### Citation

### As Published
http://dx.doi.org/10.1109/JSSC.2009.2039270

### Publisher
Institute of Electrical and Electronics Engineers (IEEE)

### Version
Final published version

### Accessed
Wed Mar 16 07:40:23 EDT 2016

### Citable Link
http://hdl.handle.net/1721.1/69949

### Terms of Use
Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

### Detailed Terms
The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
All-Digital Circuits for Measurement of Spatial Variation in Digital Circuits

Nigel Drego, Member, IEEE, Anantha Chandrakasan, Fellow, IEEE, and Duane Boning, Fellow, IEEE

Abstract—Increased variation in CMOS processes due to scaling results in greater reliance on accurate variation models in developing circuit methods to mitigate variation. This paper investigates spatial variation in digital circuit performance; we describe a test-chip in 90 nm CMOS containing all-digital measurement circuits capable of extracting accurate variation data. Specifically, we use replicated 64-bit Kogge–Stone adders, ring oscillators (ROs) of varying gate type and stage length and an all-digital, sub-picosecond resolution delay measurement circuit to provide this data. Measurement data from the test-chips indicate that 1) relative variation is significantly larger in low-voltage domains, 2) within-die variation is spatially uncorrelated, and 3) die-to-die (or global) variation is strongly correlated, but degrades toward uncorrelated as the power-supply voltage is lowered. Lastly, extended analysis of the data reveals that systematic effects such as layout pattern dependencies or circuit structure can be misinterpreted as random but spatially-correlated variation. This suggests that circuit designers will reap more benefit from design tools capable of modeling systematic, position-dependent variation rather than spatially correlated, distance-dependent variation.

Index Terms—Delay measurement, digital circuits, spatial correlation, variation.

I. INTRODUCTION

PROCESS variation in advanced technology nodes poses an increasing challenge to both process engineers and circuit designers [1]. To effectively address this challenge, digital designers require variation data and models at familiar levels of abstraction. For example, models that can accurately capture variation at the gate-level can be directly plugged into the relevant tools, such as gate- or block-level timing and power analysis tools. Going a step further, characterization of variation in common digital circuits provides designers the intuition necessary to guide robust design as well as concrete physical data capable of verifying the results of automated tools.

As decomposition of the spatial components of variation increases in importance, understanding circuit performance correlation enables first, more accurate modeling, as evidenced by the emergence of statistical timing models which incorporate spatial correlation [2]–[4] despite not having manufacturing data to validate such models, and second, design of appropriate mitigation techniques. Specifically, if nearby circuits are strongly correlated (|ρ| ≈ 1), techniques involving “replica” circuits can be employed as both monitor circuits (“canaries”) as well as within feedback loops capable of active compensation. However, when circuits are weakly or not at all correlated (|ρ| ≈ 0), “in-situ” techniques are likely necessary, as replica circuits may not accurately mimic the performance of the actual circuits under consideration. Furthermore, correlation data, and in particular any spatial dependencies, enable determination of spacing criteria between monitor circuits and potential critical paths in the design to ensure high correlation.

This work, an extension of [5], provides both circuits capable of quantifying spatial variation and analysis and decomposition of the measured data. First, we outline relevant related work in Section II and continue in Section III by further detailing the all-digital test-chip architecture specifically designed to extract spatial variation data of common digital circuits. Extended data analysis is presented in Section IV, including further comparison of systematic, position-dependent variation versus spatially-correlated, distance-dependent variation. Furthermore, we show how systematic variation components can be misrepresented as spatially-correlated variation if not properly subtracted. Section V provides concluding thoughts and insights relevant to modeling and mitigation of variation in digital circuits.

II. RELATED WORK

Spatial variation in semiconductor processes has been studied for the past two decades, beginning with the work of Kibarian et al. [6], [7], Guo et al. [8], and Stine et al. [9]. All of these studies involve the mathematical framework necessary for appropriate decomposition of variation across various
length scales (i.e., wafer-to-wafer, die-to-die, within-die) and systematic versus random components. However, data is presented only in relation to individual process steps such as etch, deposition or chemical-mechanical polishing.

More recently, the study of process variation has accelerated as scaling has resulted in increasingly limited control of individual process steps. Much of this work has focused on individual device parameters such as channel length (e.g., [10], [11]) and threshold voltage, primarily in the context of SRAM cells which are particularly sensitive to \( V_T \) variation. Agarwal et al., Rao et al., and Mukhopadhyay et al. at IBM [12]–[15], Fischer et al. [16], and Wang et al. [17] all have developed efficient measurement and characterization of arrayed structures, but only a subset of these present data regarding spatial correlation analysis of the measured results. Agarwal et al. present spatial correlation analysis of SRAM-sized devices in a 65 nm SOI process, concluding that little spatial correlation in \( V_T \) within each die exists at 65 nm. Fischer et al. also provide \( V_T \) variation data and autocorrelation analysis of 1-M SRAM cells at both the 90 nm and 65 nm nodes, with the results again showing no spatial correlation. Our own previous work, studying \( V_T \) variation, has also confirmed these results [18].

Studies of variation at the circuit level have typically focused on small ring oscillators. Ring-oscillator frequency is used to characterize variation at both within-die and die-to-die levels [19]–[22], due to the simplicity of the circuit and ease of frequency measurement. Pang et al. show ring-oscillator frequency measurements demonstrating weak spatial correlation in circuit performance [20]. It is unclear whether these results stem from systematic components of variation or are due to truly spatially-correlated random variation. Liu et al. have developed a statistical framework to use post-fabrication measurements of on-chip test structures to reduce delay uncertainties (distributions) of actual critical paths [23]. Their framework is general enough to include spatially correlated components, but the experimental results are based solely on simulation of correlated paths.

In the following sections, we not only characterize variation in larger circuits (e.g., adders), but also variation of components within these circuits. We begin by detailing the circuits that make such characterization possible.

III. Test Circuits and Chip Architecture

The variation test-chip is composed of a large number of replicated blocks containing various digital circuits and all-digital measurement circuitry. The following subsections further detail the architecture of each component.

A. High Frequency Test Circuits

To extract spatial correlation of high-frequency circuits, we array 80 nominally identical “adder-blocks” arranged in a \( 9 \times 9 \) matrix as shown in the die photo of Fig. 1(a) and occupying a 4 mm\(^2\) area. Each of these blocks contains common circuits found in modern product designs, namely a 64-bit adder, canary ring oscillators of various types and frequency measurement circuits as shown in Fig. 1(b). Details of each of these components are provided below.

1) Oscillating 64-bit Kogge–Stone Adder: The first circuit included in each “adder-block” is a 64-bit Kogge–Stone (KS) adder. Kogge–Stone adders belong to a class of adders known as parallel-prefix architectures which enable fast, efficient addition by pre-calculating carry signals [24]. The critical path in such adder structures is logarithmically related to the number of bits being added rather than linear in the worst-case, thus making wide adders feasible. Such adder structures are common building blocks in the datapaths of modern ASICs and microprocessors, making them good candidates for spatial variation analysis.

To enable simple and efficient measurement of maximum operating speed, the adder inputs are configured such that the outputs oscillate, requiring only a simple frequency counter for adequate characterization of performance. This is accomplished by setting one of the inputs to all ones \( (A|{63,0}\rangle = 1) \), the other input to all zeros \( (B|{63,0}\rangle = 0) \), and connecting an inverted version of the carry-out signal to the carry-in input \( (C_{in} = \overline{C_{out}}) \), as shown in Fig. 2.

All gates in the design are a custom standard cell implementation, each sized (using the HSpice optimizer [25]) such that they

---

1 We note that the design of this chip is flexible enough to allow any circuit capable of being configured in an oscillating condition, not just the ones chosen here, to be characterized in the same manner.
met an output rise/fall time constraint (50 ps in this implementation) for a given multiple of the minimum size load (i.e., a 1X cell must have a 50 ps output rise/fall time for a load of 5 fF; for a 2X cell the load is increased to 10 fF, etc.). Process design rules ensured that all devices within standard cells were oriented in the same direction. Carry-propagate cells (AND gates) and carry-generate cells (AND-OR gates) constitute the largest part of the adder, with the remainder of the gates being primarily XOR gates to perform the addition. The standard cell implementation allows for ease of layout due to the fixed pitch and the ability to abut cells without design rule violations. Although layout and routing were both manually done, this is representative of modern design flows for high performance blocks, as custom layout and routing often results in higher performance than completely automated flows.

2) Canary Ring Oscillators: To capture correlation between different circuit structures, we place 16 ring oscillators (ROs) around each adder, since they are often used as monitor circuits due to their simplicity and small area overhead. The ROs used are: INV[9,11,13,15], NAND[9,11,13,15] and NOR[9,11,13,15], denoting the type of gate and number of gate delays. The INV9, NAND11, NOR13 and INV15 are duplicated for a total of 16 ROs per adder. The NAND and NOR gates are two-input gates with inputs shorted to produce an inverting gate. These differ from the AND-OR gate used in the KS adder, which allows for quantification of correlation between disparate gate types and transistor stacks. The 16 ROs are divided into four blocks and interspersed between other digital logic, as shown in Fig. 1(b), to ensure that these “canary” circuits are within the context of circuits normally found in modern digital designs.

3) Frequency Measurement Circuits: Two 32-bit asynchronous frequency counters, made up of simple toggle flip-flops (Fig. 3), measure the oscillation frequency of the adder and ROs. An additional two frequency counters are also included for use in measuring the delay of each individual adder bit relative to the first bit in the adder, detailed in the following section. These frequency counters provide a simple solution to high-resolution digital measurement, allowing completely digital read-out of all relevant frequencies and delays. The outputs of the frequency counters are multiplexed onto a single bus spanning each row of the chip, and these row buses are multiplexed onto the chip outputs.

B. High-Resolution Digital Delay Measurement

Capturing variation data with higher spatial and temporal resolution than possible with simple ROs requires alternative techniques. Specifically, we seek a highly-scalable, all-digital measurement technique capable of sub-picosecond delay resolution occupying small area. In the following sections we describe the theory and operation of a random-sampling technique, in the context of measuring delays between bits of the adder, meeting these criteria.

1) Random Sampling: The critical path in the adder is from \( C_{in} \) to \( C_{out} \), resulting in all bits oscillating at an identical frequency but with delays determined by the logarithmic structure of the KS adder, giving 64 out-of-phase oscillator taps. Quantifying this phase-delay is equivalent to quantifying the difference in delay between each bit, and allows for variation analysis with improved spatial resolution. Delay measurement is done by randomly sampling the signals and counting the number of occurrences when one of the signals is logic high and the other simultaneously low, or vice-versa, dividing by the total number of samples taken and multiplying by the signal period as shown in (1). For example, counting the number of times \( C_{in} \) is logic high \( (N_{up}) \) while \( C_{out} \) is simultaneously logic low \( (N_{dn}) \) would give the delay between \( C_{in} \) and \( C_{out} \):

\[
D = \tau_{perr} \frac{N_{up} - N_{dn}}{N_{tot}}.
\]

Uses of random-sampling techniques for this purpose are not new [26]– [28]. However, in each of those implementations, an XOR gate is used to determine when one of the signals is logic high and the other logic low. XOR gates can limit the minimum detectable delay since some minimum delay is necessary to register a clear logic level at the output. In our implementation, we transform the measurement to be that of the difference in two pulse widths using a phase-frequency detector (PFD), discussed in the following section, which eliminates this constraint.

For this technique to be accurate, the random samples must be uniformly distributed across all points in the sampled signal cycle, as non-uniformity in this distribution results in non-linearities in the measured delays. Mathematically, the operation shown in Fig. 4(a), where the UP and DN signals are being randomly sampled, can be modeled as

\[
S_{j+1} = S_j + X_j = \sum_{i=1}^{j} X_i
\]

where \( S_j \) is the timing of the jth sample, \( X_i \) is the time between samples and is a random variable, and \( \tau_{perr} \) is the sampled signal period. We also define \( P_j = \text{mod}(S_j, \tau_{perr}) \) as the position of the jth sample in the sampled signal cycle. If the average sampling period \( \langle \tau_X \rangle \) and the sampled signal period \( \langle \tau_{perr} \rangle \) are co-prime, it is clear that the distribution of \( P_j \) will be uniform even without random edges. However, if they are not co-prime, (2) describes a random walk, which we simulated in Matlab using a distribution of random periods taken from
Fig. 4. Graphical depiction of random sampling. (a) Random sampling of the UP and DN signals to determine pulse widths. The dashed arrows represent sampling instants and the solid line depicts the evolution of $S_j$ over time. (b) Convergence of sampling instant to a uniform distribution.

Fig. 5. PFD and associated timing diagram. (a) PFD [30]. (b) Timing diagram.

HSpice simulations of the LFSR-controlled RO. The results of this simulation, in Fig. 4(b), empirically show convergence to a uniform distribution as $N \to \infty$. With $N > 10^8$ samples the non-uniformity of the distribution, and thus non-linearity in the measured delay values, is sufficiently small. To be quantitative, the framework in [28] can be used to find standard errors and confidence intervals on the time resolution of this technique given some number of samples. Conversely, it can also be used to compute the number of samples required for a desired resolution given desired confidence intervals. Assuming equal probability of a sampling instant falling anywhere in a clock cycle, (3) shows how the number of samples required, $N$, depends on $z_c = \sqrt{2}\sigma f^{-1}(CI)$ which gives the area underneath an appropriate Gaussian curve for a certain confidence interval, $CI$, and $p$ and $P$, which are the actual and observed duty cycles, respectively.

$$N \approx \left(\frac{z_c}{p - P}\right)^2 P(1 - P). \tag{3}$$

As shown in Fig. 4(b), $N > 10^8$ samples will satisfy the uniform distribution condition. Section IV-C will show that the number of samples taken in our random sampling implementation is $\approx 1.7 \times 10^8$, in which case the non-uniformity is $\ll 1\%$, ensuring good linearity. It should also be noted that the time resolution of this technique can be arbitrarily increased with large enough $N$, which would result in smaller errors between actual and observed duty cycle ($p - P$).

2) Measurement Circuits: Most high-resolution, sub-picosecond measurement techniques require complicated circuitry capable of exquisite timing. The nature of random sampling techniques, in making use of a large number of samples spread over many cycles, relaxes the measurement circuit constraints substantially. In this implementation there are two major circuits: a PFD and a random sampling clock generating circuit.

Removing the constraint on minimum measurable delay is achieved by transforming the measurement to that of the difference in two pulse widths. Since we are attempting to measure the delay difference between bits of the KS adders, and because the adders are in an oscillating configuration, the oscillating frequency and pulse widths remain constant for the duration of operation. This situation lends itself nicely to sampling over a large number of cycles as the inherent properties of the sampled signal remain constant, or stationary, over time.

A PFD as shown in Fig. 5(a), normally used in phase-locked loops, is used here to convert between a delay difference and a pulse-width difference. The circuit is comprised of two modified true single-phase clock (TSPC) flip-flops and a self-resetting circuit. For simplicity in explanation, we assume that the signal that arrives first, or the early signal, is $Sum(x)$, the input to the
top TSPC flop; however, the following explanation, graphically depicted in Fig. 5(b), can be completely reversed with no loss in circuit functionality. When the early signal arrives it triggers a rising edge on the UP signal. Similarly, when the late signal arrives at some point later, a rising edge is triggered on the DN signal. The arrival of the late signal simultaneously triggers the self-resetting circuitry connected to the outputs of both flops, resulting in simultaneous falling edges on both the UP and DN signals.\footnote{Variation and mismatch between the flops result in a non-simultaneous reset and a corresponding static offset. However, with the measurement technique discussed in Section IV-C, this offset can be accounted for and eliminated.} As the timing diagram shows in Fig. 5(b) by virtue of the simultaneous falling edge, the circuit has transformed a delay measurement into a pulse-width difference measurement. Moreover, the difference of these two pulse widths can be arbitrarily close to zero, eliminating any bound on minimum measurable delay.

Since the UP and DN pulses have differing pulse-widths, it is these signals that are randomly sampled using a simple D-flip-flop, clocked by a random clock. Generating this random clock is done by attaching 32 of the outputs of a 63-bit linear feedback shift register (LFSR) to the tri-state controls of inverters in four stages of a 5-stage ring oscillator as shown in Fig. 6. Each stage of the ring oscillator is variable drive-strength; since the LFSR produces pseudo-random bit sequences, the tri-state controls of the ring oscillator are enabled in a pseudo-random fashion, in effect generating the random clock we require. Since LFSRs can repeat the pseudo-random pattern if not designed with enough bits, this implementation is comprised of 63 bits, which theoretically is sufficient to avoid repetition over many years of operation at the designed operating frequencies.

HSpice simulations show the distribution of clock periods generated by this circuit (Fig. 7). The generated clock periods are indeed randomized with a distribution that can be closely approximated by a Gaussian distribution except in the tails. Nevertheless, we are more interested in the distribution of sampling instants within a clock cycle, which the previous section showed to be uniform. Despite the relatively simple circuitry and the fact that the clock periods form an approximate Gaussian distribution, with enough samples, the “drifting” nature of these periods results in a uniform distribution of the sampling instants within a clock cycle.

A single PFD is included within each replicated “adder-block” and shared between all bits of the adder. The first sum bit of the adder is directly connected to one of the inputs of the PFD and all 64 bits of the adder are mux-able into the other input of the PFD. This setup allows subtraction of any offsets (e.g., due to multiplexing or variation in the two halves of the PFD) or other non-idealities by measuring the first sum bit versus the directly connected version of itself and subtracting that result from all other measurements.

IV. DATA ANALYSIS

The test-chip described above was fabricated in IBM’s 90 nm triple-well process technology, as pictured in Fig. 1(a), and frequency measurements were carried out on 41 chips from two wafers. This section describes the measurements and quantifies their spatial decomposition. Given the limited spatial sampling of chips across the wafers, we do not study spatial wafer-level patterns, and instead focus on within-die variation with comparisons to aggregate wafer-scale variance.

A. Variation Measurements

Variation measurements show that the effect of variation is larger at lower power-supply voltages (Fig. 8), concurring with previous studies [29]. As \( V_{DD} \) is decreased below 0.7 V, variation increases dramatically. Moreover, the within-die component becomes a larger and larger component of the overall variation. While noise is typically a concern at low voltages, all of the measurements are taken by allowing the circuits to run freely for \( > 10^8 \) cycles, averaging out any (assumed white) noise.

As expected, ROs with smaller gates (e.g., INV) and fewer number of delay stages show greater within-die variation (due to less averaging, see below) than ROs with either larger gates.
(e.g., NOR) or more delay stages. At the extremes, the INV9 RO is most variable ($\sigma_{\text{WD}} = 4.9\%$ at 0.5 V and $\sigma_{\text{WD}} = 1.13\%$ at 1.2 V) and the adder, being the largest circuit, is least variable ($\sigma_{\text{WD}} = 2.8\%$ at 0.5 V and $\sigma_{\text{WD}} = 0.82\%$ at 1.2 V). Die-to-die and total variance show the same general trends except that the NAND ROs are more variable than the INV ROs—we believe this may be due to larger NMOS than PMOS global variability as NAND gate delays are typically dominated by the series NMOS stack.

Separating the die-to-die and within-die components of the variation is a matter of appropriately subtracting various means. To extract within-die variance for a particular circuit type, the die mean for that circuit is subtracted from each measurement, as in (10). The square of these differences is averaged over all circuits within a die and over all die. Computing die-to-die variance involves subtracting the grand mean over all measurements from each of the die means, as shown in (11), and averaging the square of the differences over all die. Spatial correlation computations make use of exactly identical mean subtraction except that the NAND ROs are more variable than the INV ROs—we believe this may be due to larger NMOS than PMOS separation distance.\(^3\)

Decomposing the variation into die-to-die and within-die components in this manner reveals that not only does the fraction of variation attributable to within-die variation increase with decreasing voltage, but perhaps more importantly, for some circuits, namely the INV ROs, the within-die and die-to-die components are roughly equal at those low voltages. Even in the “best” cases, the within-die component increases from 25% of the die-to-die component when $V_{\text{DD}} > 1$ V to 50% at low voltages ($V_{\text{DD}} = 0.5$ V). The asymmetric increase in variation components strongly impacts the design of variation mitigation schemes, especially those operating in low-power and low operating voltage regimes or dynamically scaling systems, as detailed further in Section V.

\(^3\)Further details of the computations for each of these components as well as the spatial correlation computations in the next section can be found in the Appendix.

B. Spatial Variation and Correlation

The large number of replicated circuits within each die allows extraction of spatial correlation. Within-die spatial correlation is computed using (4), where $x$ is a particular parameter value (here, it is either delay or frequency) with the mean of each die subtracted from $x$. All circuits separated by a distance, $d$, over all die are included in the computation which increases the number of data points per distance, thereby increasing the statistical significance of the computed results. At the closest separation distance of 170 $\mu$m between adjacent adders, there are nearly 1476 (41 chips $\times$ 36 adjacent adders) independent pairs of adders that contribute to the correlation computation, while at the farthest separation distance of 2 mm there are only 82 (41 chips $\times$ 2 pairs) independent pairs of adders.

$$\rho_{x,x+d}(d) = \frac{\text{Cov}(x,x + d)}{\sigma_x^2},$$  (4)

We find that even at high $V_{\text{DD}}$, there is no discernible within-die spatial correlation [Fig. 9(a)]. The dip at the $V_{\text{DD}} = 1.2$ V and 2 mm separation distance corner is likely due to limited data, as there are only 82 pairs of adders at this separation distance over 41 die. Furthermore, all points are within a 95% confidence interval, giving additional credence to this argument. While our die is medium sized at 2 mm per side, we do not see any indication of within-die spatial correlation that would be visible for either large or small die if significant separation distance decaying spatial correlation existed. Lack of within-die spatial correlation does not imply a lack of systematic within-die variation: rather, it simply means such variation is not a function of separation distance. Instead, this implies that the statistics of adder frequency (e.g., mean and standard deviation relative to chip-mean) is dependent on position within the die, as evidenced in Fig. 10. Indeed, a systematic within-die pattern in adder frequency is noticed in Fig. 10(a). To elucidate this further, consider equidistant pairs of adders $\{(6,6), (6,7)\}$ and $\{(7,6), (7,7)\}$ that exhibit both unequal and directionally opposite frequency differences.

\[\text{Fig. 8. Variation as a function of } V_{\text{DD}}.\]
$\Delta F(\{6, 6\}, \{6, 7\})$ and $\Delta F(\{7, 6\}, \{7, 7\})$, which implies lack of correlation.

Since this systematic pattern introduces position-dependent variation statistics, the pattern should be removed prior to any spatial correlation computation to ensure that the residue is stationary. This is necessary, as computing correlation implicitly assumes that all samples are from the same probability distribution. When the systematic pattern shown in Fig. 10(a) is subtracted (after subtracting out die means), the newly computed spatial correlation is shown in Fig. 9(b). Comparing the two figures, there is little difference owing to the mostly random nature of the systematic pattern. The small peaks that were present due to the slow bottom row of adders have now been flattened out, showing conclusively that there is no spatial correlation in within-die variation.

Nevertheless, we can study die-to-die correlation, which is defined mathematically in (14). Intuitively, die-to-die spatial correlation in this context can be thought of as the level of allowable inference between two devices or circuits separated by some distance, $d$, on die $i$ given information about only one of those devices on die $i$, complete information about devices or circuits on die $j$ separated by the same distance and the difference in die means. Fig. 9(a) shows strong die-to-die correlation at high $V_{DD}$ but decreasing with lower $V_{DD}$, indicating that the effect of random variation increases at lower power-supply voltages. This is consistent with $\sigma_{D2D} \gg \sigma_{W2D}$ at higher voltages, but the relative fraction decreasing at lower voltages as seen in Fig. 8. Since the within-die variation is random (spatially uncorrelated and even the systematic component not containing any apparent order), die-to-die correlation should decrease as the within-die component increases in relative strength. These results are also consistent with the effect of threshold voltage ($V_T$) variation, which is dominated by Random Dopant Fluctuation (RDF), increasing as gate overdrive decreases [18].

Furthermore, die-to-die correlation shows only weak dependence on separation distance. Qualitatively, this means that when die-to-die correlation is strong, knowledge of how two identical circuits differ from one die to another enables strong inference when comparing any other circuits (of the same type) between those die, regardless of how far those circuits may be separated from the original circuit. However, if voltages (and thus correlation) are decreased, such inferences become increasingly weak.

While Fig. 9(a) only shows adder correlations, the spatial correlation results for all ring oscillators are similar, with the notable exception that die-to-die correlation decreases with decreasing $V_{DD}$ more quickly with smaller circuit size. As an example, the INV9 based RO has a correlation coefficient, $\rho$, of 0.55 at $V_{DD} = 0.5$ V compared with $\rho = 0.65$ for the INV15 (shown in Fig. 11) and $\rho = 0.75$ for the adder, consistent with smaller circuits being more susceptible to random variation sources as less averaging occurs. On a die-to-die scale, these results closely match the results of our previous work in [18], which showed that correlated channel length variation could result in strong correlation at high $V_{DD}$ but degrades rapidly as $V_{DD}$ is scaled downward to save power.
Fig. 11. Inverter spatial correlation plots showing different decreases in die-to-die correlation with decreasing voltage based on number of stages. (a) 9-stage inverter RO. (b) 15-stage inverter RO.

Fig. 12. Scatter matrices of each pair-wise combination of circuit types showing no cross-circuit correlation at $V_{DD} = 1$ V. All circuit pairings are within the same “adder-block”.

Since ring oscillators are commonly used “canary” devices, thought to predict the performance of circuits situated nearby, analyzing the cross-circuit correlation is important as well. Extrapolating from the earlier results showing weak or no within-die spatial correlation, a lack of cross-circuit correlation is expected. Fig. 12 contains scatter matrices for each pair-wise combination of circuit types. Perfect cross-circuit correlation ($\rho_{xy} = 1$) would be manifested by a tight distribution, along a positive diagonal line ($y = x$), of all the scattered points within individual axes. Similarly, perfect anti-correlation would be manifested in the opposite manner, i.e., a distribution along a negative diagonal line ($y = -x$). Lastly, circular scatterings reflect no correlation at all, which is the case in the majority of the pair-wise circuit combinations. Only 7 (out of 289) circuit pairs have correlations greater than 0.4 and all of these circuit pairings involve the longest and largest ROs; the largest correlation, of $\rho_{xy} = 0.53$ is noticed between the NAND13 and NOR15 ROs, which are located adjacent to each other. This
indicates that increased averaging is occurring in these larger circuit structures. Scatterings that form tighter distributions along either the horizontal or vertical axes indicate only that there is less frequency variation in one of the circuits in the pairing (as evidenced by the accompanying histograms), not any correlation. These results are for high voltage ($V_{DD} = 1$ V) and the closest separation distance possible for each circuit pairing; each circuit is compared to the circuits in the same “adder-block.” Repeating this calculation for larger separation distances or other voltages reveals identical results, as is expected from the earlier “same-circuit” spatial correlation results.

C. Adder Bit Delays

Using the all-digital delay measurement circuits described in Section III-B, we measure the delay of each bit within each KS adder, relative to $Sum(0)$ of the same adder. All 64 bits, including $Sum(0)$, are muxed into a single PFD. By using the same PFD for all bits, offsets and other non-idealities due to mismatch in the PFD structure can be measured (by measuring the delay between the muxed and non-muxed versions of $Sum(0)$) and subtracted from subsequent measurements.

Measurements over all 80 adders on 40 chips are shown in Fig. 13 with a post-layout extracted simulation for comparison. There is good agreement between measured delays and simulation: the upward shift between measured data and simulation indicates a slower process than nominal simulation, and is consistent with 20% slower frequency measurements than nominal post-layout simulations. The measured delay pattern is consistent with simulation for all but three of the bits, $Sum(16, 32, 48)$, which are typically the fastest due to the logarithmic structure of a KS adder. However, in this layout, these three bits contain longer wires than the other bits, corresponding to the peaks in the post-layout simulation. While these three bits do have larger measured delays than nearly all of the other bits, the difference is not as large as in simulation, possibly due to slower transistors but “faster” wires, which would decrease the delay peaks formed by these three bits.

Each data point in Fig. 13 consists of no less than $1.7 \times 10^8$ random samples. Using the theoretical analysis in [28], we compute a possible observed error of approximately 200 fs, with a confidence level of 99.9999%. Combined with the general agreement between measured data and simulation, this computation gives high confidence in sub-picosecond accuracy of the measured data.

Due to the correlated structure of the adder, care must be taken when doing spatial correlation analysis using the measured bit-slice delays. Since the critical path of the KS adder involves all bits, intuitively all bits will be correlated with each other to some degree. The logarithmic nature (in particular, radix 2, or log-base-2) of the adder implies that bits two away, four away, eight away, etc. from each other will be more correlated than other combinations of bits. Fig. 14 shows the cross-correlation of bit-slices within the same adder and indeed reveals this logarithmic structure. The prominent, darker diagonals starting on the x-axis at bits 9, 17, 25, 33, 41 and 49 show stronger correlations between bit-slices 32 and to a lesser degree 8 and 4 away from each other. The 32 bit-slice separation is most strongly correlated, as it is an input nearer the end of the higher bit-slice’s path and arrives later than outputs of bits nearby, therefore more strongly influencing the delay of that bit-slice. Also noticeable, and expected, are strong cross-correlations in
bit-slices 1–16 and 33–48 by virtue of the carry signal connections to logarithmically pre-compute the overall carry-out signal.

Since the adder circuit structure largely determines the correlation of bit-slices, within-adder correlation does not reveal much about spatial correlation due to process variation. Nevertheless, looking at bit-slices relatively close to each other but not as strongly connected, for example bit-slices 16 and 20, reveals much weaker correlation: $\rho = 0.4$ compared to $\rho > 0.8$ for bit-slices separated by 32. Such results show that the relative significance of circuit structure and connectivity is far greater than process variation, and is likely the only significant source of spatial correlation.

Knowing that circuit structure and connectivity significantly impact the apparent correlation, the correlation between each bit-slice and its own adder performance should be strong. However, Fig. 15 shows that each bit-slice is only moderately anti-correlated with the performance of the adder it partly constitutes. The anti-correlation is simply due to comparing delays with frequencies and the inverse relationship between the two. The moderate, rather than the expected strong, correlation is likely an artifact of the comparison being made: the phase delays are all relative to $\text{Sim}(0)$ while the adder frequencies are absolute. Unfortunately, since we are not able to measure the absolute delay of the $\text{Sim}(0)$ bit-slice, we can only speculate that with absolute delays a stronger correlation would be extracted.

Further spatial correlation analysis of bit-slices across different adders is also possible and, as expected from all previous results, shows no significant spatial correlation. Fig. 16 shows the within-die correlation between two bit-slices picked at random, revealing strong correlation only at the distance corresponding to the two bits being part of the same adder. At larger separation distances, there appears to be slight anti-correlation. However, as we noted previously with respect to the adder frequencies, there are far fewer independent bit-slice pairs separated by these distances; the stronger anti-correlation here is likely due to random chance with smaller numbers of samples, as evidenced by the much larger confidence intervals. Despite showing the cross correlation for only two bit-slices here, any arbitrary combination of bits reveals similar results.

V. CONCLUSION

We have designed and implemented a test-chip to characterize variation in common digital circuits, including the explicit ability to decompose the various spatial components of observed variation. Variation measurements highlight the fact that averaging of random variation mitigates the impact of variation on larger gates and circuits. However, at lower voltages this becomes less true: the magnitude of variation increases as does the relative contribution from within-die components.

As a result, variation mitigation strategies must be a function of the voltage/power domain in which circuits are operated. In high-performance domains where gate over-drive is sufficiently large ($> 2 V_T$), although within-die variation is random and uncorrelated spatially, the absolute variance is also small. Mitigation schemes involving small “monitor” circuits, such as ring

\footnote{Although there are smaller distances on the plot, these smaller distances correspond to these bit-slices being part of different adders due to the layout of the adder.}
oscillators or replica critical paths, require little margin—perhaps as little as 1%, as determined by the absolute variance and desired confidence level—to be effective. In low-performance, low-power domains in which $V_{DD} \approx V_T$, variation as a percentage of the mean is significantly increased, as much as $5\times$ relative to high-performance voltage domains. Furthermore, within-die variation is a significant component of the overall variation seen. In combination, these two factors necessitate in-situ circuits capable of accurate measurement of timing data as the basis of a robust mitigation strategy.

Further analysis reveals that spatial correlation can often be confounded with systematic variation or even correlation arising from circuit structure. When these confounding components are properly accounted for, there appears to be no spatial correlation present in within-die variation. These results suggest that incorporation of spatial correlation in statistical static timing models is not necessary. Rather, incorporating positional dependencies (systematic components) may be more valuable, although quantifying these dependencies is likely difficult due to the apparent randomness of the systematic components. Simulation and layout tools could be explored to identify and compensate for such positional systematic variations, and to understand the impact and effectiveness of design approaches, such as common centroid design, given specific types of within-die variations.

**APPENDIX**

The frequency of any RO is denoted as $f_{ijk}$, where:

$$i = \text{die #, } 1 \ldots D$$

$$j = \text{RO type, } 1 \ldots J$$

$$k = \text{replicate #, } 1 \ldots K.$$  \hspace{1cm} (5)

The die mean and grand mean for a given RO type is computed according to (8) and (9):

$$\bar{f}_{ij} = \frac{1}{K} \sum_{k=1}^{K} f_{ijk}$$

$$\bar{f}_{ij} = \frac{1}{DK} \sum_{i=1}^{D} \sum_{k=1}^{K} f_{ijk}.$$  \hspace{1cm} (8)

The within-die variance for a given RO type is computed using

$$\sigma^2_{WID} = \frac{1}{D(K-1)} \sum_{i=1}^{D} \sum_{k=1}^{K} (f_{ijk} - \bar{f}_{ij})^2.$$  \hspace{1cm} (10)

Similarly, the die-to-die variance per RO type is given by

$$\sigma^2_{DD} = \frac{1}{D-1} \sum_{i=1}^{D} (f_{ij} - \bar{f}_{ij})^2.$$  \hspace{1cm} (11)

When computing spatial correlations, separation distance must be included in the notation. Accordingly, we represent the frequency of each RO as $f_{ij}(x,y)$ where $(x,y)$ denotes the location of a particular replicate of RO type $j$ on die $i$. The separation distance is defined in the normal way:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$  \hspace{1cm} (12)

Since correlation is computed on a per-distance basis, the set of independent pairs of ROs, $f_{p1}(x_1, y_1)$ and $f_{p2}(x_2, y_2)$, separated by $d$ is given by $S(d)$, and the cardinality is $|S(d)|$. The within-die correlation is then computed by

$$\rho_{WID}(d) = \frac{1}{\sqrt{|S(d)|}} \sum_{i=1}^{D} \sum_{j=1}^{K} \frac{(f_{p1} - \bar{f}_{ij}) (f_{p2} - \bar{f}_{ij})}{\sigma_{WID}^1 \sigma_{WID}^2}.$$  \hspace{1cm} (13)

where $f_{p1}$ and $f_{p2}$ are the frequencies of each of the ROs in a pair and the variances of each side of the RO pair set are defined as in (13) for $\sigma_{p1}$, and similarly for $\sigma_{p2}$.

Lastly, die-to-die spatial correlation is computed in a similar manner, with the grand mean for each RO type replacing the die mean.

$$\rho_{DD}(d) = \frac{1}{\sqrt{|S(d)|}} \sum_{i=1}^{D} \sum_{j=1}^{K} \frac{(f_{p1} - \bar{f}_{ij}) (f_{p2} - \bar{f}_{ij})}{\sigma_{DD}^1 \sigma_{DD}^2}.$$  \hspace{1cm} (14)

$$\sigma^2_{DD} = \frac{1}{D-1} \sum_{i=1}^{D} (f_{ij} - \bar{f}_{ij})^2.$$  \hspace{1cm} (15)

**ACKNOWLEDGMENT**

The authors acknowledge the support of the Focus Center for Circuit & System Solutions (C2S2), one of five research centers funded under the Focus Center Research Program, a Semiconductor Research Corporation program. The authors would like to thank D. Lim, K. Balakrishnan, N. Verma, and Y. Ramadass for their aid in chip testing and invaluable discussions regarding data analysis and implications.

**REFERENCES**


