An Adaptive Approach to Minimize System Level Tests Targeting Low Voltage DVFS Failures

Adit D. Singh
Deptepartment of Electrical and Computer Engineering
Auburn University
Auburn, AL, USA
adsingh@auburn.edu

***Abstract*— Traditional low cost scan based structural tests no longer suffice for delivering acceptable defect levels in many processor SOCs, especially those targeting low power applications. Expensive functional system level tests (SLTs) have become an additional and necessary final test screen. Efforts to eliminate or minimize the use of SLTs have focused on new fault models and improved test generation methods to improve the effectiveness of scan tests. In this paper we argue that given the limitations of scan timing tests, such an approach may not be sufficient to detect all the low voltage failures caused by circuit timing variability that appear to dominate SLT fallout. Instead, we propose an alternate approach for meaningful cost savings that adaptively avoids SLT tests for a subset of the manufactured parts. This is achieved by using parametric and scan tests results from earlier in the test flow to identify low delay variability parts that can avoid SLT with minimal impact on DPPM. Extensive SPICE simulations support the viability of our proposed approach. We also show that such an adaptive test flow is also very well suited to real time optimization during the using machine-learning techniques.**

Keywords—DPPM, system level test, optimization, low voltage failures.

I. INTRODUCTION

While ICs have long been tested using low cost scan based structural tests to screen out defects following manufacture, scan structural tests alone no longer appear sufficiently effective at detecting malfunctioning parts to achieve the quality levels required by many smartphone, notebook and other processor SOC applications. Consequently, expensive System Level Tests (SLTs) are increasing being used as an additional final screen against manufacturing defects, before the parts are shipped for assembly into systems. SLTs involve temporarily mounting the SOC being tested on a test board that closely replicates the target application hardware, including all loading and parasitic electrical parameters at the device interfaces. A wide range of functional tests are then run at operational clock speeds to mimic the full range of anticipated workloads and operating conditions. These tests can take over an order of magnitude longer than traditional scan tests, up to 15 minutes or more. SLTs add a completely new test insertion step in the test flow as shown in Figure 1. The long test time and high cost of the test hardware, both contribute to make SLTs very expensive. Note, however, that while SLTs are effective in screening out a significant

This research is supported in part by the National Science Foundation under grant nos. DEB-1212345

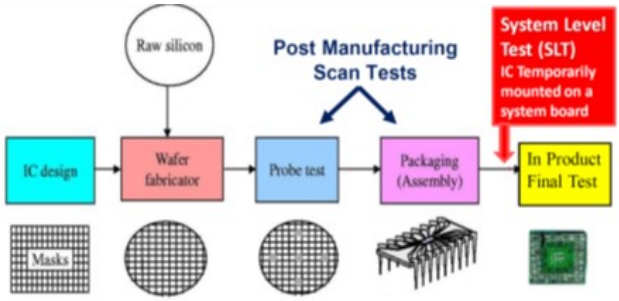

Figure 1: The VLSI Manufacturing and Test Flow

number of additional defective parts, they do not provide a complete test in themselves. Functional tests have long been known to miss many manufacturing defects reliably detected by fault model based scan tests. Also, scan tests applied during wafer probe detect and scrap a large majority of the defective die that is manufactured, and thereby help avoid the wasteful packaging of non-functional parts. These benefits rule out any possibility of eliminating conventional structural testing when functional SLTs are used as a final screen. SLTs therefore imply an added test step and significantly increased test costs, greatly motivating industry to try to eliminate, or at least greatly reduce, the use of SLTs.

Towards this goal, in this paper, we propose an innovative adaptive test strategy that aims to avoid performing SLTs for a significant fraction of the manufactured parts with little or no effect on defect levels. Our approach only selects those SOCs for system level test that have a likelihood of SLT failure exceeding the target threshold. This selection is made based on parametric and scan test results from the earlier test steps. Samples from the deselected parts can also be subject to SLT to ensure that the overall target defect levels (DPPM) are being met in the shipped product. Additionally, this sample data from the deselected parts can also be used in conjunction with the rest of SLT test data for optimizing the setting of selection thresholds for SLT minimization. As we discuss in the paper, such data driven adaptive test methods are particularly well suited to optimization using machine learning methods, where explicit rules for die selection, and the setting of thresholds, are no longer necessary.

The main contributions of this paper include a detailed study, supported by SPICE circuit simulations and recently published industrial data, of how process variability driven low voltage timing failures contribute to SLT fallout. Such failures are

typically experienced in processor SOCs during power saving, low voltage operating modes. This analysis helps us to identify the characteristics of circuits most vulnerable to such failures. These are generally die that display relatively high, although within specification, average threshold voltages due to subtle variations in the manufacturing process. This observation leads to the development of our proposed adaptive methodology for classifying parts using parametric and scan test results. Importantly, once it is validated that there is useful information in the parametric and scan test data that can be used to reliably classify parts for effective adaptive SLT tests, explicit classification rules are not necessary. Real time deep learning techniques, working on the parametric, scan and SLT test data (for the parts that do undergo SLT), can use test results for an individual part from early in the test flow to continuously optimize the selective application of SLT tests so as to achieve desired DPPM levels at minimum test cost.

The key novelty of the work presented here is that while most academic and industrial research is focused on developing improved scan tests using advanced fault models [1-6] (e.g. Cell Aware, Gate Exhaustive, Small Delay Defect, Timing Aware Cell Aware, etc.), we take the view that it may be impractical to detect enough of the SLT fallout using enhanced scan tests alone to eliminate the need for SLTs. This is because, while improved fault models can help further minimize test escapes from hard defects, scan tests remain challenged by timing failures, particularly in low voltage operation. Instead, we focus here on an adaptive approach for minimizing SLT costs by selectively skipping SLT for a significant fraction of the manufactured parts that are unlikely to fail the test. We show that there is information in the parametric and scan tests performed prior to SLTs that can probabilistically predict SLT timing failures. We then present a methodology to efficiently exploit this information for adaptive SLT optimization using machine learning techniques.

The rest of the paper is organized as follows. The next Section provides an overview of the low voltage DVFS failures targeted in this research. We show, using recently published industrial data, that such failures contribute disproportionately to SLT fails. Section 3 surveys other efforts towards eliminating SLTs, mainly by improving the coverage of scan tests towards detecting localized defects. In Section 4 we motivate our alternative adaptive approach for minimizing SLTs by analyzing low voltage delay variability driven failures in detail, with the help of circuit simulations. Of particular interest here is the random variation in device (transistor) parameters, whose timing impact is greatly amplified at low voltages. Section 5 presents and validates our proposed adaptive system level test strategy. We conclude in Section 6.

II. System Level Tests and Low Voltage DVFS Timing Failures

The key to the success of the proposed adaptive test strategy clearly lies in developing an effective selection process for parts that are likely to fail SLT, and therefore must be SLT tested; the rest of the parts with an acceptably low SLT failure probability can bypass the SLTs. For this, it is important to understand the key common characteristics of unique SLT failures, namely faults that pass all other tests, including the scan tests, and only fail SLTs. Until recently, these have remained the subject of much speculation because of the lack of hard data.

A. Analyzing SLT failures

For faults detected by scan tests, the failing input and output scan patterns are known. This is extremely helpful in diagnosing and locating the fault on the die. Fault location is further enhanced if there are multiple failing test patterns. Functional SLT failures, on the other hand, are extremely difficult to accurately diagnose and pin down to a specific location within the die. This is because of the lack visibility into the internal circuit state associated with the failure. Since a functional failure is generally only detected when an error propagates to a SOC output, which can be thousands of cycles after the fault is activated, even the first failing clock cycle often remains unknown. This greatly limits the ability to pinpoint the source and type of SLT failures through physical failure analysis (PFA), or, for delay faults, noninvasive timing measurement.

However, recently at ITC 2018 [1], Intel presented results from volume production test experiments on a processor SOC implemented in a 14nm FinFET process that also included scan test results for 156 observed SLT fails. These confirm what has been a growing consensus within industry that the SLT fails are broadly of two types: (1) "hard" fails that escape traditional scan tests due to incomplete coverage and are detected by the SLTs. These can potentially be detected by improved scan tests that better target the faults. They were reliably detected in the Intel data by Cell Aware Tests (CAT) under nominal test conditions. (2) Timing failures resulting from circuit timing marginalities caused by the combination of process variations and degraded power supply voltages. Recall that gate delay increases sharply as supply voltage is significantly lowered. The vast majority (146 out of 156) of the SLT failures studied in [1] passed the CAT delay tests at nominal voltages and only failed at low voltages quite close to Vmin. Thus, the significant contribution of timing errors, triggered in low voltage operation, to SLT failures appears validated. Such timing failures are the primary focus of this paper because of the difficult challenge of developing and applying high quality scan delay tests to screen them out. We assume that the much smaller number of hard defects contributing to SLT fails (< 7% in the Intel experiments [1]) can be further reduced through improved scan testing that exploits the ongoing development of advanced fault models such as CAT, gate exhaustive, and Timing Aware Cell Aware (TA-CAT).

B. Increasing timing failures in low voltage operation

Figure 2 reproduced from [1] shows minimum failure detection voltages Vmin at different operational frequencies for a population of *defective parts* in the Intel 14nm FinFET experiments. Failing Vmin values for individual parts for new timing aware (TA-CAT) scan tests are compared against those for (timing unaware) traditional TDF tests. Note that the tests were conducted at different test frequencies, with the lowest frequency F1 (plotted in red) at the left bottom of the plot, and the highest F5 at the right top. Observe that for most of the failing parts, the Vmin value is very close for the two tests, suggesting that the TA-CAT tests do not significantly enhance defect detection over TDF tests. However, for several defects

TA-CAT tests detect the failure at much higher supply voltages. These results indicate the activation of a timing fault along some long path sensitized by the timing aware tests that is not detected by the traditional TDF tests along a similar path. Observe in Figure 2 the much larger number of such outliers at lower operating frequencies. Whatever be the effectiveness (coverage) of the TA-CAT tests in detecting all timing failures in the tested parts (there is no way of estimating the TA-CAT test escapes from this data), it is clear from Figure 2 that many more timing errors, and of larger magnitude, are observed in low voltage operation, where processors are operated at reduced clock frequencies to save power. As the experimental SLT failure data shows, these are the major source of SLT fails.

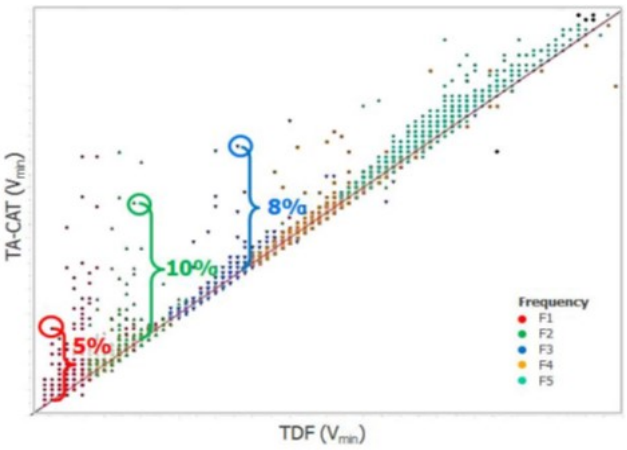

Figure 2: Vmin for TA-CAT versus TDF for timing fails from [1]

The results from [1] discussed above also explain the critical reliance of smartphone and notebook processor SOCs on SLTs [2]. Modern battery powered multicore SOCs deployed in stringent power constrained applications extensively employ aggressive dynamic voltage frequency scaling (DVFS) and adjust voltage and clock frequency settings to minimize power under low workload conditions. It is well known that energy use by circuits is minimized in near threshold operation [3]. Consequently, during low power operating modes, power supply voltages (VDD), already less than a volt for 10 and 7nm designs, are often lowered to within a couple of hundred millivolts of the threshold voltages to save power, with clock frequencies appropriately slowed. Additionally, power supply noise can add an additional 10% VDD degradation. Unfortunately, as we show through simulations presented later in the paper, it is not just nominal delays that are significantly increased at such low voltages. *Random gate delay variability due to normal manufacturing processes variations is even more greatly amplified.* This can make it prohibitive to provide large enough timing margins that can guarantee the accommodation worst case slow paths in every manufactured part. The practical timing margins chosen are forced to allow some possibility of parts from the tail of the variability driven delay distribution failing timing when worst case paths are activated. These "defective" parts must be screened out by manufacturing tests. Unfortunately, widely used Transition Delay Fault (TDF), and even Small Delay Defect (SDD) [10] tests, target lumped delays at circuit nodes and are unable to reliably detect such delay failures caused by the accumulation of random delays along circuit paths due to the distributed gate delay variability. Meanwhile, path delay scan tests [4,5] have also not proven practical. Consequently, at-speed functional system level tests (SLTs) appear to be the only effective option to detect and screen out such timing failures.

III. Related Research towards Eliminating System Level Tests

A. *Reducing test escapes from scan tests though improved fault models*

Much of research focused on eliminating system level tests is focused on improving the defect coverage of scan structural testing by generating improved test sets using advanced fault models. Traditional Stuck-At (SA) and Transition Delay Fault (TDF) models target faults at the nodes of the circuit. These nodes are the interconnections that connect the gates and flip flops (standard cells) and other electrical components in the design. However, modern designs incorporate a significant number of complex CMOS gates in the logic cells. It is increasingly observed that some faults within these complex cells remain undetected when only the input and output nodes of the cells are explicitly targeted for test generation by traditional SA and TDF tests. Recently, this has led to the development of the Cell Aware Tests (CAT) [6] that additionally also target all short and open defects within the standard cell layouts. The initial success of CAT in screening out several hundred additional DPPM in volume production [7], beyond those detected by traditional SA and TDF tests, raised the hope that cell aware tests may reduce defect levels enough to eliminate the need for SLTs. Indeed, CAT has been enthusiastically adopted by companies in the automotive sector with "zero defect" DPPM requirements. However, other applications appear to see less benefit from CAT and have been slower to adopt it. In particular, it was reported at an ITC 2017 panel discussion [9] that power constrained smartphone processor SOCs do not appear to experience a large reduction in SLT fallout when cell aware tests are introduced in the structural test flow.

B. *Many SLT failures manifest only in low voltage operation*

The key to this difference appears to lie in the differing operational power supply profiles of the two applications. Historically, automotive applications have not been power constrained. (This may change as Advanced Driver Assistance Systems (ADAS), requiring orders of magnitude greater computational throughput, become operational.) The focus in automotive systems so far has been on reliability and safety, with little motivation to save power through low voltage operation that can increase the risk of errors. Smartphone processors on the other hand, must employ aggressive dynamic voltage frequency scaling (DVFS), nearly always operating at minimum voltages, to maximize battery life. Such circuits can experience timing errors in low voltage operation from increased delays from process variability. As already discussed in the previous section, 146 out of the 156 SLT failures studied in the recent Intel experiment [1], passed all tests at nominal voltages and only failed at supply voltages close to Vmin. Such timing failures appear to dominate SLT fallout for smartphone

and notebook processors, but understandably do not significantly impact automotive applications. Consequently, automotive parts see a significant DPPM improvement from tests such as CAT that are primarily target against timing independent "hard" defects; smartphone applications dominated by low voltage timing failures appear to see less benefit.

C. *Improved scan tests for targeting low voltage timing failures*

In an attempt to also improve the effectiveness of scan tests in detecting timing failures, timing aware scan delay tests are also being investigated. The new TA-CAT tests [1] attempt to detect a target delay defect using the longest possible defect activation path (from an input to the defect site), and longest output propagation path (from the defect site to the output). Activating the longest path containing the defect ensures that the smallest possible delay defect capable of causing functional failure is detected. Note however, that the target defect itself is still assumed to be a single localized lumped delay at the target node, within or outside a standard cell. Timing aware scan tests of all flavors (TDF, CAT etc.), broadly referred to as small delay defect (SDD) tests [10], all assume a single localized delay. They often miss timing errors arising from the accumulation of multiple delays along circuit paths, as can occur in low voltage operation where individual gates display high gate delay variability. The latter require path delay tests [4], and, to be maximally effective, even more stringent robust path tests [5], for reliable detection. Unfortunately, path delay tests have so far not proven viable for various reasons that include the large number of near critical circuit paths, limitations on the delay patterns that can be applied through scan DFT, and the presence of a large number of timing hazards in CMOS.

D. *Test noise in scan timing tests for SDDs*

Additionally, because scan timing tests (including node oriented TDF and CAT delay tests) are not functional mode tests, their accuracy suffers from "test noise". Observe that these tests do not measure circuit delay in normal (continuous) operation, but attempt to mimic two cycles (a single transition) of functional operation in an attempt to capture circuit timing between launch and capture clock edges. Unfortunately, all the circuit electrical conditions in normal operation are generally not accurately replicated during the surrogate scan timing test. Factors that can cause the scan test timing to deviate from actual circuit timing in functional operation include transient noise in the power rail [11], die temperature variations [12], and "clock stretching"[13]. In the publications cited above, this scan "test noise" has been shown to introduce errors in the observed timing of up to 20%. Observe that in a design where the critical paths contain 30 gates, 20% of the path delay is the equivalent of 6 gate delays. Thus, the mismatch in timing between scan test results and actual functional operation can be of this magnitude. Timing margins in the design can absorb and hide an additional 10% (3 gate delays) increase in path delay without flagging an error. Consequently, even timing aware SDDs, such as TA-CATs, are only able to reliably detect localized lumped delay defects that increase the delay at the target gate output by more than 3-5X; in fact, they often fail to detect even larger delay defects. Meanwhile, due to the lack of practical path delay tests, the ability of scan tests to detect failures from the accumulation of distributed gate delays along circuit paths is even more significantly compromised.

Given these limitations of scan delay testing, it is not surprising that the only reliable solution that industry has found for screening out variability caused timing failures in low voltage operation are at-speed system level tests (SLTs) that extensively exercise the part in a true functional environment across various power modes and other operating conditions.

E. *Experimental results [1] for timing aware scan tests*

The effectiveness of Timing Aware Cell Aware Tests (TA-CAT) relative to TDF and CAT tests has also been investigated in the experiments reported by Intel [1] for a design in 14nm FinFET technology. TA-CAT tests were generated for approximately 25% of the CAT defects targeted with CAT-delay (two pattern) tests; these were defects where simulation of the cell library suggested possible small delay defect behavior (although no specific delay threshold for selecting these is reported). TA-CAT tests further significantly increase test pattern counts over CAT-delay tests. For the experiments reported, test pattern counts for TDF, CAT delay, and TA-CAT were approximately in the ratio X, 2X, and 2.5X, respectively. The test results in DPPM shown in Figure 2 are reproduced from [1]. Note that defects detected by all three test sets at intersection of the three circles in the Venn diagram are marked IP in Figure 2. This number is not disclosed since it could reveal proprietary yield information, but it can be realistically expected to be in the tens, or even hundreds of thousand DPPM. The results show that CAT delay patterns detected an additional 983 DPPM after TDF tests. (This number is broadly consistent with results published earlier for bulk CMOS processor technologies [7].) Adding the TA-CAT patterns resulted in the unique detection of a further 250 DPPM.

It is not obvious, however, how many of the 250 unique TA-CAT fails really are small delay defects requiring timing aware tests. To better understand this, observe from the Venn diagram in Figure 3 that TDF tests also detected 33 unique faults not detected by the CAT patterns. This occured even though the CAT delay patterns also target all TDFs in the design and have (at least) the same TDF coverage. This can happen because the test patterns covering the TDF faults in the CAT delay test set are not exactly the same patterns as those in the TDF test set itself. These different patterns can cover the same logical (TDF) faults but display differences in the actual physical defects detected in practice. Now if a large number of additional test patterns are applied, experience indicates that they always find some unique faulty parts in the production flow. The TA-CAT patterns are almost 2.5X larger in number than the TDF patterns. Assuming just random detection, they can be proportionally (2.5 x 33) expected to detect approximately 80 additional unique faults. This would be true even if these were additional (but different) TDF patterns, increasing say the N-detect capability of the TDF test set, and generated without any consideration of timing. Sensitizing long paths, as is done with the TA-CAT tests, typically leads to the detection of more TDFs per test, since more circuit nodes experience a logic transition. Thus, many (perhaps even a majority) of the 250 unique detects from TA-CAT tests in Figure 3 may have been detected fortuitously just because of

the large number of additional diverse two-pattern TDF type tests in the TA-CAT test set, and not necessarily because these tests were timing sensitive. While the authors in [1] say that the three test pattern sets were "executed multiple times using different VDD voltages and different test frequencies", they do not disclose if, or how many, of the unique TA-CAT detects were actually timing sensitive defects. Thus the reported data does not convincingly establish that TA-CAT tests are exceptionally effective in screening out timing sensitive failures, despite the large increase in test set size.

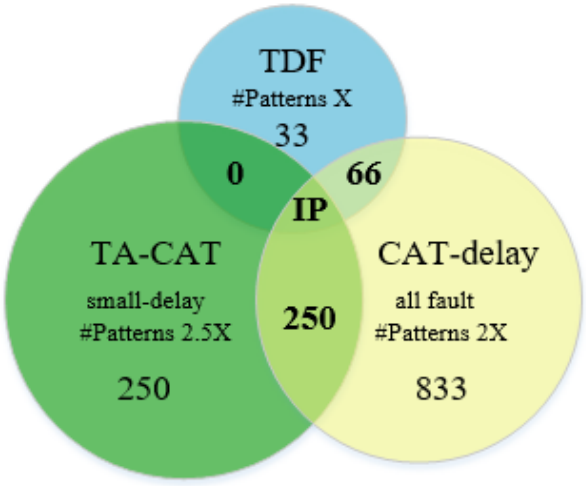

Figure 3: Results reproduced from [1] showing TDF, CAT-delay and TDF fault detections. The circles reflect the approximate size of the corresponding test set.

IV. The Motivation for an Alternate Approach for SLT Cost Savings

The 250 (approximately 25%) DPPM reduction by the TA-CAT over standard CAT delay tests can still be significant and meaningful in some applications, despite the high test cost from more than doubling of the already large CAT test set. However, it is unlikely to eliminate the need for system level tests, as many processor designs are reported to experience significantly greater fallout from SLT [2]. This is not unexpected. As discussed above, TA-CAT tests still target localized lumped delays, and not the accumulated delay faults along circuit paths due to process variations. Thus, they do not address the failures that contribute to the majority of SLT fallout in low voltage DVFS operation.

In the absence of accurate path delay tests, it currently does not appear practical to detect enough of the SLT fallout using scan tests alone to eliminate the need for SLTs. Instead, in this paper we focus on an adaptive approach for minimizing SLT costs by selectively avoiding SLT for a significant fraction of the manufactured parts. We achieve this by identifying those that are extremely unlikely to fail SLT based on test results for the part earlier in the test flow. To understand how this may be possible, we first study low voltage variability timing failures in some detail, with the help of SPICE circuit simulations.

A. *Impact of Reduced Voltage Operation and variability on Circuit Timing*

Recall that battery powered multicore processor SOCs, individual cores are operated over a wide range of voltages and power levels to always deliver the required performance at any given time with minimum power consumption. For error free operations, system clocks must be appropriately slowed down for any core operating at reduced power supply voltages. Up to 50% reduction in nominal VDD may be employed, since near-threshold operation is well known to minimize computational energy needs where performance is not an issue. However, operation at such low voltages significantly increases gate delays and has a major impact on device performance. Perhaps even more important, operation at reduced voltages, particularly near threshold operation, greatly increases path delay variability that can lead to timing errors. Unfortunately, because path delays cannot be reliably tested by scan tests, and adding sufficient timing margins to eliminate all possibility of timing errors can result in prohibitive performance degradation, screening out such failures requires expensive SLTs.

The challenge of setting appropriate clock rates for error free operation under aggressive dynamic voltage frequency scaling in the face of random process variations is illustrated with the help of a simple example. Recall that for scaled devices under the carrier velocity saturation, the gate delay can be roughly approximated by:

$$\text{Gate Delay} = K / (VDD - V_{th}). \quad (1)$$

The term ($VDD-V_{th}$) is informally called the gate overdrive, the amount by which the gate voltage exceeds the minimum voltage (V_{th}) voltage required to turn ON the transistor. It is obvious from equation (1) above that for low voltage, near threshold operation, when the gate overdrive is small, gate delays increase nonlinearly as VDD is reduced. Also, the gate delay is more significantly impacted by small random manufacturing variations in V_{th} at low VDD. For example, for V_{th} (nominal) = 0.4 V, dropping VDD from 1.2 to 0.6 volts increases the nominal gate delay by a 4X factor. This requires the clock to be slowed down by a corresponding amount, greatly affecting throughout.

In practice, however, the clock must be further slowed down by a significant additional amount to allow for disproportionally larger timing margins required at the lower voltages. This to provide sufficient slack for added circuit path delays that can be caused by random manufacturing variations. These variations are observed not only in the transistor threshold voltages V_{th}, but also in other key device parameters associated with short channel effects (SCEs). Critically, many of these SCEs, e.g. DIBL (drain induced barrier lowering), are greatly influenced by the doping profiles at the drain, source and channel of individual transistors. Given the multiple doping masks used to engineer highly scaled transistors, and the inevitable random variations in delivering small numbers of dopant atoms through ion implantation, statistical variations in several other key device parameters are inevitable. For example, DIBL is known to display a log normal distribution [14], which has a long tail, with a much higher likelihood of large values than a normal

(Gaussian) distribution. Unfortunately, statistical modeling of circuit timing to account for these variations at low operating voltages, where they have the greatest impact, is not well developed because transistors are normally not expected to operate in conditions of weak inversion. In the rest of this section, we only consider Vth variations in evaluating timing, which for simplicity, are considered independent of other device variability factors (as is common practice). We recognize that our analysis may be optimistic –the actual worst case circuit delays from variability may be larger.

V_{th} random variability is generally observed to be Gaussian, with a typical standard deviation (sigma) in 15-40 mV range. For our example, based on the closed form approximation of equation (1) above, the gate delay variation for a 1-sigma (+30mV) change in V_{th} ranges from -3.6% to +3.9% at VDD = 1.2V. It has a much wider range from -13% to +17% at VDD=0.6V. Observe that statistically one in six transistors in the circuit will have V_{th} greater than 1-sigma and will therefore display a delay of 17% or greater beyond the nominal delay at the lower voltage. With millions of circuit paths, the probability that some long path has mostly slow transistors is not insignificant. Similarly, for a 2-sigma (+60mV) V_{th} change, the delay range is -7% to +8% of the nominal at VDD=1.2V, but as much as -23% to +42% at VDD=0.6V. Thus in practice, low voltage operation requires much larger timing margins as a percentage of the clock period to accommodate statistical worst-case slow paths. The typical 10-20% timing margins used at nominal voltages are not sufficient.

Observe from the above example calculations that the gate delay distribution corresponding to a normal random V_{th} distribution is not symmetric, but is non-Gaussian. For the same magnitude V_{th} variation, the increase in gate delay is greater than the speed-up. This asymmetry is greatly exaggerated at low VDD. For example, for a +60mV threshold voltage change in both directions, at VDD=0.6 volts, the gate delay increases by 42%, but the speed-up is only 23%. Conventional wisdom holds that delay variability generally averages out for long paths, and so delay variability is not a major problem. While this is largely true at nominal VDD (large gate overdrive) where the gate delay distributions are more symmetrical, it does not hold for low voltage operation.

To study this issue in greater depth we present Monte Carlo HSPICE timing simulations for inverter chains implemented in 32nm technology. The nominal V_{th} for the PMOS and NMOS transistors are -0.416V and +0.401V respectively from the technology files. We have modified these nominal values with a Gaussian distribution to account for random variability that has a standard deviation of 30mV. Figure 4(a) shows the delay distribution for a single inverter in the middle of the chain for VDD=1.2V and 0.5V to illustrate extreme low voltage operation. The delays are in picoseconds. Observe that the inverter operating at the low voltage of 0.5V has a dramatically wider and more skewed delay distribution, with a long tail compared to the very narrow VDD=1.2V distribution to the left. We have used an aggressively scaled voltage in these simulations to highlight the variability in the plots. However, the trends are similar for somewhat less scaled voltages, as can be seen for VDD=0.6V in Figure 4(b).

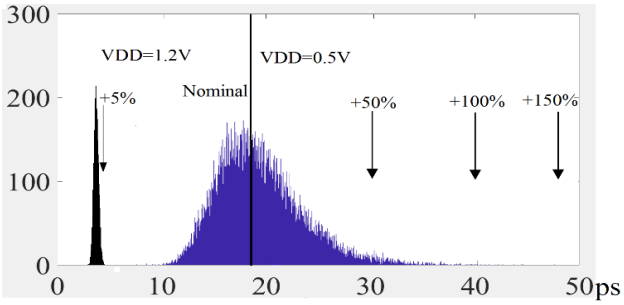

Figure 4(a): Delay distribution for an inverter for VDD = 1.2 V and 0.5V

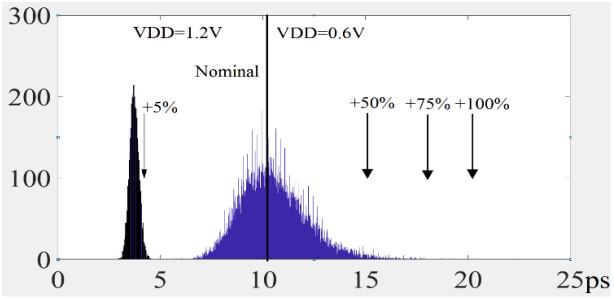

Figure 4(b): Delay distribution for an inverter for VDD = 1.2V and 0.6V.

To study a simplified a circuit path comprising multiple gates, Figure 5 shows the delay distribution for a 20 inverter chain for the same two voltages. The vertical markers shown indicate the nominal delay (i.e. all transistors at their nominal values, with no variations), and added timing margins of 5, 20 and 30% of the nominal delay, for each voltage. It is clear that low voltage operation requires much larger timing margins than nominal 1.2V operation for sufficient to ensure no timing failure for any path.

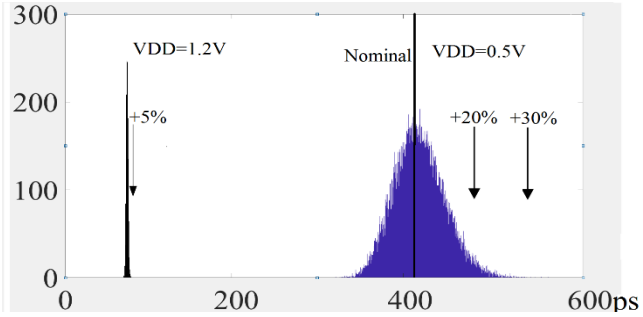

Figure 5(a): Delay distribution for 20 inverter chain for VDD= 0.5V.

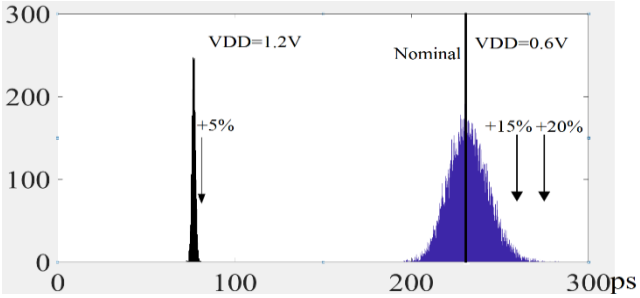

Figure 5(b). Delay distribution for 20 inverter chain for VDD= 0.6V

It is instructive to note that only random V_{th} variations are considered in the above simulation and discussion. There are also several other sources of variability and noise that can add additional delays in the circuit paths. Systematic V_{th} variations across the wafer and wafer lots also play a very significant role in timing failure. These are discussed in more detail in the next section. Additionally, circuit noise, e.g. power supply droop [11,13], can further reduce power supply voltages and introduce more delay variability and therefore require additional noise margins. This is the reason that, while the simulations in Figure 5 suggest that, based only on random process variations, the inverter chain can operate reliably with about a 5% timing margin, in practice timing margins at nominal voltage are set closer to 15% to allow for all these other sources of noise and variability. Applying a similar 3X factor to the 30% variability margin needed at 0.5V operation in Figure 5 would suggest nearly a 100% timing margin. Anything even close to this number would clearly be prohibitive in terms of performance loss, which, as can be seen in the simulations, is already very significant in low voltage operation for nominal V_{th} without any variations. (Consistent with the simplified calculation in the introductory example, nominal path delay in the simulations slows down by more than a 3X factor.)

V. The Proposed Adaptive System Level Test Strategy

The simulations in the previous section show that to accommodate all the statistical outlier paths that display very large delays in low voltage operation due to manufacturing device variability requires the design to incorporate correspondingly large timing margins. This can cause unacceptable performance loss in all manufactured ICs. The alternative that appears to define actual practice, even if it not explicitly stated, is to use a more aggressive clock to avoid this extreme performance loss, and detect the few timing fails from the tail of the delay distribution during manufacturing test. Unfortunately, these timing failures are not reliably detectable by scan timing tests and necessitate the use of SLTs.

Our proposed adaptive test flow only applies SLTs to a subset of manufactured parts, as shown in Figure 6. This requires classification of manufactured parts between those with a significant likelihood of failing SLTs, and those extremely unlikely to fail SLTs. The latter can skip SLTs without significantly impacting DPPM in the shipped product. We assume that all parts arriving at the classification stage in Figure 6 have been extensively tested using scan structural tests at both wafer probe and during post packaging tests, including using tests generated by advanced fault models such as Cell Aware. This ensures that virtually all parts with hard short and open defects, as well as gross delay defects have been screened out. As discussed in the previous sections, the source of most of the remaining failures are timing errors caused by accumulated delays along paths in low voltage operation from threshold voltage (Vth) variations and other device variability factors related to short channels effects. In the following discussion we again primarily focus on Vth variability, although the impact of possible non-Gaussian variability such as DIBL is also briefly discussed later in the paper.

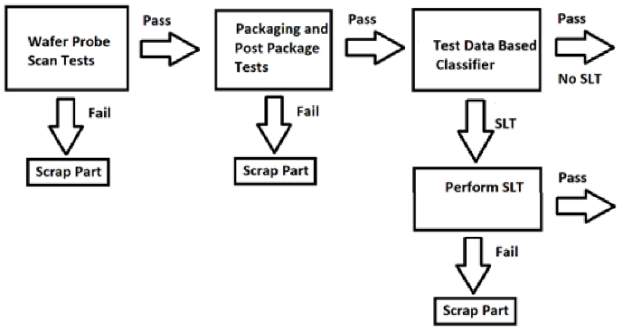

Figure 6: The adaptive test flow to minimize SLTs

Recall that the variability in transistor threshold voltages, defined with respect to the nominal for the technology, has two components: a systematic variability component, plus a random variability component. Systematic variability affects all devices on a die in the same way, and reflects the die-to-die (inter-die) variation in Vth (and other device parameters), whereas random variability captures the intra-die variation within individual dies. Historically, significant systematic variations were mostly observed between die from different fabrication lots because of subtle differences during the manufacturing steps from the use of different equipment, different batches of raw materials and chemicals, differences in operator handling, etc. However, in highly scaled technologies, die on the same wafer also display significant inter-die variation. We propose to exploit this systematic variation in Vth between manufactured die to classify and bin die (package parts by this stage in the manufacturing flow) for SLT.

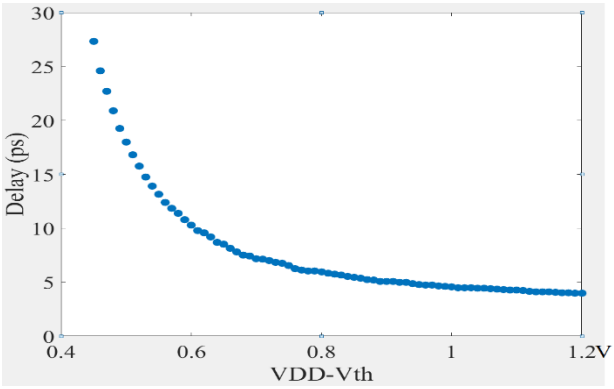

Figure 7: Inverter delay versus gate overdrive VDD-Vth

Figure 7 shows simulation results for inverter delay versus the overdrive voltage VDD-Vth for the same 32nm technology discussed earlier. Observe the rapid increase in delay as VDD approaches Vth, i.e. the overdrive approaches to zero. When VDD-Vth is down to a couple of hundred millivolts, as can happen through a combination of power rail droop and DVFS in extreme power saving modes, small changes in Vth can result in large delay variations. It can be observed from Figure 7 that when systematic variability lowers the average threshold voltage in a die, increasing the overdrive VDD-Vth, the transistors will on average operate much faster than nominal devices in low voltage operating modes. Even in the presence of additional random Vth variations in both directions, such die are less likely to fail timing than die where the systematic variation raises the

mean threshold voltage above the nominal. Thus, the mean threshold voltage of a die is clearly one parameter that can be used to classify die in terms of their likelihood of failing SLTs.

A. *Simulation results for Vth based classfication*

We next present some simulations to illustrate Vth based binning. Let us assume that the systematic component results in a constant offset ΔVth*sys* from the nominal threshold voltage for the process, Vth*nom*, for the entire die. The random variability Vth*ran*, typically displays Gaussian statistics with standard deviation σ = in the 15 to 40 mV range. Thus, for any transistor Vth = Vth*nom* + ΔVth*sys* + ΔVth*ran*. Noting that gate delay is approximately proportional to 1 / (VDD –Vth), for the same random variability one would statistically expect the largest gate delays, relative to the nominal, in circuits that have the largest ΔVth*sys*. This is seen in Figure 8, which shows simulated path delay distributions for the same 20 inverter chain discussed earlier for VDD = 0.5V and ΔV*sys* = - 20 mV, 0 mV and +20 mV. Systematic Vth variations over multiple production lots and months of manufacturing can range up to 10% of the nominal Vth, (± 40 mV for the simulated) circuits, or even more.

Observe in Figure 8 how the path delay distribution for a systematic positive shift in Vth moves significantly to the right (larger delays) and also spreads out with a longer slow tail. The functional clock rate for low voltage operation (in the most aggressive power saving mode) is typically chosen based on nominal circuit speed plus some timing margin. Consequently, for the same distribution of random process variability, parts

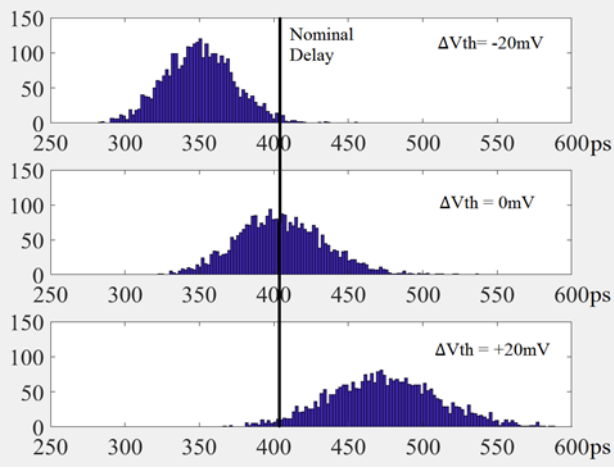

Figure 8: Impact of systematic Vth shift on path delay at VDD = 0.5V

with the largest positive systematic variation in Vth are much more likely to fail timing. The large majority of SLT timing failures can therefore be expected from parts with a significant positive shift in Vth*sys*. This suggests that if manufactured parts are classified based on increasing ΔVth*sys*, SOCs which high positive ΔVth*sys* will experience significantly greater SLT fall out. Those with negative ΔVth*sys* will display much lower SLT failure rates and can even skip SLT depending on the target DPPM. This trend is greatly accentuated in low voltage operation because of the high sensitivity of delays to Vth variations. Note that if the Vth*sys* distribution across dies is symmetric, well less than half the incoming parts will have a large positive Vth*sys* shift.

B. *Estimating timing error probability for critical paths*

The probability of a critical path failing to switch within a clock period, and thereby causing a timing error at VDD=0.5V, can be estimated for the three Vth*sys* distributions in Figure 8. This probability clearly depends on the timing margin added to the clock period, beyond the 405.5 picosecond nominal critical path delay observed with nominal device parameters, i.e. when ΔVth*sys* = 0. The larger the timing margin, the smaller will be the tail of the distribution to the right of the timing limit, with a corresponding reduction in the chance of an error. From the simulation data, we can readily obtain the mean and standard deviation for each of the distributions. Once the timing slack between the nominal path delay for each distribution and the clock period is expressed in terms of its standard distributions, the error probability can be read off published charts of tail probabilities for Gaussian distributions. The results are shown in Table 1.

TABLE I. PROBABILITY OF TIMING ERROR FOR VARYING MARGINS

ΔVth*sys*	Mean Delay	SD (σ)	Timing Margin			
			20%	30%	40%	50%
-20 mV	350.16	23.83	8.7 E-9	2.6 E -14	≈0	≈0
0 mV	405.50	31.93	5.1 E-3	6.9 E-5	2.8 E-7	4.1 E-10
+20 mV	474.43	38.75	3.8 E-1	8.8 E-2	7.7 E-3	2.8 E-4

Observe the huge influence of ΔVth*sys* in the critical path timing error probabilities. For a 30% timing margin, the error probability for a part with ΔVth*sys* = +20mV is a thousand times larger than that for ΔVth*sys* = 0, and almost a trillion times larger than that for ΔVth*sys* = -20 mV. For a (more realistic) 50% timing margin, the timing error probability for ΔVth*sys* = +20 mV is a million times higher than for ΔVth*sys* = 0. *These numbers suggest that for practical timing margins, virtually all of the timing failures occur in parts with elevated ΔVthsys*. This is also apparent from Figure 8.

Large high performance SOCs can contain thousands, even hundreds of thousands, of near critical paths, all with similar delays due to timing closure during design to achieve aggressive clock rates. Unless the probability of timing error for any individual path under nominal conditions is well below one in a million (1 E -9), the probability of some path failing timing becomes quite significant, and leads to unacceptable manufacturing yield loss. Hence the need for large timing margins. Note however that the near critical paths in such circuits may not all be independent. Therefore, the timing error probabilities for a single critical path, such as shown in Table 1, cannot be easily used to directly estimate the timing error failure probability for the complete circuit.

C. *Estimating systematic Vth shifts*

Based on the above discussion, our methodology for classifying manufactured SOCs into the two categories, those that require SLT and those that do not, as shown in Figure 6, can be built upon estimating ΔVth*sys*, the systematic threshold

voltage variation relative to the nominal, for each part. While this parameters is not commonly directly measured for each die during production testing, a number of surrogate measures can be employed to estimate it. For example, ring oscillators are often used to measure the "speed" of the logic for an individual instance of a manufactured part. Since the delays in the many individual inverters comprising the ring oscillator are strongly influenced by their transistor threshold voltages, the ring oscillator frequency reflects the average (NMOS and PMOS) Vth, and correlates well with Vthsys. Additionally, measurements from several other test structures are commonly used in practice to characterize and classify individual SOCs in terms of process corners (e.g. T: TYPICAL, S: SLOW, F: FAST). These can be used as additional data inputs towards classifying parts for the adaptive optimization of SLTs.

For large die, systematic variations within different regions of the same die can be a concern. This can be managed by obtaining performance measures (e.g. using ring oscillators) from multiple areas of the die, and conservatively making the classification decision based on the worst-case measures.

D. Real-time machine learning based adaptive test flow

Our adaptive approach for optimizing SLTs as proposed above requires some mechanism for aggregating the surrogate measures of ΔVth*sys* in the test data obtained earlier in the test flow, and setting thresholds to select parts that can avoid SLT. Explicit algorithms to perform this computation can be complex, and may have to be adjusted over time to react to process changes, product mix, etc. However, once it is experimentally validated that there is useful information in the parametric and scan test data that can be used to reliably classify parts for effective adaptive application of the SLT tests, explicit classification rules are not necessary. Deep learning techniques, working on the parametric, scan and SLT test data (from parts that do undergo SLT), gathered at all stages in the test flow, can use test results for an individual part from earlier in the test flow to continuously optimize the selective application of SLT tests in real time. Such an approach can be designed to adaptively achieve desired DPPM levels at minimum SLT test cost. Product volumes for commodity SOC parts, such as smartphone processors, run into the millions. These can provide ample test data for accurate classification based on machine learning techniques.

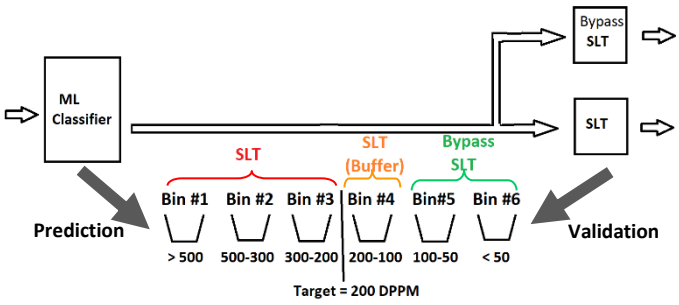

Figure 9: Predicted DPPM based binning for selecting parts to bypass SLT

A possible classification strategy employs real time supervised learning to predict the SLT failure rates (in DPPM) based on the parametric and scan test data measured from test structures, along with the pre and post packaging tests for each die. All incoming parts are classified and binned before the SLT tests into multiple (5-10) contiguous bins that span DPPM rates around the target defect level. An example six-bin arrangement for a target defect level of 200 DPPM, is shown in Figure 9, along with the predicted DPPM range for each bin. After SLT is performed, the number of actual SLT failures for the dies classified in each bin can be compared to the predicted DPPM to aid learning and improve the classification.

Initially all incoming parts are subject to SLTs until the supervised learning based prediction matches the observed DPPM levels in each bin. Observe that a well designed learning algorithm should always converge to make the best possible predictions based on the available data. Let us suppose that the residual DPPM after scan testing in 600 DPPM. Consider two extreme hypothetical scenarios. In the first we assume that there is nothing in the available scan test data that can actually predict SLT failures, i.e. SLT fails are completely independent of the earlier test results. In this case, a correctly working learning algorithm will converge to predict equal SLT failure probability for all incoming parts. All parts will be placed in the same bin by the classifier. Since the incoming DPPM is 600, based on the results of the SLT tests in the supervised learning mode, the classifier will place all the parts in Bin #1 (>500DPPM). Clearly this bin will require SLT (of all parts) to achieve the target 200 DPPM. In a second scenario, assume that the scan tests results contain all the information to precisely predict the SLT fails (although how exactly to use this information to make the right prediction in an explicit rule based manner may be unknown). In this case, after complete learning, the classifier will take advantage of all available information to correctly predict the SLT fails and place all the faulty parts in Bin #1, and all good parts in Bin #6. Because only 0.06% of the parts are faulty (corresponding to 600 DPPM), virtually all the parts will be placed in Bin #6. This would almost completely eliminate the need for SLTs.

Any practical scenario will clearly reside in between the above two extremes. As discussed earlier in this paper, there is information in the tests performed prior to SLTs to make some probabilistic, although not perfect, prediction of timing failures detected by SLT. Therefore, in practice, all the bins in Figure 7 will be assigned some parts based on the DPPM prediction by the classifier. Further, once the real time supervised learning has stabilized, these predictions will match the actual DPPMs in each bin following SLT. If, as assumed, the target DPPM is 200, once learning has stabilized, parts classified in Bin #5 and Bin #6 can clearly be allowed to skip SLT without any risk to defect levels. While in principle, Bin#4 can also skip SLT, it may be safe to continue to test and monitor it as a buffer to gather additional SLT data and to ensure that the parts that avoid SLT are indeed within the DPPM targets. In practice, the number of bins, their DPPM ranges, the selection of this safety buffer, etc. will all depend on depend on the specifications and failure statistics of the target design. Also, the bin selection for SLTs shown in Figure 7 is conservative. Even if Bin #4 was not selected for SLT, the average DPPM is the Bypass SLT parts would be the average of Bins #4, #5 and #6. Depending on the number of parts in each of these bins, this could be well under 200. Thus, once stable statistics for each bin become available,

the dividing threshold between SLT and Bypass SLT parts can be statistically set higher than the target 200 DPPM. The setting of this threshold to achieve a desired defect level can also be handled though machine learning.

E. Discussion

How much cost savings can be achieved by such an adaptive SLT strategy is clearly limited by how much information actually exists in the test data available prior to SLT to meaningfully predict the SLT failures. The target here are process variability driven timing failures; we assume based on recent experience with CAT, that scan test escapes of hard defects will continue to be minimized by improved fault models. Most process variability (not just Vth) has both a random and a systematic component. The proposed machine learning approach exploits any and all measures of systematic variability. This can be sensed at wafer probe, for example, by ring oscillators operated at (perhaps multiple) low voltages. Ring oscillators naturally average out device parameters over a large number of inverters. Given the advanced state of deep learning techniques, and the extent of test data available from high volume parts, it is not unreasonable to assume that virtually all the available information can in fact be exploited by a machine learning based classifier.

Finally, transistor threshold voltage is well known to have a Gaussian distribution. A concern, unaddressed so far in the paper, is the possibility of other variability parameters that affect circuit timing having a more skewed distribution. For example the transistor parameter DIBL is known to have a log normal distribution with a long tail [14]; even a die with near nominal average DIBL can have a significant probability of some transistor having an large outlier value. Thus, a part binned "Bypass SLT" based on some measure of the average DIBL, can contain an extreme outlier transistor, which might alone cause a timing error. Note however, that such a timing fault would mostly be concentrated at the output of the single gate containing the weak transistor, and will correspond to a lumped delay fault. This type of timing faults is already detected by node oriented scan timing tests such as TDF, and especially by timing aware tests such as TA-CAT. The much more difficult challenge faced by scan timing tests is the accumulation of delays contributed by many individual gates along a path. These occur in sufficient numbers only when the variability distribution is close to a normal distribution such that there is a likelihood of elevated delays in a significant number of gates. Our proposed adaptive methodology effectively addresses these timing failures.

VI. CONCLUSION

Efforts to eliminate or minimize the use of SLTs in industry have primarily focused on new fault models and improved test generation methods to improve the effectiveness of scan tests. In this paper we have argued that given the limitations of scan timing tests, such an approach may not be sufficient to detect all the low voltage failures caused by circuit timing variability that appear to dominate SLT fallout. Instead, we have proposed an alternate adaptive approach for meaningful cost savings that adaptively avoids SLT tests for a subset of the manufactured parts. This is achieved by using parametric and scan tests results from earlier in the test flow to identify low delay variability parts that can avoid SLT with minimal impact on DPPM. We also shown that such an adaptive test flow is also very well suited to real time optimization during the using machine-learning techniques. Future research efforts will be focused on collaborating with industry to validate the proposed adaptive test methodology in volume production.

REFERENCES

[1] W. Howell et *al*., "DPPM Reduction Methods and New Defect Oriented Test Method Applied to Advanced FinFET Technologies," *2018 IEEE International Test Conference (ITC)*, Phoenix, AZ, 2018.

[2] Sajjad Pagarkar, "Component SLT" Industry Test Challenges Meeting (coordinated by Phil Nigh) at the 2017 International Test Conference. Unpublished.

[3] Dejan Markovic, Cheng C. Wang, Louis P. Alarcon, Tsung-Te Liu, Jan M. Rabaey, "Ultralow-Power Design in Near-Threshold Region" Proceedings of the IEEE, Vol. 98 , No. 2 , Feb. 2010.

[4] G. L. Smith, "Model for Delay Faults Based Upon Paths", in Proc. International Test Conference, 1985, pp. 342-349.

[5] Lin, C.J. and Reddy, S.M., 1987. On delay fault testing in logic circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 6(5), pp.694-703.

[6] F. Hapke et al., "Cell-Aware Test," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 33, no. 9, pp. 1396-1409, Sept. 2014.

[7] F. Hapke et al., "Cell-aware Production test results from a 32-nm notebook processor," 2012 IEEE International Test Conference, Anaheim, CA, 2012, pp. 1-9.

[8] E. J. Marinissen et al., "Adapting to adaptive testing," 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010), Dresden, 2010, pp. 556-561.

[9] Panel 2: "Cell Aware ATPG: Beyond the Hype", 2015 International Test Conference. Unpublished.

[10] X. Lin et al., "Timing-Aware ATPG for High Quality At-speed Testing of Small Delay Defects," 2006 15th Asian Test Symposium, Fukuoka, 2006, pp. 139-146.

[11] J. Saxena, K. M. Butler, V. B. Jayaram, S. Kundu, N. V. Arvind, P. Sreeprakash and M. Hachinger, "A case study of ir-drop in structured at-speed testing", in Proc. International Test Conference, 2003, pp. 1098-1104.

[12] T. M. Mak, A. Krstic, K. T. Cheng and L. C. Wang, "New challenges in delay testing of nanometer, multigigahertz designs", IEEE Design & Test of Computers, vol. 21, 2004, pp. 241-248.

[13] Jeff Rearick, Richard Rodgers, "Calibrating clock stretch during AC scan testing". Proceedings 2005 International Test Conference.

[14] N. Damrongplasit, L. Zamudio, T. K. Liu and S. Balasubramanian, "Threshold Voltage and DIBL Variability Modeling Based on Forward and Reverse Measurements for SRAM and Analog MOSFETs," in IEEE Transactions on Electron Devices, vol. 62, no. 4, pp. 1119-1126, April 2015.

[15] J. Wang, D. M. H. Walker, A. Majhi, B. Kruseman, G. Gronthoud, L. Elvira Villagra, P. van de Wiel and S. Eichenberger, "Power Supply Noise in Delay Testing", in Proc. International Test Conference, 2006, pp. 1-10.

[16] K. Takeuchi et al., "Understanding Random Threshold Voltage Fluctuation by Comparing Multiple Fabs and Technologies," 2007 IEEE International Electron Devices Meeting, Washington, DC, 2007, pp. 467-470.

[17] A. Asenov, "Simulation of Statistical Variability in Nano MOSFETs," 2007 IEEE Symposium on VLSI Technology, Kyoto, 2007, pp. 86-87.

[18] M. L. Bushnell and V. D. Agrawal, "Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits", Springer, 2000, pp. 39-40.