscillator running in phase with the clock at half the frequency. If it has had sufficient time to propagate through the CUT after half a clock period, then the two registers will have equal values. If so, we know that the propagation delay over the CUT is equal to or less than half the clock period. However, if the clock period is shorter than the CUT delay, the input will not be able to propagate through the CUT, the outputs of the two registers will differ on the falling edge of the clock, and the AND gate in the Error Detector Circuit will register the error. This will trigger an increment of the Error Counter. For each frequency in our experiments, the input and output of the CUT are compared for $N_{samples} = 2^{15}$ transitions. A failure is reported if at least half of the comparisons are mismatches. We measure timing at this 50% failure point since that is where the results will be most statistically significant. To support this, we use a 14 bit counter. When the count exceeds 2^{14}—half of the number of comparisons we run—it indicates that the CUT has failed at this frequency. In our experiments, we took measurements at both the rising and falling edges of the input clock signal and observed similar effects. For simplicity and brevity, we present only the effects seen at the falling edges of the clock.

Differences in placement and routing of resources could potentially impact delays. Consequently, we took care to control the exact wires and switches used in our experiments following the methodology from [7]. The elements in the measurement circuit, including the CUT, are placed in the same positions, and they use identical switches and wire tracks to connect them. The control structures (shown on the left in Fig. 3) are placed optimally to shorten the connections to the measurement circuits, and the routing from the measurement circuits to the control structures depend on the placement of the CUT; nonetheless, the routes from the control structures are always the same when a CUT is placed in the same position on the array. Boundary registers isolate the routes between the CUTs and the control circuitry and provide fixed locations for these routes. To the right of these registers, the routing is strictly controlled to ensure consistency between measurements and reduce potential crosstalk between wires. We use QUIP to extract and control placement and routing [1].

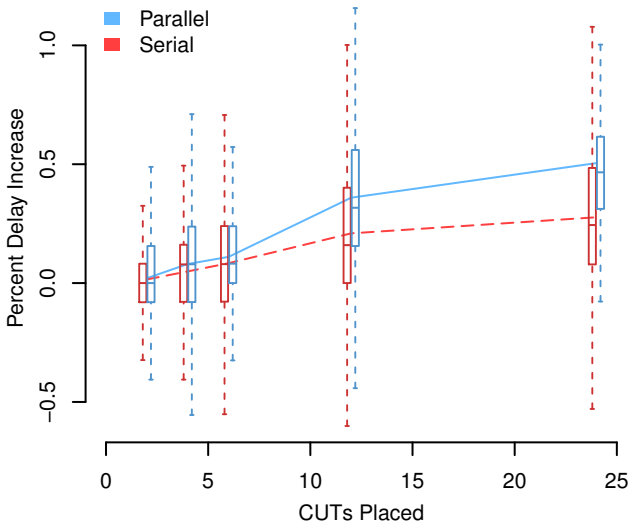

Figure 2: Impact of Simultaneous CUT Placement and Activation on Measured Delay. Lines added to highlight the mean values.

Each tests starts with a binary search to identify the bounds on the circuit operating frequency. Once the binary search has found an approximate frequency, the controller uses a linear search, decreasing the clock period by 1.6 ps until the frequency of failure is found. Five sets of the 2^{15} comparisons are taken from each CUT, and only the result of the last set is used. This allows the circuit to reach a steady-state temperature where the effects of self-heating can be measured uniformly. Previous work established that this methodology gave consistent measurements that were repeatable and independent of the order in which the CUTs were measured [6].

3. SIMULTANEOUS MEASUREMENT

With this setup, we performed an experiment where we first measured a single path in isolation using a single CUT and then measured it with additional CUTs placed at least two LABs away. As noted, the placement and routing for the reference CUT are constrained to be identical across the tests. In the isolation case, aside from control structures (shown on the left of Fig. 3), only a single CUT is placed on the FPGA. For the multiple CUT cases, sets of 2, 4, 6, 12, and 24 CUTs were placed on the FPGA and distributed over at least two rows. Separate measurements were taken for the cases where only one CUT was activated and measured at a time (serial) and cases where all CUTs were simultaneously active and conducting measurements (parallel). An activated CUT will toggle its input to generate a series of transitions that propagate through the path, and, possibly, toggle error counters, while the input to a non-activated CUT does not switch. We collected data across 14 Cyclone III (EP3C16F256C8N) components on Arrow BeMicro boards. We measured CUT delays between 1.911 ns and 2.110 ns.

Fig. 2 shows the percentage increase of the non-isolated measurement from the isolated measurement. The top (solid blue) line shows the mean percent increase in the parallel case where the other CUTs were simultaneously activated,

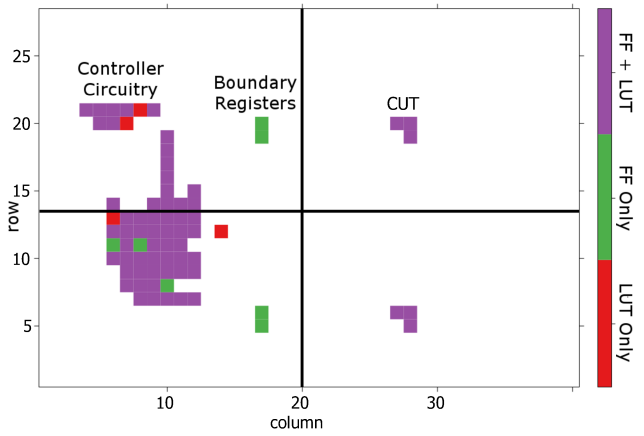

Figure 3: Measurement Test Setup on Cyclone III EP3C16F256C8N

while the bottom (red dashed) line shows the percent increase in the serial case where the added CUTs are not enabled, leaving only the reference CUT active. Both curves include boxplots that show the distribution of the values measured across the various multiple CUT cases; the boxes show the middle two quartiles, while the whiskers show the full range of values measured except for outliers. From these data, we can see that placing additional CUTs increases mean delay. Comparing the serial and parallel trend lines, we can see that more than half of the delay increase comes from the placement of the CUTs, the serial case, rather than their activation. We also see that the effect is not uniform, with some delays decreasing and the largest non-activated delays being as high as the simultaneous activation delays. We saw one, repeatable outlier on one chip whose delay changed by 124 ps.

Depending on the intended use of the measurements, these results may be encouraging or disappointing. The fact that they are within ±1.5%, says that the simultaneous measurements do not change the delays significantly. However, the fact that the delays can change by ±33 ps, even when only one CUT is activated at a time, means the precision of the measurements made is far worse than the precision expected from the clock resolution of 1.3 ps claimed by [10] and 1.6 ps used here and claimed by [7]. Furthermore, the in-LAB LUT chain measurement spreads in [10] are less than 80 ps, and, when measured at the nominal voltage, spreads in [7] are less than 100 ps. This means the delay contribution from simultaneous placement could be 60–80% of these measured delay spreads.

4. WHAT'S HAPPENING?

The CUT used for testing carefully isolates a logic path as a single pipeline stage between registers (Fig. 1). Between the isolated and simultaneous tests, no change is made to the measured CUT logic or physical layout. The only change is that additional, disconnected logic is added elsewhere on the chip. What might cause the measured delay differences?

4.1 Voltage Fluctuation and Self-Heating

Even though the circuit and layout do not change, the delay of the individual circuit elements that make up the CUT may be impacted by environmental changes, including the local temperature and supply voltage for the circuit as demonstrated in previous work [12]. For MOS gates, the switching time is roughly:

$$\tau \approx \frac{C_{load} V_{dd}}{I_{ds}} \tag{2}$$

Due to voltage drops on the power distribution lines that supply individual gates on the chip, the local V_{dd} seen by a gate will be smaller than the package V_{dd} and will vary based on the current draw of other gates sharing portions of the power distribution network ($I \times R$ voltage drop). The on-transistor source-drain current, I_{ds}, is impacted by the local supply voltage, V_{dd}, as well as, the threshold voltage and mobility (and hence saturation velocity, v_{sat}) of the MOSFET, both of which are temperature dependent.

$$I_{ds} = W v_{sat} C_{ox} \left(V_{dd} - V_{th} - \frac{V_{d,sat}}{2} \right)^{\gamma} \tag{3}$$

As circuitry on the FPGA switches, it dissipates energy (e.g., CV^2 switching energy) as heat. The energy dissipation will heat the die area in the vicinity of the switching, increasing the temperature seen by surrounding circuits and changing their current flow and hence speed. For example, Zick and Hayes show 2.5% change in frequency based on temperature and the ability to control local temperature by controlling switching activity [12]. Consequently, switching activity on the chip can potentially impact the delay of a measured CUT. As expected, the impact increases with the total volume of activity, as Fig. 2 shows. Nonetheless, we might be surprised to see that the magnitude of the effect can be equally large even when circuitry is placed but not activated.

4.2 Configurable Clock Architecture

Modern FPGAs have configurable clock networks that allow portions of the clock distribution tree to be deactivated when not in use in order to save power. Placing a clocked circuit that *could* be activated demands that the clock network be configured to deliver a clock signal to the flip-flops on the circuit. The mere presence of a clocked circuit creates activity in the clock network, even when the circuit is not activated. Furthermore, the buffers in the clock distribution network are typically large in order to drive large clock loads and minimize delay and skew on the clock network, likely much larger than the buffers on logic in a LAB. We believe the activation of different portions of the clock network contributes significant activity that impacts circuit delay in the serial cases where additional CUTs are placed but not simultaneously activated with the measured CUT. Differential loading on clock network in the different clock configurations could also be a contributing factor.

In particular, the Cyclone III clock architecture provides independent control of the clock supplied to each quadrant of the chip, including the ability to disable the clocks to a quadrant [2]. Our measurements suggest that row clock drivers can also be independently disabled on the Cyclone III, similar to more recent Altera architectures [3, 4]. We were able to confirm this conjecture with an Altera architect [5].

5. IMPACT OF SECOND CUT

To better characterize the effects of self-heating on measured CUT delay, we performed a controlled experiment with a single second CUT. A CUT placed in the LAB at

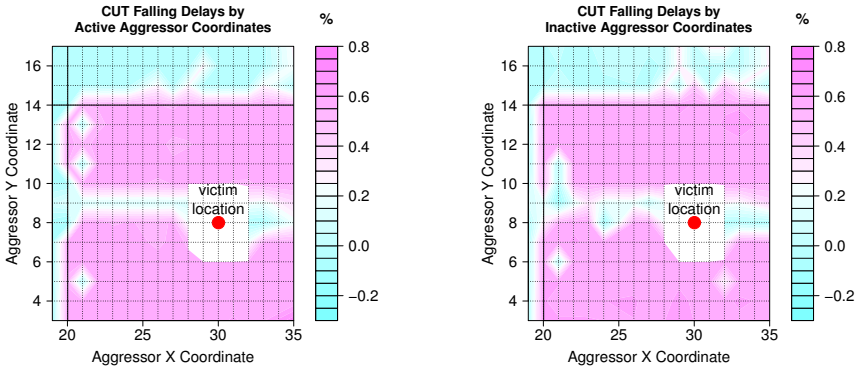

Figure 4: Impact of Relative Position of Two CUTs

(30,8) was used for measurement. It was measured in isolation for a reference point. It was then tested with a second CUT placed at a range of positions around it, both within and outside of its clock quadrant. We call the measurement CUT the *victim* because we are measuring how it is affected by the second CUT, which we call the *aggressor*. The test was conducted once with the aggressor CUT inactive and once with it active (Fig. 4).

Each coordinate in the Fig. 4 map is the location of the aggressor CUT. The value associated with the coordinate is the percent delay increase of the victim CUT with respect to its measurement in isolation. No trials were run where the aggressor was placed in the 3×3 region surrounding the victim. Because the measurement circuits occupy three LABs on the Cyclone III, we left a buffer zone so that the victim and aggressor would never overlap.

The quadrant where the victim CUT is located has its boundaries on x=20 and y=14. When the aggressor CUT is placed outside these boundaries, the victim runs faster—usually close to the delay of the isolated case. When the aggressor is placed in the same quadrant, the victim delays increase by up to 0.74%. Across chips and trials, we see the same pattern—placing the aggressor in the same quadrant increases the victim's delays more significantly than placing it in a different quadrant. This suggests it may be possible to place and measure CUTs simultaneously as long as they reside in separate quadrants.

When the aggressor is placed in the same, or a nearby, row as the victim, the aggressor typically has less impact on the victim. It is possible that the two CUTs share an enabled row clock driver in this configuration, so there is no additional heat or activity generated by activating an *additional* row clock driver within the quadrant. If this effect is robust, it could suggest another option for obtaining low-noise parallel measurements in the same configuration.

6. QUADRANT EXPERIMENT

The previous section suggests there is a strong impact on timing when two CUTs are placed in the same quadrant, but a much smaller effect when they are placed in different quadrants. To further understand this effect, we provide a more directed quadrant experiment. Since the control circuitry lives in the left-hand quadrants (Fig. 3), we limited this experiment to the upper-right and lower-right quadrants. For the same-quadrant experiment, pairs of CUTs were placed in the lower-right quadrant seven rows apart. For the different-quadrant experiment, pairs of CUTs were placed with one CUT in the lower-right quadrant of the chip and the second placed in a similar position in the same column in the upper-right quadrant as illustrated in Fig. 3.

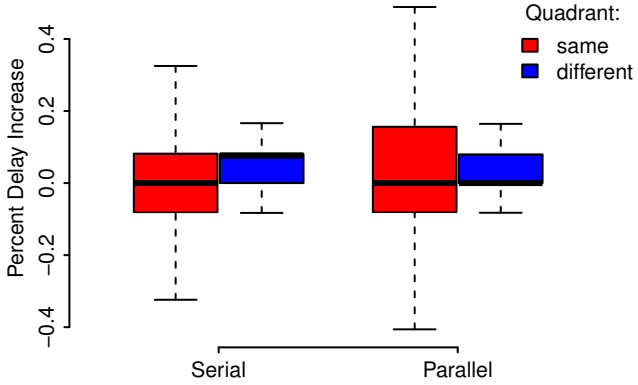

Figure 5: Comparing Impact of Same Quadrant vs. Different Quadrant Aggressors

When the simultaneously activated CUTs (parallel) are in different quadrants, the difference between simultaneous and isolated measurements is highly concentrated at zero, much more so than when the CUTs are in the same quadrant (Fig. 5). There are still a small percentage of cases where the same-quadrant measurements differ by as much as the different-quadrant measurements as both have occurrences around ±1.5%. Leaving the aggressor CUT inactive (serial) produces similar distributions, also with outliers around ±1.5%.

To determine whether the effect was systematic or unique to particular chips, we differentiated the data based on the measured chip (Fig. 6), using a different symbol for each chip. Some chips are faster than others, as we expect from die-to-die variation. However, no small subset of the chips is uniquely to blame for outliers, nor do they show a tendency to produce results shifted up or down on the graph.

7. DISCUSSION AND FUTURE WORK

The presence of logic, even inactive logic, does impact timing results in on-chip measurements experiments. Vendors know this, and it is one of the timing margins included in their timing analysis [7]. It is necessary to understand how

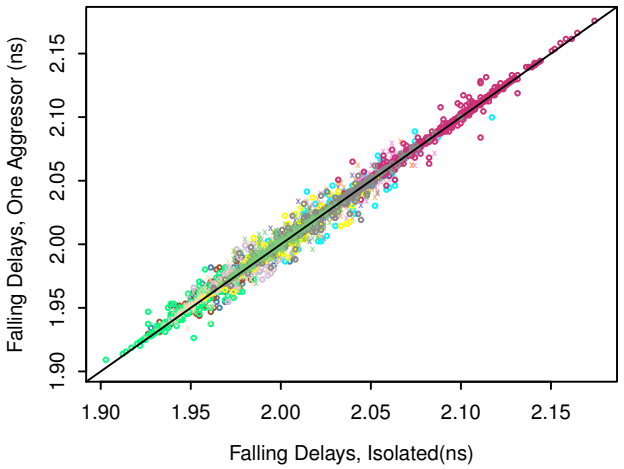

Figure 6: Isolated vs. Simultaneous, Two-CUT, Out-of-Quadrant Measurements by Chip

this simultaneous logic can affect the measurements in order to interpret the results of on-chip timing measurements. The non-isolated measurements are not necessarily wrong—in typical usage scenarios much of the nearby logic will be used, so large portions of the clock network will be enabled as well. When the activity effects are relatively uniform, such that there is a good correlation between component path delay measurements and application-circuit delays, the measurements can still be useful in identifying the relative delay of the resources. *The pitfall comes in comparing the delay of two resources that are differentially impacted by their environment.* As we see in Figs. 2, 5, and 6, while many measured delays change little, some change by much larger amounts, and this differential effect can be misleading. When the magnitude of the difference exceeds the intrinsic delay differences, the measurements can mislead CAD tools (e.g., [8]) and variation characterization.

The primary contribution of this short paper is to identify this issue and provide an initial characterization of the magnitude of the effect. As such, it raises a host of questions that will need to be addressed in future work. For example, how significant are these effects on other FPGA models with different clock architectures? Ideally, we would like to develop a timing model that accounts for thermal and other coupling effects, including modeling the effects of infrastructure logic such as the clock network. E.g.,

$$\tau_{use}(A,B) = \tau_{int}(A,B) + \sum_x \sum_y f(x,y,act(x,y),T_{ext}) \quad (4)$$

This would allow CAD tools to account for coupling delay effects directly. At least, it is necessary to develop a better understanding of the effects upon on-chip delay measurements and develop best practices for collecting delays that are useful and predictive for CAD. It will also be useful to understand how much the coupling effects themselves are subject to variation at various scales and over time.

8. CONCLUSIONS

Even nominally disjoint and quiescent logic placed in proximity to circuitry configured on an FPGA can impact its delay. On a 65 nm Cyclone III we identified effects as high as ±33 ps or about 1.5% of the delay of the paths we were measuring. The effects of the quiescent logic can be explained in terms of circuit activity when we account for the configurable clock network. We show that the average delay effect can be reduced by keeping simultaneous logic in different quadrants or, perhaps, in the same row within a quadrant. Our preliminary experiments suggest there is a rich area to explore to characterize the nature of these coupling effects.

9. ACKNOWLEDGMENTS

Timothy A. Linscott was supported by the National Science Foundation (NSF) Research Experience for Undergraduates (REU) program under contract EEC-1359107. Parts of the research were supported by DARPA/CMO contract HR0011-13-C-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the official policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government.

10. REFERENCES

[1] Altera. QUIP. http://www.altera.com/education/univ/research/quip/unv-quip.html, 2005. 2

[2] Altera. Cyclone III Device Handbook Volume I. http://www.altera.com/literature/hb/cyc3/cyclone3_handbook.pdf, 2011. 1, 4.2

[3] Altera. White paper 01148-2.0: Reducing power consumption and increasing bandwidth on 28-nm FPGAs. https://www.altera.com/en_US/pdfs/literature/wp/wp-01148-stxv-power-consumption.pdf, March 2012. 4.2

[4] Altera. Cyclone V Device Handbook. https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/hb/cyclone-v/cv_5v2.pdf, 2015. 4.2

[5] V. Betz. Cyclone 3 configurable clock architecture. *personal communications*, December 2015. 4.2

[6] B. Gojman. *GROK-FPGA: Generating Real On-chip Knowledge for FPGA Fine-Grain Delays using Timing Extraction*. PhD thesis, University of Pennsylvania, 2014. 1, 2

[7] B. Gojman, S. Nalmela, N. Mehta, N. Howarth, and A. DeHon. GROK-LAB: Generating real on-chip knowledge for intra-cluster delays using timing extraction. *ACM Tr. Reconfig. Tech. and Sys.*, 7(4):5:1–5:23, Dec. 2014. 1, 1, 2, 3, 7

[8] N. Mehta, R. Rubin, and A. DeHon. Limit Study of Energy & Delay Benefits of Component-Specific Routing. In *FPGA*, pages 97–106, 2012. 7

[9] T. Tuan, A. Lesea, C. Kingsley, and S. Trimberger. Analysis of within-die process variation in 65nm FPGAs. In *ISQED*, pages 1–5, March 2011. 1

[10] J. S. Wong, P. Sedcole, and P. Y. K. Cheung. Self-measurement of combinatorial circuit delays in FPGAs. *ACM Tr. Reconfig. Tech. and Sys.*, 2(2):1–22, June 2009. 1, 1, 3

[11] H. Yu, Q. Xu, and P. H. Leong. Fine-grained characterization of process variation in FPGAs. In *ICFPT*, pages 138–145, 2010. 1

[12] K. M. Zick and J. P. Hayes. On-line sensing for healthier FPGA systems. In *FPGA*, pages 239–248, 2010. 1, 4.1, 4.1