Silent Error Corruption: The New Reliability and Test Challenge

Adit D. Singh
Department of Electrical and Computer Engineering
Auburn University
Auburn, AL 36849 USA
adsingh@auburn.edu

***Abstract*— Large commercial datacenters have recently highlighted a new and significant test and reliability challenge: manufacturing test escapes that cause silent data errors during operation. While there are many potential sources that can cause these failures, research is pointing to timing errors from random process variations, accentuated in power saving low voltage operation, being a significant contributor. Screening out such failures will require new and better timing tests that can reliably detect outlier circuits with path delay failures.**

Keywords—Silent errors, timing failures, process variations

I. Introduction

Traditional scan structural tests frequently fail to achieve DPPM (defective parts per million) targets for SOCs fabricated in state-of-the art technology nodes despite the introduction of advanced new fault models for test generation in recent years. Most notably, these new test generation methods include defect-oriented Cell Aware Tests (CAT) that target shorts and opens inside standard cells in the design, and timing aware tests that reduce timing slack for better testing of delay defects. Consequently, functional system level tests (SLTs) are increasingly needed as a final test screen to eliminate scan test escapes. However, there is growing evidence of a significant number of malfunctioning ICs are evading even these SLT screens and finding their way into parts deployed in operation. Particularly intriguing are a new class of test escapes that cause silent data errors (SDEs). Google recently reported [1] that some processors in their large date centers do not always perform as expected. "These machines were credibly accused of corrupting multiple different stable well-debugged large-scale applications. Each machine was accused repeatedly by independent teams, but conventional diagnostics found nothing wrong with them." Such "mercurial cores" were found to miscompute occasionally, under varying operating conditions, suggesting defect levels of such faulty or unstable hardware as high as 1000 DPPM. Other companies, including Facebook (Meta), operating large server farms, have also reported similar failures.

II. Testing Timing and SDE Failures

Any manufacturing fault in a circuit that is very rarely activated in operation, and perhaps additionally, also has a limited impact on the functional logic in terms of the number and importance of the flip-flops it corrupts, can potentially cause SDEs. Possible sources of such failures include permanent logic faults caused by open and short defects, and delay faults that impact circuit timing. In advanced technologies, the sources of timing errors include not only manufacturing defects such as shorts, but also systematic and random changes in circuit parameters caused by process variations.

As new improved fault models such as CAT, supported by better test generation and design-for-test (DFT) methodologies, have significantly increased coverage of the permanent faults targeted by scan tests, a growing share of test escapes from post manufacturing testing are being observed to be timing failures [2]. This is likely because two-cycle transition delay fault (TDF) scan delay tests target lumped delays at the circuits nodes caused by defects, and not the distributed delays that can accumulate over multiple gates along circuit paths from process variations. Modern processors have billions of circuit paths, many of which have very similar delays because of careful timing closure during logic synthesis and layout to meet aggressive clock rates. When subject to random processes variations, any one of these long paths can potentially accumulate sufficient extra delay from multiple random slow gates to become the slowest path in a specific instance of the manufactured IC and cause timing failure. Thus all, or at least a large fraction, of the paths in each manufactured part must be tested to ensure that passing ICs meet specified timing requirements that guarantee failure free operation in the field. However, testing such a large number of circuit paths would require an unacceptable increase in the number of tests required, because of prohibitive test application time and resulting test cost. Because of this, and a number of other technical limitations, scan-based path delay fault (PDF) tests are impractical. Targeted high coverage scan tests to screen out slow paths resulting from random process variations remains elusive.

Observe that at most only about a million two-cycle timing tests are currently applied when scan testing even the largest ICs and SOCs that typically contain many billion circuits paths. Given this relatively small number of test patterns compared to the number of paths to be tested, the likelihood that an isolated slow path with the potential to cause timing failure will be activated and detected by random chance is extremely small. Consequently, alternative timing tests to screen out timing failures due to process variations need to be developed and deployed.

979-8-3503-2597-3/23/$31.00 ©2023 IEEE

III. Functional System Level Tests

Over the past few years, industry has responded to this challenge with a brute force solution by introducing functional system level tests (SLT) as a final test screen. During SLT, the packaged device is temporarily mounted on a test board that closely replicates the intended end application hardware. This complete system assembly is then extensively tested anywhere from 15 minutes to as long as 2 hours at the rated frequency, over a range of user applications and operating conditions. A processor operating at GHz frequencies goes through more than a trillion clock cycles in 15 minutes, each applying different inputs to the internal logic. This greatly increases the likelihood that many, even most, of the circuit paths are randomly exercised and thereby tested for timing failures. However, SLT requires entirely new test systems and infrastructure as compared to traditional ATE. This is because SLT testers must be capable of testing hundreds of test boards in parallel to ensure sufficient throughput of tested parts in the face of the very long functional test times that are typically three orders of magnitude longer than scan tests. Additionally, the temperature of circuit under test can increase unacceptably when operated at-speed for long durations confined in a temporary test socket. This requires an extensive cooling capability associated with each test board and socket, and active thermal management that can accurately control the temperature of the circuit being tested to mimic realistic worst case operating conditions in functional deployment. As a result, SLT adds significantly to test costs. Consequently, new low-cost scan tests that can reliably detect timing failures caused by random process variations remain of great interest to industry.

Unfortunately, even functional SLTs are unable to detect all failures. There is no known systematic approach available for developing high coverage functional tests with respect to any fault model; indeed, the fault coverage metric is rarely used with respect to functional tests. Thus, the path delay fault coverage of the tests applied during SLT is unknown. Some faulty devices inevitably escape, contributing to the DPPM counts for the part.

IV. Low Voltage Operation

Developing better tests for the timing failures from process variations requires understanding of the circuit conditions under which delays are accentuated and therefore more likely to cause failure [3]. As discussed earlier, random parameter variations can cause each circuit path, and therefore every input transition, to display a different and unique delay for different instances of the same manufactured design. This makes testing for the worst-case timing delay extremely challenging.

Much of the impact of process variations on timing can be modeled through changes in transistor threshold voltages (V_{th}) [3]. V_{th} values for the billions of transistors in a chip are found to display a normal distribution around an average (nominal) value. Slower transistors, with weaker current drives, are those with threshold voltages higher (in magnitude) than the nominal. Lower V_{th} values speed up the switching delay for the logic gate driven by the transistor. This delay depends on the "overdrive" voltage, $V_{OL} = V_{DD} - V_{th}$, indicating how strongly the transistor is turned on. The gate delay is approximately proportional to $1/V_{OL}$. Note from the above that as the supply voltage V_{DD} is decreased, variations V_{th} will have a bigger impact on gate delay, accentuating the influence of process variations on path delays. Today, on-chip thermal management systems in processors employ dynamic voltage frequency scaling, at times operating only ~150mV above nominal V_{th} to minimize power dissipation under high workloads. A statistically slow transistor due to random process variations with a somewhat extreme 100mV elevation in V_{th} above the nominal can see a 3X increase in gate delay. Given the billions of transistors in modern processors, such transistors are not so rare. Note that the standard deviation (σ) of the V_{th} process variations is usually around 20mV, making the probability of encountering a 5σ transistor is approximately 300 in billion. Thousands of very slow transistors can therefore be expected in every processor. Furthermore, with hundreds of billions of paths in the circuit, a few paths can also be expected to have mostly slow transistors and much longer delays. A single statistical outlier slow path, with a total delay that exceeds the clock period, results in a failing chip that exhibits timing errors.

V. The Risk from Timing Marginalities

In addition to the parts that fail during post-production timing tests, either scan or functional SLT, there are potentially many others that contain one or more statistical outlier slow paths that narrowly pass timing under the test conditions. These marginal parts may occasionally fail and cause SDEs in deployment under worse case operating conditions caused by power supply droop and/or die temperature changes, or increases in V_{th} from transistor aging. SDE failure rates are reported to increase as systems age, which is consistent with the perception that many of these are timing failures caused by slow transistors.

VI. Conclusions

Large commercial datacenters have recently highlighted a new and significant test and reliability challenge: manufacturing test escapes that cause silent data errors during operation. While there are many potential sources that can cause these failures, research is pointing to timing errors from random process variations, accentuated in power saving low voltage operation, being a significant contributor. Screening out such failures will require new and better timing tests that can reliably detect outlier circuits with path delay failures.

Acknowledgement

This manuscript is based upon work supported by the National Science Foundation under Grant CCF-1910964. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

[1] Hochschild, Peter H., et al. "Cores that don't count." Proceedings of the Workshop on Hot Topics in Operating Systems. 2021 https://dl.acm.org/doi/10.1145/3458336.3465297

[2] W. Howell et al., "DPPM Reduction Methods and New Defect Oriented Test Method Applied to Advanced FinFET Technologies," Proceedings of the IEEE International Test Conference 2018, Phoenix, AZ.

[3] Adit D. Singh, "Understanding VDDmin Failures for Improved Testing of Timing Marginalities", Proceedings of the International Test Conference, 2022, 1-10, Anahiem CA.