Within-Die Delay Variation Measurement and Power Transient Analysis Using REBEL

Fareena Saqib, Dylan Ismari, Charles Lamech, and Jim Plusquellic

***Abstract*— Variations in path delays are increasing with scaling, and are increasingly affected by neighborhood interactions. To fully characterize within-die variations, delays must be measured in the context of actual core-logic macros using embedded test structures (ETSs). In this brief, we propose an ETS called regional delay behavior (REBEL) that is designed to measure path delays in a minimally invasive fashion. REBEL provides capabilities similar to an off-chip logic analyzer. We implemented REBEL in a 90-nm test chip and present results on within-die path delay variations in a floating point unit. We also analyze the impact on delay when the chips are subjected to industrial-level temperature and voltage variations.**

***Index Terms*— Embedded test structure (ETS), path delay, process variations (PVs).**

I. Introduction

Both random and systematic within-die process variations (PV) are growing more severe with shrinking geometries and increasing die size. Embedded test structures (ETSs) continue to play an important role in the development of models of PVs and as a mechanism to improve correlations between hardware and models. Their availability can be used to improve yield, and the corresponding profitability and product quality of the fabricated ICs [1]. Unfortunately, accurate models of within-die PVs are becoming more difficult to derive because of their increasing sensitivity to design-context. Stand-alone ETSs, such as ring-oscillators (ROs) are becoming less effective for characterizing delay variations because they are typically placed around the layout region of the macro and are not exposed to, e.g., the same types of distortions which are introduced by photolithography interference patterns. More recently proposed ETS, such as those that measure delay characteristics of the macro itself [3], offer the best solution, but are challenging to integrate without having an adverse impact on area overhead, yield loss, performance, I/O interface, test cost, and so on, of the product design.

In this brief, we investigate an ETS, called regional delay behavior (REBEL), which is designed to measure path delays in macros while minimizing these types of adverse effects. The proposed ETS is designed to serve applications, such as model-to-hardware correlation [4], detection of hardware Trojans, design debug processes, detection of small delay defects, and physical unclonable functions. Each of these areas requires accurate measurements of path delays and/or the ability to differentiate at high resolutions between delays of neighboring paths.

The REBEL ETS leverages the scan chain architecture to measure delay variations. In particular, it uses a special configuration of flush delay mode that is available in Level Sensitive Scan Design (LSSD)-style scan chains. In [5], we demonstrated the premise of capturing regional delay variations using a special launch–capture (LC) timing sequence applied while in flush delay mode. We extend this technique here by allowing output signals from a design macro to be inserted into the flush delay chain for path delay measurements.

A key feature of the work presented in this brief is the evaluation of REBEL in multiple copies of a custom designed test chip fabricated in IBM’s 90-nm technology. The macro in which REBEL is integrated is an IEEE-754 compliant floating point unit (FPU), with five pipeline stages. Random test patterns are applied to the combinational logic within each of the pipeline stages and the measured delays are analyzed, with emphasis on evaluating the magnitude of within-die variations as a function of path length. A second important component of our experiments is the evaluation of delay variations while the chips are subjected to industrial-level temperature and voltage (TV) variations.

Our results show that the magnitude of within-die delay variations is dependent on the length of the path and TV conditions, and can vary up to 35%. A second result is the high sensitivity of delay to the power transient effect introduced by the LC clock event. The impact of the power transient is particularly evident when the LC interval is dynamically changed as a means of obtaining high resolution delay measurements. A calibration process for applications in Design for Manufacturability area is proposed that is designed to eliminate this environmental source of delay variation.

II. Background

Within-die and die-to-die delay analysis of delay variations continues to be an active research area. RO-based test structures have been successfully used to characterize within-die delay variations in application-specified integrated circuits and field-programmable gate arrays (FPGAs) [2], [3].

Stand-alone RO-based delay measurements lack the ability to account for circuit context. Macro embedded RO schemes, such as path RO [3], can only be applied to hazard free robust paths, which restricts their coverage. Analog measurement systems and time-to-digital convertors (TDCs) have large area overheads.

Several on-chip analog measurement systems are proposed in [6], wherein the authors characterize rising and falling delay variations on the I/O waveforms of individual gate using an on-chip sampling oscilloscope. Various TDCs have been proposed for on-chip delay measurements [7] with resolution as high as 5 ps and with low thermal sensitivity. In [8], a within-die variation characterization system is proposed which uses an on-chip sampling oscilloscope. An LSSD-style scan chain-based ETS is proposed in [4] and [5] for obtaining single-shot measurements of path delays in product macros. This brief investigates this scheme (REBEL) on a set of 90-nm test chips.

III. Overview of ETS REBEL

In this section, we describe the modifications needed to integrate REBEL into a clocked-LSSD-style scan architecture. The macro-under-test (MUT) in Fig. 1 is the combinational logic from a core logic macro. A row of scan flip-flops (FFs) is shown along the top which serve to launch transitions into the MUT. The bottom row is used to capture transitions that propagate through the MUT.

REBEL ETS components include row control logic (RCL) and front-end-logic. Transitions can be launched into the MUT using standard manufacturing delay test strategies such as launch-OFF-capture and launch-OFF-shift. In either of these two scenarios, the scan chain is loaded with the initial pattern of the two-pattern test and the system clk (Clk) is asserted to generate transitions in the MUT by capturing the output of a previous block or by doing a 1-bit shift of the scan chain, respectively.

Manuscript received June 28, 2013; revised February 5, 2014; accepted March 25, 2014.

F. Saqib, D. Ismari, and J. Plusquellic are with the University of New Mexico, Albuquerque, NM 87131 USA (e-mail: fareenasaqib1@gmail.com; dismari@unm.edu; jimp@ece.unm.edu).

C. Lamech is with Intel Corporation, Santa Clara, CA 95054 USA (e-mail: charleslamech@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2014.2318307

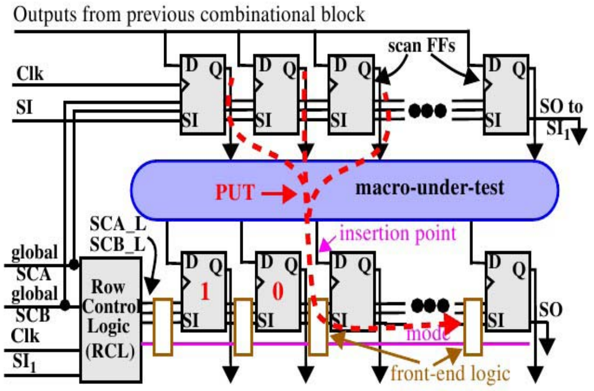


Fig. 1. REBEL integration strategy.

The transitions that propagate through the MUT emerge on some of its outputs. REBEL allows only one of these transitions from a path-under-test (PUT) to be measured at a time in the MUT as indicated by insertion point in the figure. The RCL and front-end-logic blocks convert all scan FFs to the right of the insertion point FF into a delay chain. A digital snapshot of the signal as it propagates along the delay chain can be obtained by deasserting Clk. The digital snapshot is used to obtain timing information about the PUT, and to determine if glitches occurred. Implementation details on REBEL can be found in [4].

IV. Experimental Setup, Results, and Analysis

A. Experimental Setup

A FPU fabricated in IBM's 90-nm technology is used as a test vehicle in chip experiments carried out at nine different TV corners. The FPU design has a five-stage pipeline structure, labeled P1 through P5, with MUXes, decoders, adder/subtractors, a multiplier, and so on, inserted between the pipeline registers, all wired together to form a scan chain with input SI_1 (Fig. 2). REBEL is integrated in each pipelined stage as REBEL rows labeled RR_x from 1 to 28. REBEL increases the scan overhead of Clocked Level Sensitive Scan Design by 61%, and adds one fanout load to the master latch, which has a very small impact on power and performance. The area of REBEL integrated FPU is 252 k $\mu\mathrm{m}^2$, and chip level overhead of the REBEL integration is 11.45%.

REBEL testing is carried out in four basic configurations. In the first two configurations, Cfg1 and Cfg2, the REBEL rows in pipeline stages P0, P1, and P3 are configured in functional mode while those in P2, P4, and P5 are configured in REBEL mode, as shown in Fig. 1. In configurations Cfg3 and Cfg4, the rows in P0, P2, and P4 are configured in functional mode while those in P1, P3, and P5 are configured in REBEL mode. These four configurations collectively allow paths in all of the logic blocks to be tested using the REBEL ETS.

B. LC Clocking Sequence and Clock Strobing

The RCL and front-end logic for REBEL allow critical timing events, i.e., the LC interval (LCI), to be controlled by the system clock. This is illustrated by the timing diagram shown in Fig. 3.

A scan operation is first carried out that introduces a random test and configuration data into the REBEL rows. The LCI test consists of asserting the Clk signal, which launches transitions in the combinational logic, and deasserting Clk a fixed Δt later, which halts all signals propagating along the delay chains. The delay in a combinational path can be computed using

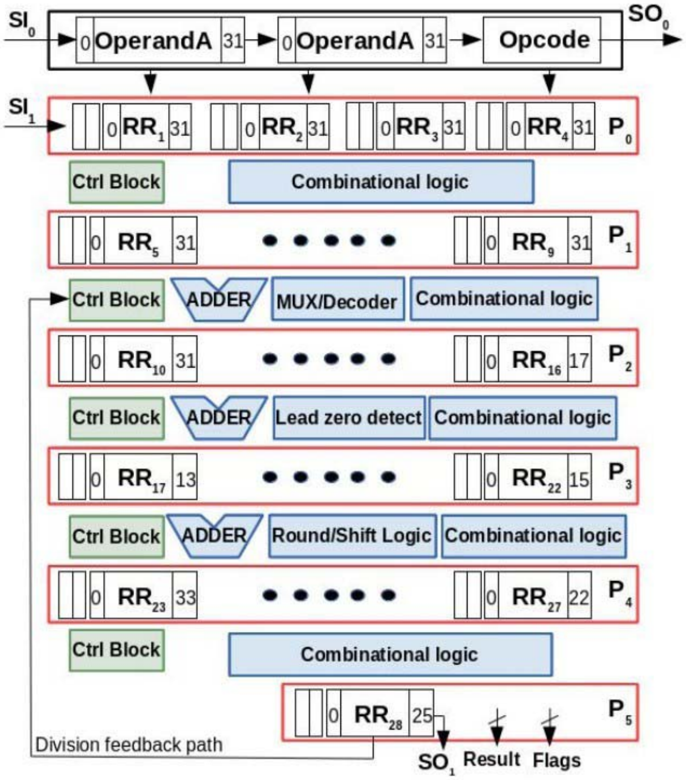


Fig. 2. Product MUT.

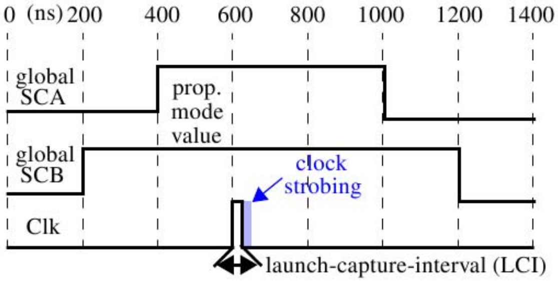

0 (ns) 200 400 600 800 1000 1200 1400
global SCA
prop. mode value
global SCB
clock strobing
Clk
launch-capture-interval (LCI)

Fig. 3. REBEL LC test sequence. Clock strobing applies a sequence of LCIs of different widths.

$$T_{\mathrm{path}} = T_{\mathrm{lc}} - T_{\mathrm{dc}} \tag{1}$$

where

T_{path} delay of the combinational path;
T_{lc} LCI delay;
T_{dc} delay through the delay chain.

The LC clock sequence is generated using the fine phase adjust (FPA) feature of a digital clock manager (DCM) on a Virtex-6 FPGA. The actual LCI is somewhat different than the programmed value because of the FPGA internal logic that creates the pulse from the DCM output. Fig. 4 shows the programmed FPA on the x-axis against the actual LCI produced by the FPGA (we round all delays to the nearest 5 ps value). In our experiments, we apply a sequence of LCI tests over the range of FPAs between 128 and 444 in FPA increments of two. This results in the application of $(444 - 128)/2 + 1 = 159$ LCI tests with actual LCIs between 2745 and 8400 ps, as given by the oscilloscope curve in Fig. 4. All references to FPAs are translated to actual LCIs using this curve.

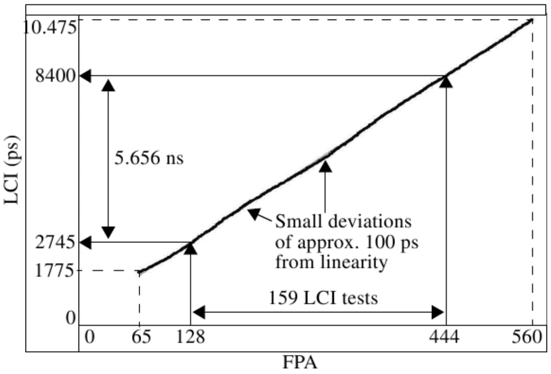
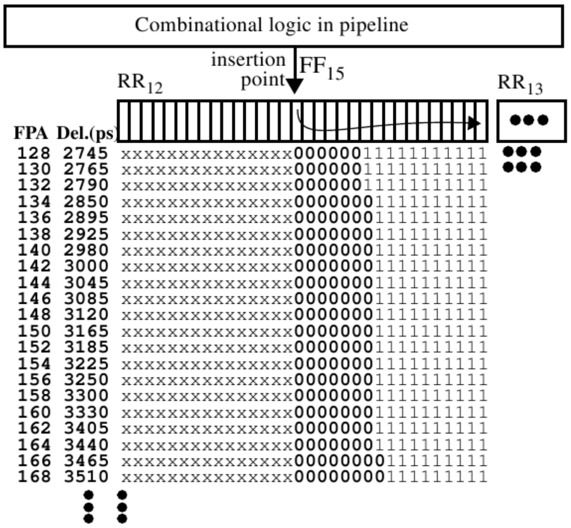


Fig. 4. Oscilloscope measured LCIs for each of the FPA values on the FPGA.



```
FPA Del.(ps)
128 2745 xxxxxxxxxxxxxxx00000011111111111 ●●●
130 2765 xxxxxxxxxxxxxxx00000011111111111 ●●●
132 2790 xxxxxxxxxxxxxxx00000011111111111
134 2850 xxxxxxxxxxxxxxx00000001111111111
136 2895 xxxxxxxxxxxxxxx00000001111111111
138 2925 xxxxxxxxxxxxxxx00000001111111111
140 2980 xxxxxxxxxxxxxxx00000001111111111
142 3000 xxxxxxxxxxxxxxx00000001111111111
144 3045 xxxxxxxxxxxxxxx00000001111111111
146 3085 xxxxxxxxxxxxxxx00000001111111111
148 3120 xxxxxxxxxxxxxxx00000001111111111
150 3165 xxxxxxxxxxxxxxx00000001111111111
152 3185 xxxxxxxxxxxxxxx00000001111111111
154 3225 xxxxxxxxxxxxxxx00000001111111111
156 3250 xxxxxxxxxxxxxxx00000001111111111
158 3300 xxxxxxxxxxxxxxx00000001111111111
160 3330 xxxxxxxxxxxxxxx00000001111111111
162 3405 xxxxxxxxxxxxxxx00000001111111111
164 3440 xxxxxxxxxxxxxxx00000001111111111
166 3465 xxxxxxxxxxxxxxx00000000111111111
168 3510 xxxxxxxxxxxxxxx00000000111111111
 ⋮    ⋮
```

Fig. 5. Partial digital snapshot of LCIs for one path from a path delay test.

C. Calibration and Power Rail Voltage Transient Effects

A large fraction of the paths tested using our random test patterns are shorter than the Δt associated with the smallest applied LCI. Although it is possible to use smaller LCIs to test these paths, thereby eliminating the delay chain elements, doing so requires testing the chip with a faster-than-at-speed clock sequence. It is well known that applying faster-than-at-speed tests results in inaccurate delay measurements because of power supply voltage transient effects introduced by two closely placed LC edges. REBEL allows accurate timing information to be obtained for these short paths without using faster-than-at-speed LC tests. However, to do so, a mechanism is needed to eliminate the delay chain components.

The digital snapshot captured in the delay chain is a string of binary bits, one string for each of the 159 LCI tests applied to test a path. Fig. 5 shows the digital snapshots for the first 21 LCI tests of a path in a vertical sequence. The insertion point in this example is FF_{15} of the REBEL row RR_{12} under Cfg_1 from Fig. 2. For FPA 128, a falling edge propagated along six elements of delay chain, i.e., through FF_{15} through FF_{20}, before being halted by the capture event.

In each subsequent snap-shot up through FPA 132, the edge continues to propagate through FF_{20} but fails to reach FF_{21} until FPA 134 is applied. The falling edge requires 33 more FPAs, i.e., 134 through 166, to propagate completely through FF_{21}. From these snapshots, it is possible to derive the approximate delay through FF_{21} as $(3465 - 2850) = 615$ ps. A similar process is used to measure the delay through the other FFs. Using eight random vectors a typical chip has approximately 825 stable paths. Each stable path allows, on average, the delays of 12 FFs to be estimated. Therefore, nearly 10 000 FF delays can be derived from the test data for each chip.

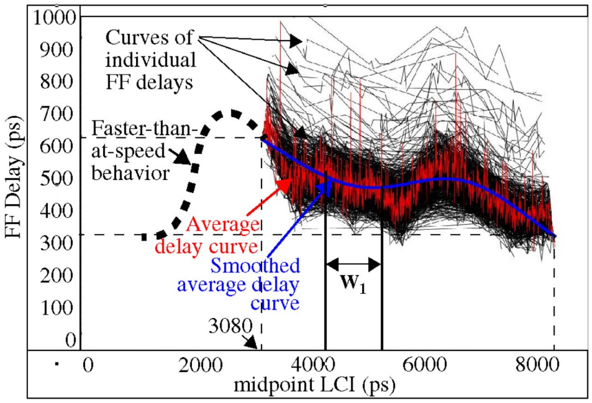


Fig. 6. Curves of individual FF delays measured at different midpoint delays. An average of all individual curves is superimposed, as well as a curve with only the low frequency components of the average delay curve.

Unfortunately, the delay through each FF varies as a function of the applied LCI. This occurs because of the power supply voltage transient that is produced by the clock strobe events. Fig. 6 shows the measured FF delays for a chip as a function of the applied LCI plotted along the x-axis. The vertical dispersion of the individual curves is caused by process variations among the FFs, i.e., the curves shown along the top of the figure belong to slower FFs. It is a remarkable fact that the delay through any given FF varies by as much as 2X over the range of applied FPAs. In particular, the range delineated by the dotted lines illustrate that the average delay changes from approximately 300 to 600 ps. The figure includes a smoothed average delay curve to show the general trend among all individual curves, which is obtained by eliminating the high order frequency components of average delay curve (also shown). The impact of the power transient is even more dramatic for LCIs less than the smallest one used in our experiments, as illustrated by the region labeled faster-than-at-speed behavior on the left side in the figure. We purposely avoid the region below approximately 4000 ps because of the large distortion in path and FF delays.

We use the digital snapshots to compute the FF delays and then assign this delay to the LCI which represents the midpoint between the FPAs used to time it. As an example, the delay 615 ps computed using FPAs 134 and 166 in reference to Fig. 5 is assigned a midpoint FPA of $(134 + 166)/2 = 150$. The measured FF delay of 615 ps is the assigned delay through this FF when the LCI used is 3165 ps (delay at 150 from Fig. 5).

D. Measuring and Calibrating Path Delays

From the analysis of FF delays, it follows that the delay along the combinational logic paths is also a function of the FPA. Therefore, to properly capture and analyze the variations which occur along paths within a combinational logic block, we limit the LCIs considered valid for path delay testing to a small region or window between 4355 and 5250 ps (approximately 1 ns), which is delineated and labeled as W_1 in Fig. 6. The FF delays in this window remain relatively constant, within 25 ps, and therefore, so will the path delays.

The digital snapshots from this window are parsed in reverse order starting with the digital snapshot associated with largest LCI (5250 ps). Parsing in reverse order ensures the last transition is used as the path delay in cases where there is glitching.

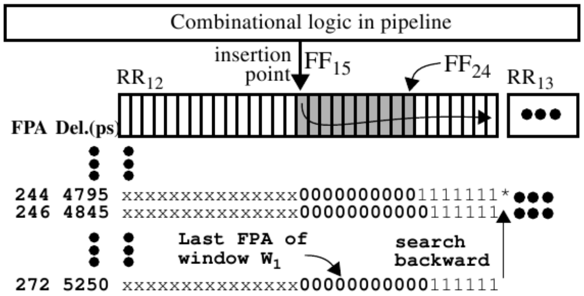

Combinational logic in pipeline
insertion point
FF15
FF24
RR12
RR13
FPA Del.(ps)
244 4795 xxxxxxxxxxxxxxx00000000001111111*
246 4845 xxxxxxxxxxxxxxx00000000000111111
Last FPA of window W1
search backward
272 5250 xxxxxxxxxxxxxxx00000000000111111

Fig. 7. Path delay using digital snapshots.

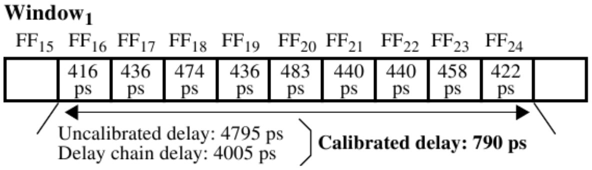

Window1
FF15 FF16 FF17 FF18 FF19 FF20 FF21 FF22 FF23 FF24
416 ps
436 ps
474 ps
436 ps
483 ps
440 ps
440 ps
458 ps
422 ps
Uncalibrated delay: 4795 ps
Delay chain delay: 4005 ps
Calibrated delay: 790 ps

Fig. 8. Delay calibration applied to path using W_1.

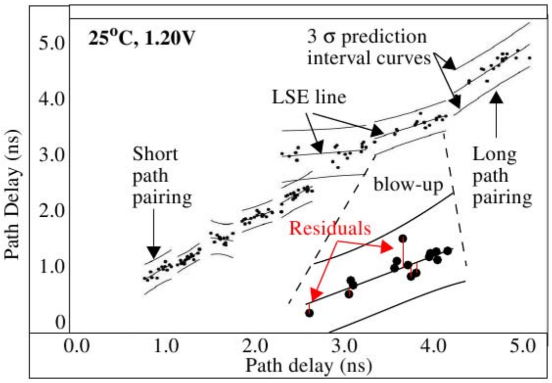


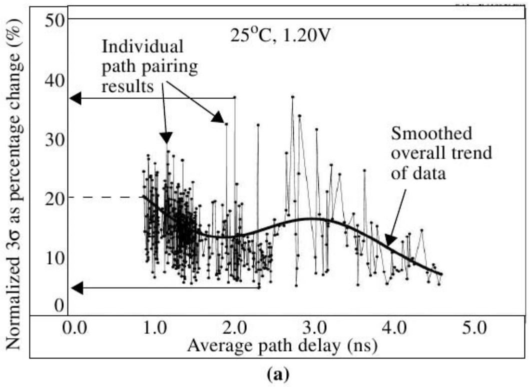
Fig. 9. Within-die delay variation analysis using regression analysis on scatter plots with data from 16 chips.

This process is shown in Fig. 7 using snapshots for the path referenced in Fig. 5 but at larger FPAs. The snapshot at LCI 5250 indicates that a falling edge is propagating through FF_{25}. Our algorithm searches backwards stopping with the snapshot where the edge is just about to enter FF_{25}, which occurs at FPA 244.

The corresponding LCI of 4795 ps gives the uncalibrated delay. The actual path delay is obtained by subtracting the delays for FF_{16} through FF_{24}. The FF delays are computed using the LCI chosen for the path delay, i.e., 4795 ps, and the corresponding FF curves shown in Fig. 6. Fig. 8 shows the computed FF delays obtained from the calibration process for this path, as well as the actual path delay, given as 790 ps.

E. Within-Die Delay Variation Analysis

Regression analysis is an effective technique for measuring and analyzing within-die variations. Linear regression is applied to scatter plots which are constructed from the delays of a group of chips using two separate paths, i.e., a path pairing. Fig. 9 plots eight path pairings in a sequence of eight scatter plots, illustrating variations that occur across the range of short (lower left) and long (upper right) path pairings. Each data point in a given scatter plot represents a pair of path delays from one of the chips. Path pairings are created by sorting the n delays from a reference chip, CHIP_1, and then pairing consecutive paths in the sorted order to create $n-1$ scatter plots. This ensures that the paths of each pairing have similar delays.



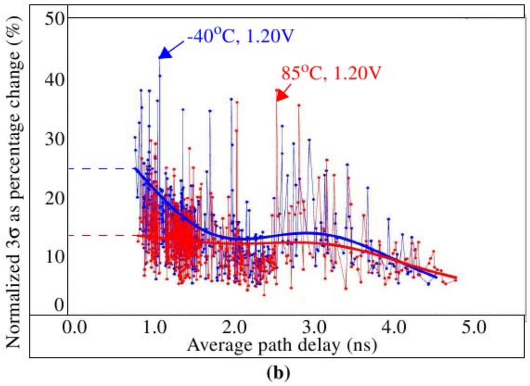


Fig. 10. Within-die variation analysis using regression at (a) 25 °C, 1.20 V and (b) −40 °C and 85 °C, 1.20 V.

Linear regression analysis first computes a least squares estimate (LSE) of a best fit line through the data points of each scatter plot separately (see [9] for defining equations). Several of the eight LSE lines are labeled in Fig. 9. The LSE line tracks chip-to-chip PVs. Within-die variations (and random noise) are represented by the vertical offsets of the data points from the LSE line. The vertical offsets are called residuals (see the blow-up illustration in the figure).

Three σ prediction interval curves are also derived for each scatter plot, and reflect the overall spread of the points around the LSE line. We compute the 3σ of the residuals and then convert this value to percentage change by dividing by the average path delay and multiplying by 100. The average path delay is the mean x-value of all data points from a scatter plot. This metric scales the 3σ according to the length of the path, making it possible to compare within-die variations of short and long paths.

The results obtained by applying regression analysis using data from the nominal TV corner is shown in Fig. 10(a). The average path delay for each of 551 path pairings is given along the x-axis, plotted against the percentage change metric described above. The law of averaging works to keep the variation of longer paths smaller. However, the level of variation per gate is much larger, and is reflected better in the short path results. The most significant deviation in this trend is the peak in the curve for median length paths of length approximately 3 ns. Fig. 10(b) gives the results at −40 °C and 85 °C at nominal voltage. Although similar trends are apparent, higher temperatures tend to reduce the level of variations and noise levels while lower temperatures exacerbate them. For example, the percentage change for the shortest paths at nominal TV are approximately 20%, while those for higher and lower temperatures are approximately 15% and 25%, respectively. The largest variations under nominal conditions exceed 35% for several paths, as shown in the figure.

V. Conclusion

In this brief, we propose an ETS called REBEL suitable for measuring within-die variations in actual product macros. The results show that path delays change according to the LCI used to time them. Therefore, calibration as proposed in this brief must be carried out to obtain an accurate analysis of within-die variations. Within-die variation analysis of test chip data shows long paths change by 5% on average while short paths change by 20% on average, and the magnitude of within-die variations increases as temperature is reduced.

References

[1] M. M. Bashir and L. Milor, “Determining the impact of within-die variation on circuit timing,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 3, pp. 385–391, Aug. 2011.

[2] S. Paul, S. Krishnamurthy, H. Mahmoodi, and S. Bhunia, “Low-overhead design technique for calibration of maximum frequency at multiple operating points,” in *Proc. IEEE/ACM ICCAD*, Nov. 2007, pp. 401–404.

[3] X. Wang, M. Tehranipoor, and R. Datta, “Path-RO: A novel on-chip critical path delay measurement under process variations,” in *Proc. IEEE/ACM ICCAD*, Nov. 2008, pp. 640–646.

[4] C. Lamech, J. Aarestad, J. Plusquellic, R. M. Rad, and K. Agarwal, “REBEL and TDC: Embedded test structures for regional delay measurements,” in *Proc. IEEE/ACM ICCAD*, Nov. 2011, pp. 170–177.

[5] J. Aarestad, C. Lamech, J. Plusquellic, D. Acharyya, and K. Agarwal, “Characterizing within-die and die-to-die delay variations introduced by process variations and SOI history effect,” in *Proc. 48th ACM/EDAC/IEEE DAC*, Jun. 2011, pp. 534–539.

[6] X. Zhang, K. Ishida, H. Fuketa, M. Takamiya, and T. Sakurai, “An on-chip characterizing system for within-die delay variation measurement of individual standard cells in 65-nm CMOS,” in *Proc. Des. ASP-DAC*, 2011, pp. 109–110.

[7] C. C. Chen, P. Chen, C. S. Hwang, and W. Chang, “A precise cyclic CMOS time-to-digital converter with low thermal sensitivity,” *IEEE Trans. Nucl. Sci.*, vol. 52, no. 4, pp. 834–838, Aug. 2005.

[8] Z. Xin, K. Ishida, M. Takamiya, and T. Sakurai, “An on-chip characterizing system for within-die delay variation measurement of individual standard cells in 65-nm CMOS,” in *Proc. 16th ASP-DAC*, Jan. 2011, pp. 109–110.

[9] (2013, Feb. 28). *Regression Analysis* [Online]. Available: http://en.wikipedia.org/wiki/Regression_analysis