Testing On-Die Process Variation in Nanometer VLSI

Mehdrad Nourani
University of Texas at Dallas

Arun Radhakrishnan
Texas Instruments

Editor's note:
Ring oscillators are not new, but the authors of this article use them in a novel, unconventional way to monitor process variation at different regions of a die in the frequency domain.

—T.M. Mak, Intel

The driving force behind current IC fabrication technology is the demand for circuit complexity and density. A challenge common to process engineers and circuit designers in trying to meet this demand is the effect of process variation (PV) on design characteristics such as functionality and performance. PV is the deviation of parameters from desired values due to the limited controllability of a process. Every process has some level of uncertainty in its device parameters. As device size continues to scale down into the ultra-deep-submicron regime (less than 100 nm), manufacturing tools are less reliable in their control of design parameters. PV usually arises from limitations imposed by the laws of physics, imperfect tools, and properties of materials that are not fully comprehended. Sources of PV include random dopant fluctuation, annealing effects, and lithographic limitations. Typical variations are 10% to 30% across wafers and 5% to 20% across dies, and such variations can change the behavior of devices and interconnects.[1] (See the "Related work" sidebar for a discussion of the various methods researchers have developed to deal with process variation.)

Measuring the variation of each design or fabrication parameter is infeasible from a circuit designer's perspective. Therefore, we propose a methodology that approaches PV from a test perspective. This methodology advocates testing dies for process variation by monitoring parameter variations across a die and analyzing the data that the monitoring devices provide. We use ring oscillators (ROs) to map parameter variations into the frequency domain. Our use of ROs is far more rigorous than in standard practices. To keep complexity and overhead low, we neither employ analog channels nor use zero-crossing counters. Instead, we use a frequency domain analysis because it allows compacting RO signals using digital adders (thereby also reducing the number of wires), and decoupling frequencies to identify high PVs and problematic regions.

Our PV test methodology includes defining the PV fault model; deciding on types, numbers, and positions of a small distributed network of frequency-sensitive sensors (ROs); and designing an efficient, fully digital communication channel with sufficient bandwidth to transfer sensor information to an analysis point. With this methodology, users can trade off cost and accuracy by choosing the number or frequency of sensors and regions on the die to monitor.

PV fault model

Tracing PVs for each individual parameter is impossible because of the size, complexity, and unpredictability of the factors involved. As in conventional testing approaches such as stuck-at fault and path delay, this approach requires a simplified PV fault model in order to devise and apply a PV test methodology. Despite this simplicity, however, the model should be generic in concept, straightforward in measurement, and practical in application. Using this criteria, we define the following:

The single-PV fault model assumes that only one faulty grid (unit area) exists in the layout of the circuit under test, where a sensor planted in that region can generate a faulty metric, z_f, instead of a fault-free (acceptable) metric, z, such that $\Delta z = |z_f - z|$ is measurable—in terms of delay, frequency, and so on. (In a taxonomy of VLSI testing, *fault* refers to a failure mechanism, and its presence requires automatic rejection. We acknowledge that our term *PV fault* stretches

0740-7475/06/$20.00 © 2006 IEEE Copublished by the IEEE CS and the IEEE CASS

Related work

Researchers have explored various ways of analyzing and dealing with process variation (PV). The solutions are generally design-for-manufacturing techniques. DFM tries to quantify the impact of PV on circuits and systems. Such a role has made DFM techniques very interesting to semiconductor and manufacturing companies. Several factors affect or contribute to PV: interconnects, thermal effects, gate capacitance, and so on.[1] PV potentially causes 40% to 60% variation for effective channel length L_{eff}, and 10% to 20% fluctuation in both threshold voltage V_{TH} and oxide thickness T_{ox}, potentially leading to a malfunction.[2]

You can classify PV monitoring and analysis approaches using different criteria. From the source perspective, variation can be intradie (within the die) or interdie (between dies). The latter can be die to die, center to edge (in a wafer), wafer to wafer, lot to lot, or fab to fab. From a methodology perspective, solutions fall into two broad categories: statistical and systematic. Examples of statistical approaches include using PV modeling,[3] analyzing the impact of parameter fluctuations on critical-path delay,[4] mapping statistical variations into an analytical model,[5] and addressing PV's effect on crosstalk delay and noise.[6]

In systematic approaches, because of the parameters' complexity, almost all researchers have traced very limited PV metrics or design characteristics. Orshansky et al. explored the effect of gate length on performance.[7] Chen et al. proposed a current monitor component to design PV-tolerant circuits.[8] Azizi et al. analyzed the effect of voltage scaling on making designs more resistant to PVs.[9] Other researchers considered PV's effect on key design characteristics.[10-13] Mehrotra studied manufacturing variation's impact on microprocessor interconnects.[10] Ghanta et al. showed PV's effect on the power grid.[11] Agarwal et al. presented a failure analysis of memories by considering PV effects.[12] Ding, Luo, and Xie investigated PV's effect on soft-error vulnerability.[13]

Owing to the nature of parameters affected by PV, such as threshold voltage V_{TH}, oxide thickness T_{ox}, and effective channel length L_{eff}, tracing and pinpointing each variation for any realistic circuit is not a viable option. Hence, it's necessary to limit the problem to a specific application domain or design metric. Such specificity appears in earlier works that consider PV for clock distribution, delay test, defect detection, PV monitoring techniques, reliability analysis, and yield prediction.

References

1. C. Dryden, "Survey of Design and Process Failure Modes for High-Speed Series in Nanometer CMOS," *Proc. 23rd IEEE VLSI Test Symp.* (VTS 05), IEEE CS Press, 2005, pp. 285-291.
2. S. Borkar, "Microarchitecture and Design Challenges for Gigascale Integration," *Proc. 37th Ann. Int'l Conf. Microarchitecture* (Micro 37), IEEE CS Press, 2004, pp. 2-3.
3. H. Sato et al., "Accurate Statistical Process Variation

continued on p. 440

the terminology a bit, because the presence of a fault might only mean low performance and not necessarily a nonfunctional die.)

Based on this definition, we elaborate some key concepts:

- *Location.* A conventional stuck-at-fault model assumes that a net (interconnect segment) is a fault's physical location. A grid, on the other hand, is a minimum size (unit area) region in the layout whose variation can be traced through simulation, measurement, and comparison.
- *Behavior.* A stuck-at-fault model assumes that the behavior in a particular net is permanently either 0 or 1. Here, the behavior is a particular metric shift, Δz. An ideal sensor spreads uniformly across the grid, accumulates all the effects of PVs, and reflects these effects in Δz.
- *Test mechanism.* In testing stuck-at faults, we have two objectives: stimulating the fault from the primary inputs and propagating it to an observable point. In PV testing, no fault stimulation is necessary, thanks to PV's self-generating and on-spot nature. But a sensor's output must proceed to the observation point. We can collect the most accurate PV test data when the die is entirely covered only by sensors. In practice, sensors and test circuitry should be a small percentage of the die. Therefore, we use the concept of *fault sampling* to remain practical while getting a good coverage estimate. According to the fault sampling, we randomly choose N_s grids where PV faults occur out of N_p, and we devise sensors to collect data on PVs. The test data goes to an observation point, where a frequency-domain analysis determines the coverage.

continued from p. 439

Analysis for 0.25-µm CMOS with Advanced TCAD Methodology," *IEEE Trans. Semiconductor Manufacturing*, vol. 11, no. 4, Nov. 1998, pp. 575-582.

4. A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical Timing Analysis for Intra-Die Process Variations with Spatial Correlations," *Proc. IEEE/ACM Int'l Conf. Computer Aided Design* (ICCAD 03), IEEE CS Press, 2003, pp. 900-907.
5. Y. Cao and L. Clark, "Mapping Statistical Process Variations toward Circuit Performance Variability: An Analytical Modeling Approach," *Proc. Design Automation Conf.* (DAC 05), ACM Press, 2005, pp. 658-663.
6. U. Narasimha, B. Abraham, and Nagaraj NS, "Statistical Analysis of Capacitance Coupling Effects on Delay and Noise," *Proc. 7th Int'l Symp. Quality Electronic Design* (ISQED 06), IEEE CS Press, 2006, pp. 795-800.
7. M. Orshansky et al., "Impact of Systematic Spatial Intra-Chip Gate Length Variability on Performance of High-Speed Digital Circuits," *Proc. IEEE/ACM Int'l Conf. Computer Aided Design* (ICCAD 00), IEEE CS Press, 2000, pp. 62-67.
8. Q. Chen et al., "Process Variation Tolerant Online Monitor for Robust Systems," *Proc. IEEE Int'l On-Line Testing Symp.* (IOLTS 05), IEEE CS Press, 2005, pp. 171-176.
9. N. Azizi et al., "Variations-Aware Low-Power Design with Voltage Scaling," *Proc. Design Automation Conf.* (DAC 05), ACM Press, 2005, pp. 529-534.
10. V. Mehrotra et al., "Modeling the Effects of Manufacturing Variation on High-Speed Microprocessor Interconnect Performance," *Proc. Int'l Electron Devices Meeting* (IEDM 98), IEEE Press, 1998, pp. 767-770.
11. P. Ghanta et al., "Stochastic Power Grid Analysis Considering Process Variations," *Proc. Design, Automation and Test in Europe* (DATE 05), IEEE CS Press, 2005, pp. 964-969.
12. A. Agarwal et al., "Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture," *IEEE J. Solid-State Circuits*, vol. 40, no. 9, Sept. 2005, pp. 1804-1814.
13. Q. Ding, R. Luo, and Y. Xie, "Impact of Process Variation on Soft Error Vulnerability for Nanometer VLSI Circuits," *Proc. 6th Int'l Conf. ASIC*, IEEE Press, 2005, pp. 1117-1121.

Using ring oscillators as PV sensors

Using ROs to probe on-die PV is a long-standing practice. PV affects the oscillators inserted into the system along with the rest of the system. Delay variation—due to PV faults, for example—in any inverters in the loop results in a detectable deviation in the oscillator's frequency. Some fabrication companies have already used ROs on wafers to monitor PVs. In fact, this monitoring often serves as a benchmark performance measure (see http://www.mosis.org/Technical/process-monitor.html). Conventionally, fabrication companies place several on-wafer ROs in a dicing line for process parameter monitoring. Unfortunately, this is insufficient for evaluating each die's variations; there is a growing demand for sensors and methodologies that allow precise PV evaluation.

Within-die variations are significantly more difficult to predict and handle than die-to-die variations on a wafer.[2] Hatzilambrou, Neureuther, and Spanos demonstrated that on-chip sensors with counters, embedded in a test chip, could detect variations in frequency.[3] They then used this frequency variation to grade the die's and the individual cores' performance. They did not intend this approach to be a pass/fail test but rather a grading test. Such PV information complements, not replaces, existing functional testing. The oscillators estimate actual delays throughout the chip and thus can estimate system speed. The grading strategy helps the test phase narrow down the frequency range under which the die can reliably operate. Because oscillators are not part of the internal circuits, the delays provided are estimates with a certain level of confidence. This confidence level depends on the type, number, and position of ROs across the die.

Several patents deal with PV probing and monitoring techniques.[4,5] For example, the monitor test element group (TEG) proposed by Ukei and Aoyagi consists of an RO and control circuitry.[4] Five TEGs, arranged in the middle and at the four corners of a die, report their signals one by one for PV and manufacturing yield analysis. Samaan's approach disposes ROs opportunistically over an IC chip depending on available layout space.[5] Only one oscillator can operate at a time. An analog frequency wire delivers test data to the counting and monitoring units.

To the best of our knowledge, there has thus far been no analytical justification for numbers, types, and positions of ROs and no systematic approach for quantifying PV metrics. Our approach mainly targets SoC designs implemented in a sub-100-nm process. By using a distributed network of several oscillators per die, we can detect PV changes that collectively cause a measurable frequency shift, Δf, in the output of an RO planted in that

region. Such measurements can also identify the problematic regions and even grade chip or die quality.

Our approach is not suitable as a PV debugging tool, nor can it report the particular causes of parameter variations. Fabs monitor many parameters to characterize a process. However, using our PV fault model, designers can simplify the problem. Rather than tracing each parameter and its effect, our approach records whether within-die PVs (local supply voltage variation, temperature fluctuations, interconnect loading and coupling variations, and so on) have collectively caused any significant change in a circuit's performance or functionality. We accomplish this by carefully designing a distributed RO network across a die, and monitoring and analyzing the behavior of PV sensors (ROs) in the frequency domain. Here, we examine this approach both analytically and empirically. This formalization justifies the use of ROs as a distributed network of sensors for PV testing across a die.

Key metrics

We implant ROs by cascading an odd number of inverters to form a loop. By using an odd-numbered loop, we ensure that the last inverter's output is the inverse of the first inverter's previous input, thus preventing the RO from stabilizing to a steady state. An RO's oscillation frequency is the reciprocal of the total delay of the inverters. That is,

$$f_{RO} = 1/(N_{inv} t_{inv})$$

where N_{inv} is an odd number of inverters, and t_{inv} is one inverter's delay. Hence, we select the frequency by choosing the number of inverters in the loop. RO implementation is a relatively mature topic; details on its various aspects are available elsewhere.[6-8]

Analytical justification

It's possible to trace the effect of critical parameter variations on an inverter by using that inverter's saturation current. We can simplify switching an inverter by assuming that the voltage changes instantaneously and that one of the transistors goes into the saturation region while the other turns off completely. With this assumption, we obtain the following well-known equations, which link an RO's frequency to some process parameters:

$$I_{avg} \approx I_D = (\mu C_{ox}/2)(W/L_{eff})(V_{GS} - V_{TH})^2(1 + \lambda V_{DS})$$
$$t_{inv} = (V_{DD} C_{Load})/I_{avg}$$
$$f_{RO} = 1/(N_{inv} t_{inv})$$

where I_{avg} is the average saturation current, I_D is the drain current in the inverter, μ is the mobility of carriers, C_{ox} is the oxide capacitance, W is the channel width, L_{eff} is the effective channel length, V_{GS} is the gate-to-source voltage, V_{TH} is the threshold voltage, λ is the channel length modulation parameter, V_{DS} is the drain-to-source voltage, V_{DD} is the supply voltage, and C_{Load} is the load capacitance. Combining these three equations and substituting $\lambda \propto 1/L_{eff}$ and $C_{ox} = (\varepsilon W L_{eff})/T_{ox}$, we get

$$f_{RO} \approx \frac{1}{N_{inv} V_{DD} C_{Load}} \left(\frac{\mu \varepsilon W^2}{2T_{ox}}\right)\left(V_{GS} - V_{TH}\right)^2 \left[1 + \left(\frac{K}{L_{eff}}\right) V_{DS}\right] \quad (1)$$

where ε is the oxide permittivity, and the value of factor K is chosen such that K/L_{eff} approximates λ. Thus, Equation 1 is a first-order approximation of the relationship between current, load capacitance, and RO frequency. Although it's possible to analytically trace frequency variation caused by contributing factors (for example, using a derivative such as $\partial f_{RO}/\partial V_{TH}$), the result is predictably inaccurate, mainly for the following reasons: First, the formula is an approximation and cannot accurately capture the complex relationships among all factors. In fact, the metrics in Equation 1 (T_{ox}, V_{TH}, L_{eff}, and so on) are just a few among a large set of factors on which a MOSFET's current depends. Second, the factors are interdependent. For example, V_{TH} has a slight dependency on L_{eff}; and the variation of T_{ox} in a region affects not only one inverter's current but also the next inverter's gate capacitance. Third, the location, affected factors, and magnitude of variations are all highly unpredictable.

Despite ambiguities in the exact variation, it's reasonable to assume that, in general, PVs collectively cause a measurable frequency shift in an RO's output. We can verify this assumption both analytically (for instance, using Equation 1) and empirically (using simulation tools such as HSpice). Like other fault models, our PV fault model rests on a simplified scenario so that using the test mechanism will be practical. Obviously, there is no guarantee that every variation will cause a measurable Δf. In such cases, we assume that PV does not affect the surrounding circuitry either. Our mechanism doesn't rely on absolute frequency f_{RO} or delay t_{inv} values. Instead, it depends on frequency shift Δf and delay changes Δt, which the existing tools can measure and trace. Observing the PV fault model and the reflection of PV on Δf, we conclude that ROs function very well as PV sensors. Moreover, it's easy to replicate ROs to collect more data and achieve higher precision.

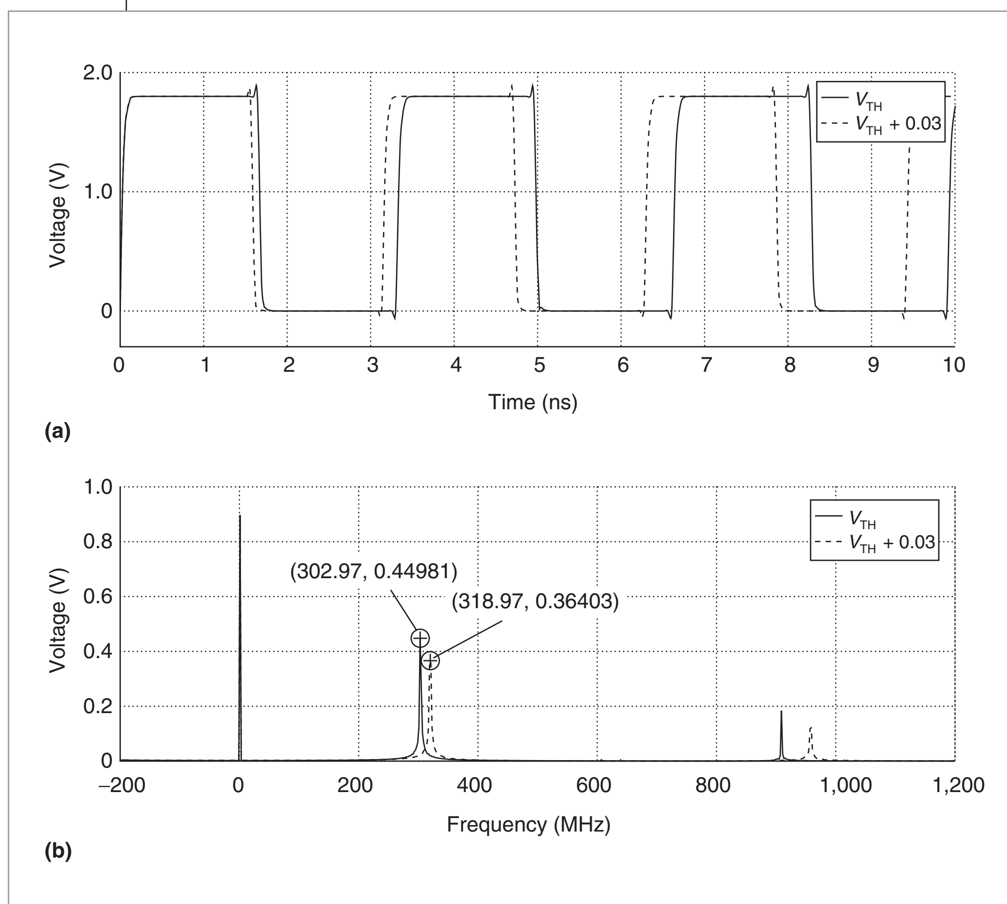


Figure 1. Effect of threshold voltage V_{TH} on the behavior of a 41-stage ring oscillator (RO) in the time (a) and frequency (b) domains. The solid and broken lines show the original signal and the signal in the presence of variation, respectively.

PV detection in the frequency domain

There are two main reasons for using the frequency domain in our methodology. First, the frequency domain provides an efficient, low-cost mechanism for combining RO outputs. Measuring frequency using counters (for example, counting the zero crossings) might be sufficient when there is one or a few ROs. However, in our method we advocate implanting tens of ROs on large dies. Frequency domain analysis is more efficient in this case because it allows compacting RO signals without using analog channels (wires) or decoupling the frequencies to identify problematic regions. As we show later, the concept of weighted addition serves to combine n RO outputs into $\log_2 n$ (or less) signals, and we process the resulting signal through fast Fourier transform (FFT) analysis to obtain the individual harmonics. The key feature of frequency domain and FFT analysis is that combining well-designed RO signals into $\log_2 n$ (or less) signals doesn't change their main harmonics. Such delivery and processing would be far more difficult and more costly in the time domain.

Second, the frequency domain enables easier separation and identification of permanent and intermittent noise. Our proposed test architecture embeds the PV test information in the RO output signals. We're interested in analyzing the shift of the ROs' main harmonics. Intermittent problems such as overshoots, coupling noise, IR drop, and excessive delay are reflected in the higher harmonics. Such common noise is usually present in very high frequencies because the noise often repeats several times over a signal period. Thus, in the frequency domain, the main harmonics are far from the noise harmonics, so it's easy to separate the two.[9]

Behavior of a PV sensor

On-chip sensors provide real parameters that the system truly experiences. In our methodology, each RO combines all the variation effects in its surrounding regions into its output signal's frequency. We demonstrated this practice by observing a 41-stage oscillator's behavior when there was V_{TH} variation. We obtained simulation results from HSpice using TSMC 180-nm technology.

Figure 1a shows the effect of inverters in the time domain on V_{TH} deviation (± 0.03 V, which is almost 10% variation). Figure 1b gives the corresponding frequency domain signal. As the figure shows, the main harmonic of the signals (303 MHz and 319 MHz) are clearly differentiable. (The main harmonic is the one we are interested in throughout this work.) A digital signal (square wave) provides several other harmonics, which are useful for signal skew or signal integrity measurements. Significant power (the largest peak)

is concentrated at the zero frequency—also called DC bias because of the voltage swing from 0 to V_{DD}, instead of from $-V_{DD}$ to V_{DD}, but we ignore DC bias in our work.

Noise insensitivity

Ideally, you should analyze a sensor's output at the point of sensing to reduce noise, but this is impractical when the sensors are deeply buried inside a die. One advantage of using frequency domain analysis is its insensitivity to noise. The environment or interconnect noise in digital systems (for example, overshoots, oscillations, crosstalk, or delays) don't affect the main test information (FFT peaks)—thus facilitating the transmission of these signals either on or off chip, with only subtle distortions. Because the noise (overshoots or oscillations) occurs several times over a signal period, it contributes to the frequency domain only at high frequencies. The magnitude of oscillation caused by noise is small compared to that of the original signal; hence, its contribution will also be small. Delay in a signal adds phase in the frequency domain. Thus, if the signal transmitted from a sensor is delayed (shifted in time)—for example, due to a long interconnect—the magnitude spectrum still remains the same.

When a die is operating, switching noise affects the power supply, which in turn affects RO frequencies. According to our definition of a PV fault model, we're interested in variation of any factor that affects performance. Our architecture treats performance degradation from switching noise as equivalent to the PV effect. We haven't attempted to separate or identify the causes. However, you could run a separate test with clocks off (and thus no switching noise) to measure the PV fault's impact on RO frequency shifts.

Test architecture

The basic PV test architecture consists of some ROs as PV sensors and an adder as a compactor, arranged on a die. PV sensors, distributed across the die or wafer, don't interfere with the chip's normal behavior. Several sensors are sprinkled (randomly assigned) across the die so that the test architecture samples PV with a certain confidence level. A compactor such as an adder efficiently combines the oscillator outputs and delivers them to the frequency analysis point. This analysis point can be on or off chip. However, throughout this article, we assume the latter. There are at least three sensor parameters that we must carefully choose: type, number, and position.

Type

Generating high-frequency signals from ROs, mixing them to save interconnects, delivering them to the observation point, and monitoring them using devices puts a limit on the frequency. Also, frequency is inversely proportional to time: $f_{RO} = 1/(N_{inv}t_{inv})$. This nonlinear relationship between time and frequency causes the frequency shift of different ROs to vary differently with the same PV.

Assuming minimum-size inverters, an RO's type depends on its number of inverters. We find the choices for various RO types, M, through the following systematic procedure:

- Choose f_{max} based on the limits of the monitoring device and mechanism (for example, the limits of the probing or load board).
- For a given f_{max}, simulate an RO running at that speed and apply the maximum expected variation based on statistical observation to determine Δf_{max}.
- Use f_{max} and Δf_{max} to determine Δt from the $f_{RO} = 1/t_{inv}$ curve.
- Find Δf_{min} based on the monitoring device's resolution. This is the minimum frequency variation that is reliably measurable.
- Map Δf_{min} and Δt together on the $f_{RO} = 1/t_{inv}$ curve to determine f_{min}.

Figure 2 illustrates the relationship of these factors. For example, an oscillator with $f_{max} = 1.2$ GHz could experience a maximum frequency variation of $\Delta f_{max} = 250$ MHz (for instance, due to a 25% change in V_{TH}). Such a frequency shift corresponds to $\Delta t = 0.5$ ns. For a monitoring resolution of $\Delta f_{min} = 35$ MHz, we obtain $f_{min} = 200$ MHz, as the figure shows.

Using these factors, we can determine the number M of various RO types to use. As Figure 2 shows, we must find f_{min} beyond which Δf can no longer be accurately measured. Given a constant Δt, we can express the bandwidth of the highest and lowest oscillator as

$$\Delta t = 1/f_{min} - 1/(f_{min} + \Delta f_{min}) = 1/(f_{max} - \Delta f_{max}) - 1/f_{max}$$

After algebraic manipulation, we can write this equation as

$$f_{max}/f_{min} = [\Delta f_{max}(f_{min} + \Delta f_{max})]/[\Delta f_{min}(f_{max} + \Delta f_{min})]$$

Assuming the values of Δf_{min} and Δf_{max} are far smaller than f_{min} and f_{max}, we can simplify their relationship to

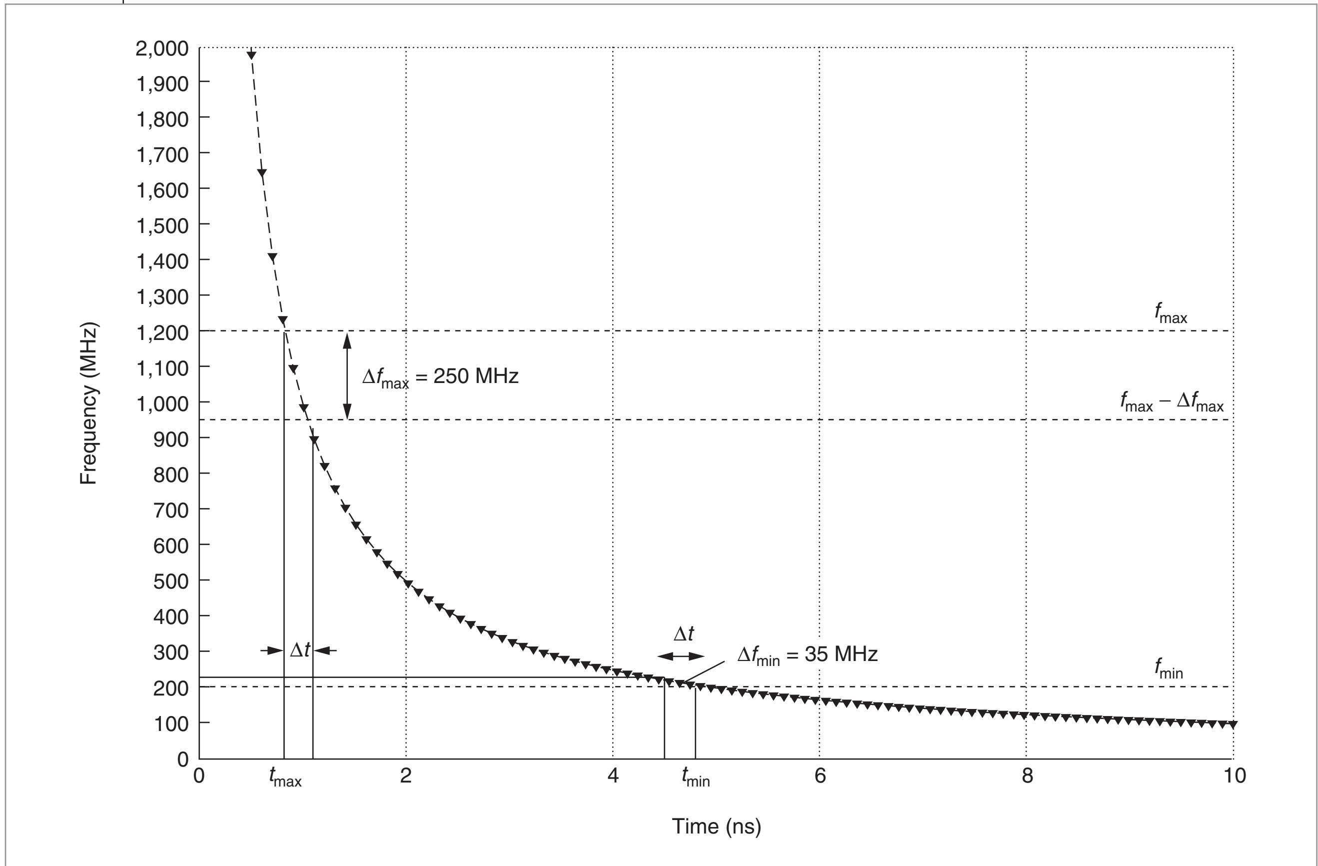


Figure 2. Nonlinear relationship between time delay and frequency variation.

$$f_{max} / f_{min} \approx \sqrt{\Delta f_{max} / \Delta f_{min}}$$

We can now determine the number of different RO types by dividing the time frame $[t_{max}, t_{min}]$ into M equal Δt intervals:

$$M \leq \frac{t_{min} - t_{max}}{\Delta t} = \frac{t_{max}}{\Delta t}\left(\frac{t_{min}}{t_{max}} - 1\right) \leq \left\lfloor \frac{1}{f_{max}\Delta t}\left(\sqrt{\frac{\Delta f_{max}}{\Delta f_{min}}} - 1\right)\right\rfloor \quad (2)$$

Note that M is only an upper bound and not an exact requirement.

As a second example, suppose that for $f_{max} = 1.2$ GHz and a maximum of 15% variation on V_{TH}, we found $\Delta f_{max} \approx 90$ MHz and $\Delta t \approx 0.5$ ns. For a resolution of $\Delta f_{min} \approx 5$ MHz, we could combine these factors using Equation 2. We'd get $M \leq 5$, meaning we could use up to $M = 5$ types of ROs.

Number

We can estimate the number of oscillators for grading a die from a probability perspective—that is, fault sampling. We randomly choose a subset of faults (the fault sample) for simulation, and we estimate the fault coverage for the complete set by simulating this subset with a set of vectors. Similarly, by using a few grids (RO positions), we can sample the PV with a certain level of confidence without actually measuring the PV across all regions on a die. But first, let's define a few modeling and sampling terms to normalize the area metrics:

- A_0 is the unit (grid) area corresponding to the unit RO.
- A_i is the area of RO_i (oscillator of type i).
- m_i is the normalized area of RO_i, which is $\lceil A_i/A_0 \rceil$.
- n_i is the number of ROs of type RO_i.
- A_{die} is the area of the die without ROs.
- N_p is the total number of grids on a die, which is $\lceil A_{die}/A_0 \rceil + \Sigma n_i m_i$.
- N_s is the number of grids covered by all ROs, $\Sigma n_i m_i$.
- C is the true PV coverage, which is the number of grids with acceptable PV divided by N_p.
- c is the estimated PV coverage with only N_s grids sampled.

- x is an estimate of c.
- Δ^2 is the estimation error, or $(|C-x|)^2$.

Theoretically, we can obtain C only if every grid is ideally analyzed by a PV sensor. We can derive the probability of obtaining an estimate of coverage x as follows: We divide the number of ways of obtaining coverage x by choosing N_s samples at a time by the number of ways of choosing N_s samples, given N_p. Thus, we get the probability density function

$$P(x) = Prob(c = x) = \frac{\binom{CN_p}{xN_s}\binom{(1-C)N_p}{(1-x)N_s}}{\binom{N_p}{N_s}} \tag{3}$$

Equation 3 is known as a *hypergeometric* probability density function because x can take values that are only a multiple of $1/N_s$. If A_0 (corresponding to a unit RO) forms a small area, then N_s is large and we can approximate this function using a Gaussian random variable with the distribution shown in the following equation (with average value μ and standard deviation σ derived from Equation 3):

$$P(x) = Prob(x \le c \le x + dx) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-C)^2}{2\sigma^2}}$$

$$\mu = C$$

$$\sigma^2 = \frac{C(1-C)}{N_s}(1-\frac{N_s}{N_p}) \approx \frac{C(1-C)}{N_s}$$

Even with a small N_s, the hypergeometric probability assumes a shape that is Gaussian but discrete; hence, we can use the same 3σ concept. When A_0 is too small compared to an oscillator's size, the initial assumption of choosing N_s samples doesn't hold, because several A_0 units are necessary to cover the area that one oscillator requires. Thus, certain choices will narrow down an oscillator's grids to surrounding areas. We choose A_0 close to the size of a standard oscillator whose frequency shift Δf is measurable. The areas of other oscillators are practically within an order of magnitude (for example, 3 to 5 times) of A_0.

For a Gaussian distribution, the probability of x being in the range of $C - 3\sigma \le x \le C + 3\sigma$ is approximately 0.997—that is, almost certain. Hence, we can assume the value of x is between $C - 3\sigma$ and $C + 3\sigma$. We obtain the sampling error by setting $x - C$ to 3σ, which we can also write as

$$\Delta^2 = (x - C)^2 = (3\sigma)^2 = 9[C(1 - C)/N_s] \tag{4}$$

Solving the quadratic function in this equation for C, we get

$$C = \left[\frac{2x + \frac{9}{N_s}}{2(1+\frac{9}{N_s})}\right] \pm \sqrt{\left[\frac{2x + \frac{9}{N_s}}{2(1+\frac{9}{N_s})}\right]^2 - \frac{x^2}{(1+\frac{9}{N_s})}} \tag{5}$$

Equation 5 provides two solutions for C, giving the upper and lower bounds of coverage. We make no assumption regarding the value of N_s, to make sure the relationship remains valid regardless of sample size. By plugging in the maximum C value to Equation 4, we can calculate the maximum error Δ^2, which is plotted in Figure 3. This figure shows different error curves as a function of N_s for a few fixed values of x. Such a plot helps estimate the number of samples N_s necessary to give a certain level of confidence, given an approximate value for die grade x. For example, for a die and process with a statistical defect coverage of $x = 0.8$, we need $N_s = 10$ or greater ROs (sampling grids) if we intend to test PV with at most a 3% error ($\Delta^2 \approx 0.03$).

Position

The placement and number of sensors depends on the user and the desired confidence level in the measurements. As the number of sensors increases (because of chip size or demand for higher accuracy), a compactor can reduce interconnect cost. Finally, the accumulated PV test information goes to the off-chip FFT analyzer to interpret the data, perform limited diagnosis, and make the final call on the pass/fail test result.

We can easily insert sensors into a system's layout at random locations by providing uniform positioning during floorplanning. Then, using the automatic layout and placement tools, we decide their final positions. Because PV faults can occur anywhere on a die, from a fault-sampling perspective the RO positions are quite random with respect to PV fault occurrence. Fortunately, almost all layout and placement tools can efficiently (randomly with respect to grids) and automatically position PV-sensing circuitry such as oscillators on a chip. The result is a layout in which the PV test circuitry—ROs and adder compactors—is implanted.

Optimization and implementation

For on-chip test analysis, we use an off-chip FFT analyzer such as Matlab. Although we do not pursue it here,

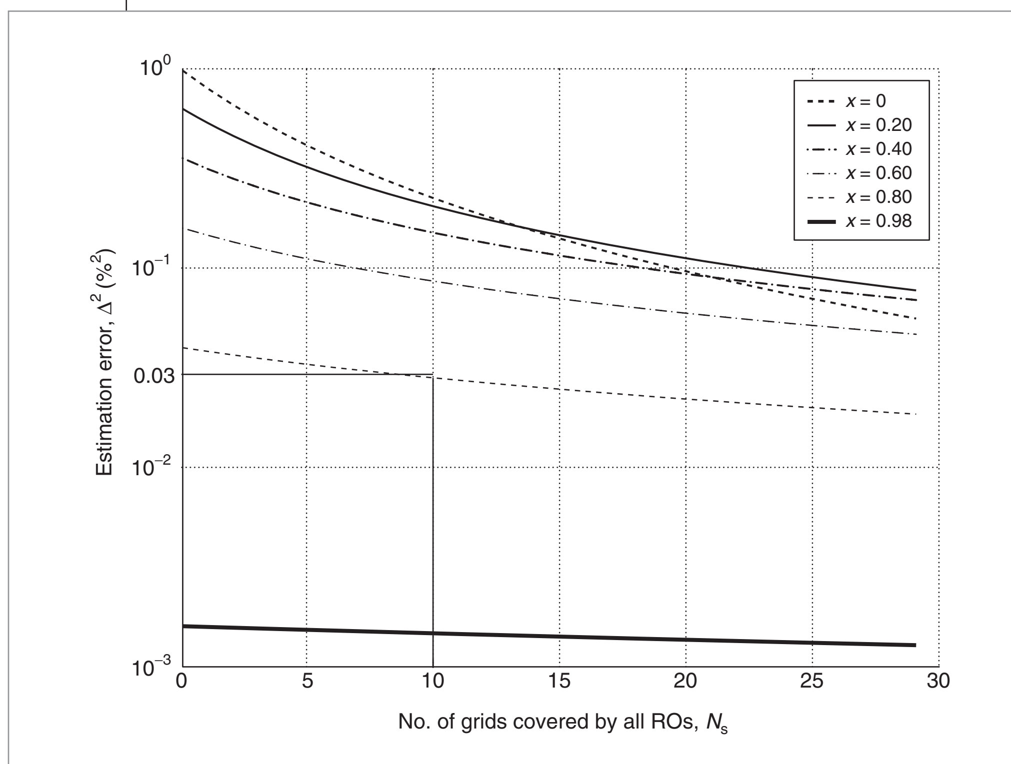


Figure 3. Estimation error as a function of number of grids N_s.

different frequencies, the same frequency, or a combination of the two. A few clarifications about these arrangements follow.

First, using oscillators of different frequencies lets the user uniquely identify each oscillator. Unique identification facilitates a level of diagnosis or even adaptation (that is, PV-aware circuits)—for example, letting the user detect that a particular region containing a video encoder might have slower than typical gates and reconfigure it to run at a lower data rate to prevent errors. The cost of unique frequencies comes into play in data-processing units such as filters or compactors that operate in the entire frequency range. Also, processing circuits work at several hundred megahertz. Hence, there is a limitation on how many unique input frequencies you can generate within that bandwidth.

Second, running all oscillators at the same frequency solves the bandwidth limit problem by eliminating the need to divide the bandwidth among sensors. The drawback is that you cannot identify a PV region if you use a single channel to communicate with the tester. As a result, you can use this technique only for a pass/fail test. In that case, the test will be based on observing frequency deviation of oscillators that ideally have identical spectrums.

Finally, most applications choose the majority of N_s regions to run ROs at the same frequency. Only critical regions receive oscillators at a different frequency. Such a hybrid scheme provides flexibility from both the sensing and detection perspectives.

using an on-chip FFT analyzer or a frequency detector is also possible. In that case, we would use a signal processor for demodulation, filtering, and so on, to perform the FFT-based test analysis on chip.

Bandwidth planning

Equation 2 gave the analytical upper limit of M. To determine the actual number of RO types, it's essential that the RO frequencies do not overlap beyond recognition. To prevent data loss, sensor design must follow certain guidelines. An RO's control parameter is its frequency, which we control by determining the number of stages (inverters) forming the loop and the sizes of the transistors. Throughout our work, we used minimum-size inverters and modified only the number of stages. While choosing the oscillators' frequencies, we must maintain a certain frequency separation between the first harmonic of any two oscillators. This gap should be the sum of the frequency deviations expected because of the PV in these oscillators. For example, we couldn't use a 300-MHz oscillator expected to vary up to 50 MHz along with a 360-MHz oscillator expected to vary 30 MHz, because both oscillators can vary at similar values (with overlapping frequencies of 330 MHz to 350 MHz). Depending on the application, we must choose oscillators to run at

Bandwidth and power

An RO output is a periodic signal that distributes finite power among all harmonics. This implies that the more harmonics there are, the more unique oscillators the adder will combine. Each harmonic's magnitude must decrease because the power is now redistributed

among more harmonics. Therefore, depending on the detector's sensitivity (resolution), we can define a minimum threshold for detectable magnitude. Given the detectable threshold, we can determine the number of unique main harmonics (unique sensors) that an adder can combine.

Data compaction

Providing good PV coverage requires many sensors. Each oscillator requires an interconnect to deliver information to an observation point or a detector to analyze the results. An alternative approach is to reduce the bits by compacting the data without significant data loss. The following equation, where $S(f) = F\{s(t)\}$ is the Fourier transform of signal $s(t)$, shows the basic superposition property of FFTs; this property also relates the time and frequency domains:

$$F\{s_1(t) + \ldots + s_n(t)\} = F\{s_1(t)\} + \ldots F\{s_n(t)\} = S_1(f) + \ldots + S_n(f)$$

If the main harmonic of the signals (oscillator outputs) are sufficiently apart, summing the signals in the frequency domain preserves the information.

Given n 1-bit digital signals $s_1(t), \ldots, s_n(t)$, we can combine them into $\log_2(n)$ bits—for example, $y_1(t), \ldots, y_{\log(n)}(t)$—by counting the total number of 1s. This addition provides a lossless compaction from n to $\log_2(n)$ because the adder can operate at a frequency higher than the input frequencies. Hence, in a way, the adder restricts the bandwidth and thus the number of unique sensors that we can employ. We can use a 1-adder (or 1s adder), an adder that adds n 1-bit numbers to produce a binary output. As n grows, the adder's maximum operating frequency deteriorates, reducing the usable bandwidth. For large n, hierarchical compaction is more efficient. In other words, we first divide the n bits into several groups and combine each group of 1s using the 1-adder. Then, at further levels, we combine the output of all the 1-adders, using normal binary adders (b-adders) for second-level compaction. A hierarchical approach can use pipelining to reduce delay. In the pipelined approach, we must trade off area for timing because registers are required at intermediate stages. In the hierarchical approach, each stage must perform more quickly than the input signal, but the end-to-end path can be slower. This doesn't change the frequency characteristics. There will be a delay in the phase spectrum but no effect in the magnitude spectrum and thus no frequency distortion.

Compaction doesn't change the frequency domain behavior or the PV information collected within the signals. In fact, it's possible to combine these $\log_2(n)$ digital signals into a single analog signal using a digital-to-analog converter (DAC). More specifically, the following equation links analog (mathematical) and digital implementations:

$$\sum_{i=1}^{n} s_i(t) \rightarrow \sum_{j=0}^{\lceil \log_2(n) \rceil} [2^j y_j(t)] \quad (6)$$

The left side of this equation shows amplitude-based analog addition of $s_i(t)$. The right side reconstructs this addition from a 2^j-weighted sum of $y_i(t)$, which we can realize using a DAC. However, for design simplicity, we don't use a DAC. After data compaction, we can deliver $\log_2(n)$ bits to an off-chip analysis unit to compare the frequency deviations with the expected values. The simplest, though not necessarily fastest, approach is to use software such as Matlab to compute the resultant signal's FFT, and then compare this FFT with a given signature.

Figure 4 shows the generic PV test architecture. The layout under PV test can be a full die or a portion of a die such as a sensitive core. We determine the types, numbers, and positions of ROs using the approach outlined earlier. In this figure, each RO symbolically represents one grid out of a total of $9 \times 9 = 81$ grids. Note that according to fault-sampling metrics, the smallest RO has almost the same layout size as one grid. Of course, larger ROs can occupy multiple grids. The layout and placement generation tools position all ROs automatically, using internal optimization methods. However, from the PV test perspective, the layout tools distribute ROs randomly across the die, and the fault-sampling statistics such as coverage and confidence level still hold.

Example

The following example demonstrates the compaction concept. We used three ROs with 39, 51, and 73 inverters as sensors. (The number of inverters and the exact frequency peak are irrelevant to this example.) We arbitrarily chose these inverters to place the frequencies between 150 MHz and 300 MHz, with a difference of about 50 MHz to 60 MHz between them. We then passed these oscillators' outputs through a 1-adder to compact them into 2 bits. Figure 5 verifies Equation 6 by showing almost identical curves and accuracy for the two cases—that is, processing all three signals individually $\Sigma_i S_i(f)$, and processing two bits (signals) from

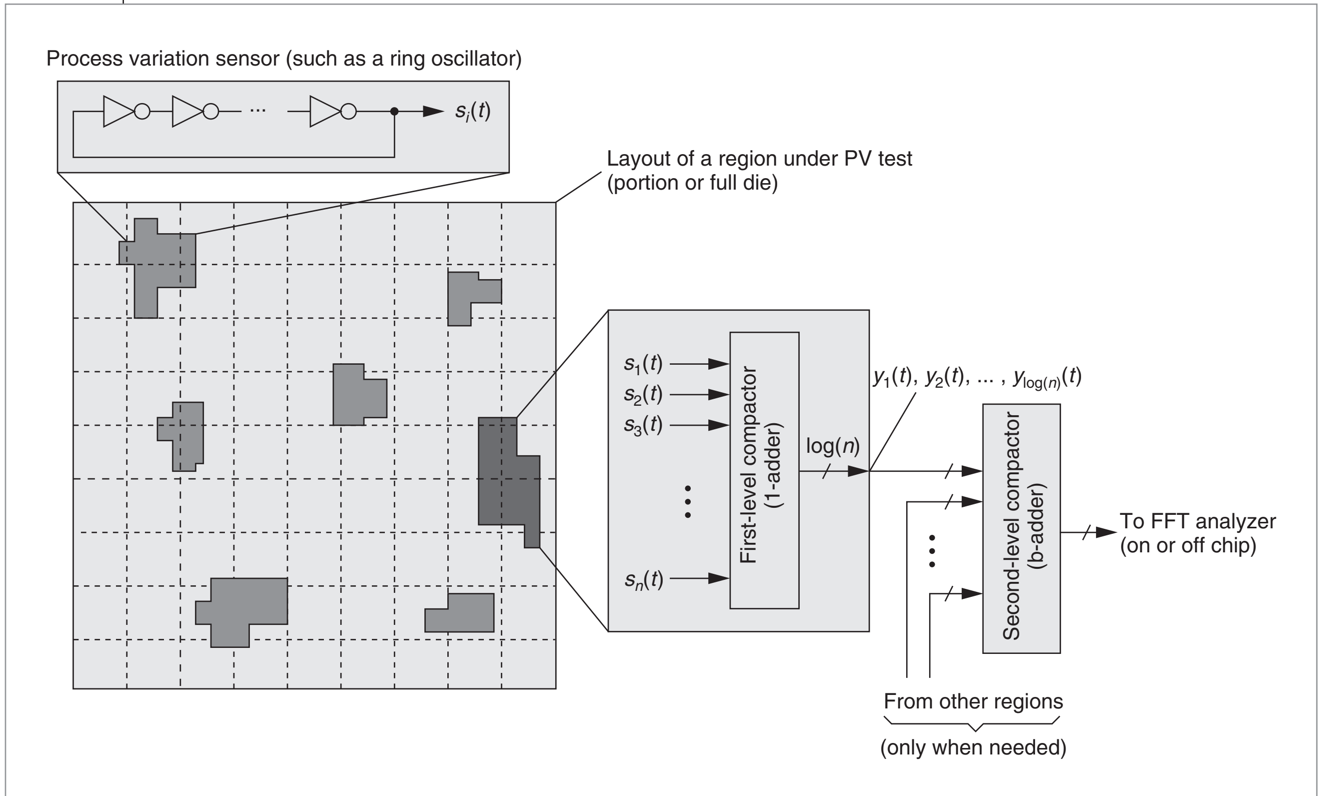


Figure 4. Basic PV test architecture with ROs and compactors.

the 1-adder compactor and then combining them using 2^j weighted sums $\Sigma_j[2^j Y_j(\mathrm{f})]$.

Limitations of our methodology

There are a few limitations in applying our methodology. Theoretically, some PV parameters could mask one another, so the frequency of some RO outputs might not change even in the presence of PV faults. This phenomenon is similar to multiple stuck-at faults masking one another, and it reflects our approach's limited capability. Also, our method is based on sampling RO outputs over a short period of time. Any PV not reflected in that short time will not be captured.

As for the accuracy of our PV measurement, the frequency shift of ROs (Δf) is not due to PV alone. Many other environmental parameters (such as temperature and power voltage fluctuations) can also contribute to this frequency shift. That is,

$$\Delta f = \Delta f_{\text{env}} + \Delta f_{\text{PV}}$$

However, in the final test analysis, we can ignore Δf_{env} as a fixed offset by making a conservative assumption that the environmental parameters affect the entire circuit uniformly. Finally, for slow external ATE that cannot directly read or analyze high-speed RO outputs, extra circuitry is necessary to sample RO outputs above the Nyquist rate, store them in a memory, and deliver them to the ATE at a low speed.

Experimentation

We have implemented our architecture with the TSMC 180-nm technology process using Synopsys and Cadence tools. Even though the 180-nm process shows less variation than 90-nm-or-below technologies, the concepts still hold true. We implemented adders of different sizes to examine compaction performance. As Table 1 shows, adder performance degrades (bandwidth decreases) as the number of inputs increases, forcing us to use more oscillators of the same frequency.

Table 2 gives the results of an optional approach. Pipelining the adder structure can significantly improve performance. The trade-off is uniqueness versus area. The hierarchical design provides wider bandwidth. Hence, we can combine oscillators of several frequencies, providing the benefit of pinpointing the area of variation.

Because RO frequency doesn't vary linearly with variation, the user can choose oscillators according to the

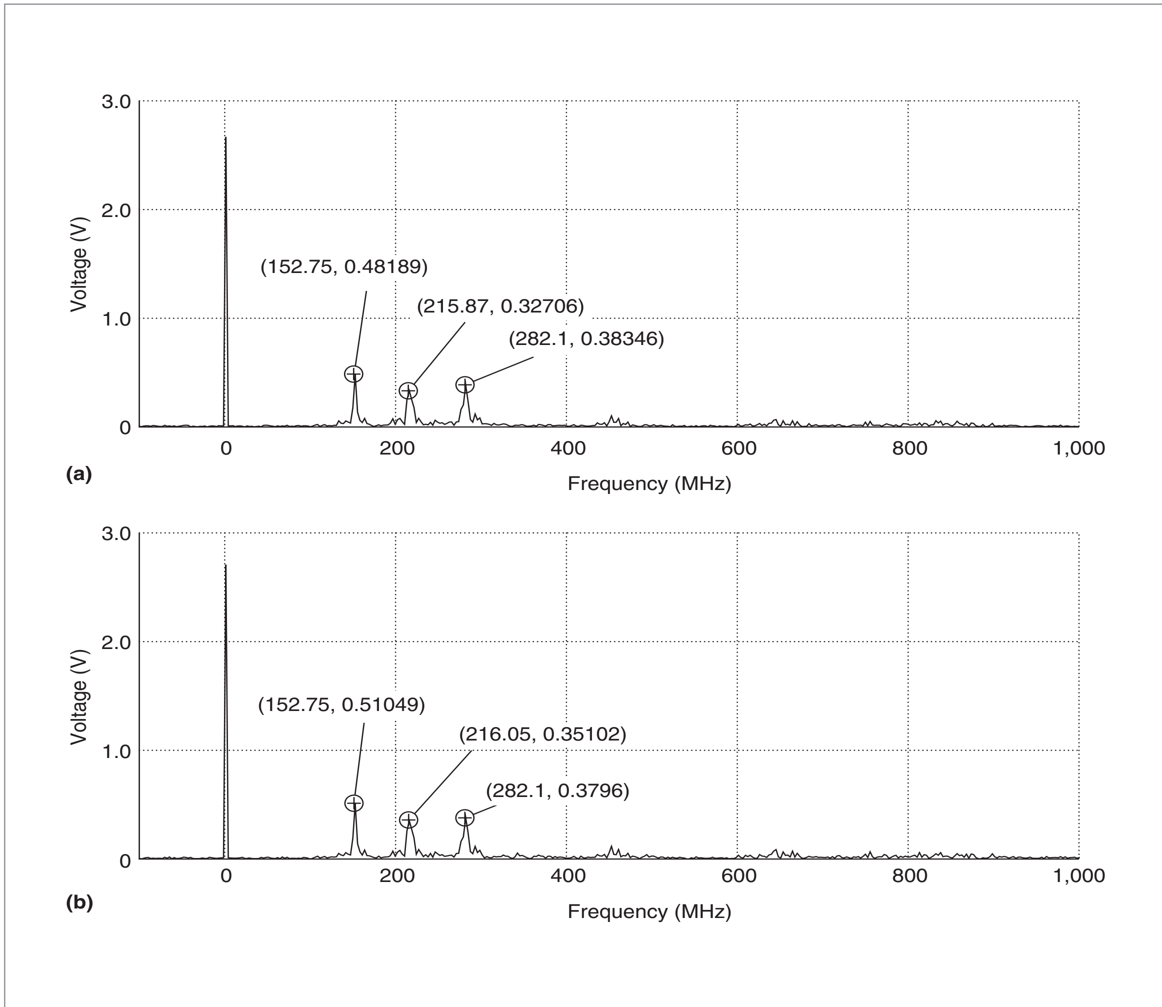


Figure 5. Fast Fourier transform (FFT) for three ROs' output signals: processed individually and directly added (a), and compacted into 2 bits through a 1-adder and then combined using weighted sums (b).

sensitivity (resolution) to PV required at different regions on the die. At the circuit level, the ideal $f_{RO} = 1/(N_{inv}t_{inv})$ relationship does not accurately hold, because of parasitics' dependence on operating frequency. As Table 3 shows, nonlinear behavior decreases as frequency increases. This is not surprising, because at high

Table 1. Comparison of different-size 1-adders.

No. of inputs	Area (no. of NAND gates)	Delay (ns)	f_{max} (MHz)
3	8	0.40	2,500
7	29	1.05	952
15	130	2.95	339
31	774	6.85	146

Table 2. Various implementations of a 15-bit compactor.

No. of stages	No. of 1-adders	No. of b-adders	No. of flip-flops	Area (no. of NAND gates)	f_{max} (MHz)
1	1 (15-bit)	0	0	130	300
2	2 (8-bit)	1 (3-bit)	6	184	900
3	4 (4-bit)	3 (2-bit)	14	223	950

frequencies parasitic capacitances contribute less impedance, providing a reduction in nonlinearity. This phenomenon lets us use oscillators at frequencies of around 1 GHz without worrying that the variation will consume several hundred megahertz of bandwidth.

To show the effect of RO usage and fault sampling on the accuracy of PV testing, we performed another series of simulations. We chose benchmark circuit c6288 from the 1985 IEEE International Symposium on Circuits and Systems (ISCAS) suite. This circuit has 2,416 gates, 32 inputs, and 32 outputs. Here is how we proceeded:

- The layout size of c6288 was A_{die} = 19,940 μm^2. We assumed that the base (unit) RO whose frequency change due to variation is detectable had A_0 = 195 μm^2, which was also the grid size.
- Applying our RO selection strategy, with $\Delta t \approx 0.5$ ns, f_{max} = 1.2 GHz, $\Delta f_{max} \approx 90$ MHz, and $\Delta f_{min} \approx 5$ MHz, we obtained $M \leq 5$, and we chose $M = 3$ RO types. Their sizes were A_1 = 195 μm^2 ($m_1 = 1$), A_2 = 384 μm^2 ($m_2 = 2$), and A_3 = 640 μm^2 ($m_3 = 3$).
- Using the fault-sampling approach elaborated earlier and assuming we wanted the 3σ sampling error not to exceed 3%, we found that $N_p = \lceil A_{die} \rceil \approx 100$, and N_s =10 (determined from Figure 3 for x = 0.8 and $\Delta^2 = 0.03$). Therefore, we used three type-1 ROs ($n_1 = 3$), three type-2 ROs ($n_2 = 2$), and one type-3 RO ($n_3 = 1$), to cover

$$N_S = \sum_{i=1}^{3} n_i m_i = 10$$

- We positioned the ROs manually in the floorplan, while trying to be uniform, and we generated the layout using the Cadence Encounter toolset.
- We chose 5% to 25% of the grids ($0.75 \leq C \leq 0.95$) randomly in the final layout to have 5% to 15% V_{TH} variation, and we repeated this procedure 100 times. The x and $\overline{\Delta f}$ columns in Table 4 list the number of faulty grids found and the average frequency shift.

Because we chose the grids with PV faults, we can consider the first column that indicates what percentage of grids are subject to V_{TH} variation as true coverage C. The position of oscillators is fixed in the layout. A randomly chosen grid might affect none, a portion, or the entire RO. Estimated coverage x is the percentage of grids out of N_s = 10 that are portions of grids that ROs occupy. In a way, the closeness of x and C indicate the accuracy of the fault-sampling method in the PV test. The data in Table 4 confirms the relationship of the factors plotted in Figure 3—that is, the error increases for fixed N_s and decreasing x. In all cases, however, the error is predictably bounded:

$$\Delta^2 = (|C - x|)^2 \leq 0.03$$

In practice, we first determine the acceptance (grading) levels of chips for a given product and process. We do this by conducting thorough testing of several samples of good dies, collecting statistics, and determining acceptable levels—for example, Δf_A and Δf_B. Then the PV test simply compares the overall frequency shift of a chip under test

$$\overline{\Delta f} = \frac{1}{6}\sum_{i=1}^{6} \Delta f_i$$

against these predefined levels to pass/fail or grade them. In a way, Δf represents the overall evaluation metric (grade) for that die. For example, one grading policy can use ranges of $\overline{\Delta f} \leq \Delta f_A$, $\Delta f_A \leq \overline{\Delta f} \leq \Delta f_B$, and $\Delta f_B \leq \overline{\Delta f}$, for A (pass, good quality), B (pass, low quality), and

Table 3. Sensitivity of ring oscillators (Δf) to variation of threshold voltage (ΔV_{TH}), with V_{TH} equal to 0.370 and –0.384 for NMOS and PMOS transistors, respectively, in the simulation.

		Δf (MHz) for ΔV_{TH} of		
No. of inverters	**Original f_{RO} (MHz)**	**5%**	**10%**	**15%**
21	620	27	43	55
41	303	13	22	27
61	208	6	15	18

Table 4. True (C) and estimated (x) coverage and average frequency shift $\overline{\Delta f}$ for different V_{TH} variations.

	ΔV_{TH} = 5%		ΔV_{TH} = 10%		ΔV_{TH} = 15%	
C (%)	**x (%)**	**$\overline{\Delta f}$ (MHz)**	**x (%)**	**$\overline{\Delta f}$ (MHz)**	**x (%)**	**$\overline{\Delta f}$ (MHz)**
95	97.8	4.2	97.1	5.2	96.5	7.9
90	92.1	5.8	93.6	6.9	94.0	12.2
85	89.0	10.0	90.5	15.4	88.3	18.6
75	81.5	14.2	81.7	25.7	82.8	32.0

F (fail), respectively. We can even use collective grades of dies on a wafer to indirectly estimate the yield.

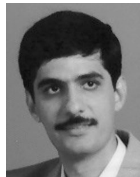
THE FLEXIBILITY of our mechanism facilitates future extensions to other aspects of PV testing. Possible extensions include analyzing higher harmonics to recognize what caused a variation, performing spatial frequency variation as a function of RO location, and adapting parameters such as power supply and speed for a PV-aware circuit. ■

Acknowledgments

This work was supported in part by National Science Foundation Career Award CCR-0130513.

■ References

1. S. Borkar, "Microarchitecture and Design Challenges for Gigascale Integration," *Proc. 37th Ann. Int'l Conf. Microarchitecture* (Micro 37), IEEE CS Press, 2004, pp. 2-3.
2. S. Nassif, D. Boning, and N. Hakim, "The Care and Feeding of Your Statistical Static Timer," *Proc. IEEE/ACM Int'l Conf. Computer Aided Design* (ICCAD 04), IEEE CS Press, 2004, pp. 138-139.
3. M. Hatzilambrou, A. Neureuther, and C. Spanos, "Ring Oscillator Sensitivity to Spatial Process Variation," *Proc. 1st Int'l Workshop Statistical Metrology* (IWSM 96), 1996; http://bcam.berkeley.edu/archive_old/iwsm96-hatz_paper.pdf.
4. T. Ukei and H. Aoyagi, *Monitor TEG Test Circuit*, US Patent 6,239,603, to Kabushiki Kaisha Toshiba, Patent and Trademark Office, 2001.
5. S. Samaan, *Parameter Variation Probing Technique*, US Patent 6,535,013, to Intel Corp., Patent and Trademark Office, 2000.
6. A. Hajimiri, S. Limotyrakis, and T. Lee, "Jitter and Phase Noise in Ring Oscillators," *IEEE J. Solid-State Circuits*, vol. 34, no. 6, June 1999, pp. 790-804.
7. L. Dai and R. Harjani, "A Low-Phase-Noise CMOS Ring Oscillator with Differential Control and Quadrature Outputs," *Proc. 14th Ann. IEEE Int'l ASIC/SOC Conf.*, IEEE Press, 2001, pp. 134-138.
8. M. Grozing, B. Philipp, and M. Berroth, "CMOS Ring Oscillator with Quadrature Outputs and 100 MHz to 3.5 GHz Tuning Range," *Proc. 29th European Solid-State Circuits Conf.* (ESSCIRC 03), IEEE Press, 2003, pp. 679-682.
9. M. Nourani and A. Radhakrishnan, "Modeling and Testing Process Variation in Nanometer CMOS," *Proc. Int'l Test Conf.* (ITC 06), IEEE CS Press, 2006, pp. 7.2.1-7.2.10.



Mehrdad Nourani is an associate professor of electrical engineering at the University of Texas at Dallas. His research interests include SoC testing, signal integrity modeling, and test and application-specific processor architectures. Nourani has a BSc and an MSc in electrical engineering from the University of Tehran and a PhD in computer engineering from Case Western Reserve University. He is a member of the IEEE Computer Society and the ACM SIGDA.

Arun Radhakrishnan is a design engineer in the High-Performance Analog Group at Texas Instruments. He completed the work on the project described in this article while he was a graduate student at the University of Texas at Dallas. His research interests include the development and maintenance of CAD tools related to VLSI design and test. Radhakrishnan has a BSc and an MSc in electrical engineering from the University of Texas at Dallas. He a member of the IEEE.

■ Direct questions and comments about this article to Mehrdad Nourani, University of Texas at Dallas, PO Box 830688-EC33, Richardson, TX 75083-0688; nourani@utdallas.edu.

For further information on this or any other computing topic, visit our Digital Library at http://www.computer.org/publications/dlib.