Testing in the Year 2020

Rajesh Galivanche (1), Rohit Kapur (2), Antonio Rubio (3),
(1) Intel Corp., Santa Clara, USA, (2) Synopsys, Mountain View, USA, (3) Technical University of Catalunya, Barcelona, Spain
E-mail: rajesh.galivanche@intel.com, rohit.kapur@synopsys.com, antonio.rubio@upc.edu

Abstract

Testing today of a several hundred million transistor System-on-Chip with analog, RF blocks, many processor cores and tens of memories is a huge task. What will test technology be like in year 2020 with hundreds of billions of transistors on a single chip? Can we get there with tweaks to today's technology? While the exact nature of the circuit styles, architectural innovations and product innovations in year 2020 are highly speculative at this point, we examine the impact of likely design and process technology trends on testing methods.

I. INTRODUCTION

Predicting the future with specific details is always difficult. What makes this paper unusual is that we specifically target the year 2020 to get the benefits of clarity. After all isn't 2020 vision perfect? While the future of test depends on where technology ends up, test has isolated itself enough from the actual semiconductor technology that whether the devices are build in CMOS, strained silicon, or some other nanotechnology may not matter much. The logic built out of the process technology mostly impacts test methods.

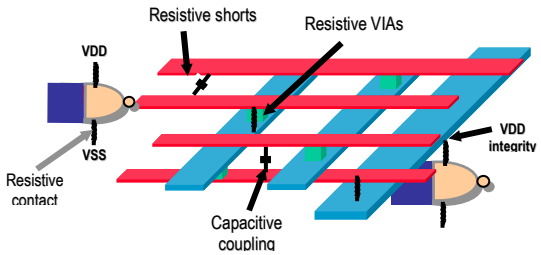


Figure 1: Defects that show up as a timing problem at the logic level.

Figure 1 shows the isolation between defects and test technology seen in ICs today. The defects shown are all targeted by test methods designed to test timing failures. This will be true for the technology of the year 2020. For example, the exact defects in trigate transistors may be very different, but what test will be more concerned about is the logical impact of the defects. The fact that the corner device turns on at lower voltages, due to the proximity of two adjacent gates, simply transforms into the test conditions necessary to observe the delay effect. Exact transistor structure or composition does not really matter much in this situation.

Fault models provide the necessary isolation of IC test to the actual technology the chip is built on. In the year 2020, we still envision some variant of the single stuck-at fault model leading the charge for test generation. Any new adaptations of the single stuck-at fault model will be determined by the new defect mechanisms of the future. In the next section, we discuss some failure modes that the test community will face in the years ahead. Process geometry scaling is providing the designers with doubling of transistors every process generation. However, utilizing all the available transistors requires some tough design choices. Many of the design choices will directly impact the future test methods and they will be discussed in the following section. Test itself is branching out in different areas. This trend is discussed in the section on test methods. Finally we discuss test automation. Scan DFT, ATPG have become the bread and butter test methods for logic test in the industry. Compression technology and power aware testing will become main stream much before the year 2020. In the section on automation we address other changes that will drive the dynamics of the tools used by the IC test community.

II. FAILURE MECHANISMS IN 2020

The likely technology scenario for year 2020 from the *International Technological Roadmap for Semiconductors (ITRS) Report* [2] points to usage of multiple-gate Si-CMOS based technology with 6 nm of physical gate length, a gate oxide thickness of 0.5 nm (t_{ox})and 0.5 volts of power supply voltage, possibly interfacing with blocks of specific functions based on new emergent devices (NED) [3]. In such a situation, we can expect the following:

- Low quality of switching components and the interconnect due to high variability in their specified

978-3-9810801-2-4/DATE07 © 2007 EDAA

(expected) behavior as a result of manufacturing process variations

- Very high device count per design
- Very low level of energy for unitary operations (a single write cycle in a memory or a logic gate transition) and
- High susceptibility to internal and external noise/perturbations.

Quality of components. MOSFETs and NEDs of the 2020's generation will be atomic scale devices. For example, the effective channel width of a 6 nm physical gate length implies a volume of 30x30x30 atoms. Moreover, the t_{ox} thickness will be formed by 2 or 3 stacked atoms. Variations and deviations of manufacturing processes and materials at this discrete level of material as well as the intrinsic discrete random distribution of dopants imply drastic variations in the electrical characteristics of the switching devices. ITRS predicts a variability of 112% in the V_{th} of the MOSFETs of that generation. Together with the manufacturing failure mechanisms of these devices, degradation factors that quantum effects will produce (leakage currents), variations-resilient circuit design methodologies and statistical validation techniques will likely be required in future.

High complexity ICs: Integration of hundred of billions of devices on a single die will be done leveraging the concept of regular structures or multi-core systems that achieve performance improvement through high level of parallelism instead of increasing the clock frequency of the circuits. The difficulty of synchronizing such complex structures will likely introduce GALS (Globally Asynchronous and Locally Synchronous)-like strategies into the main stream. This technique will result in highly asynchronous and non-correlated switching transient further leading to a complex and unpredictable electromagnetic activity. Design-for-Test features, fault tolerance and self calibrating design techniques will likely mitigate the validation challenge both during design and the manufactured silicon.

Very low energy level for unitary operation: The energy required to perform an unitary operation has been decreasing steadily decreasing over the last several decades as shown in Figure 2. The first samples have been taken from Landauer's works [4], the next from data of modern circuits and the predictions from ITRS. For the year 2020 the unitary operation energy is predicted to be lower than 10^{-18} Joules. As published by Landauer, Stein [5] and others, such low levels of energy required for unitary operations increases the probability of an error event caused by interference of external or internal noise sources. One of the side-effects of the switching device size scaling is the increased sensitivity of the circuit to noise. In the case of just thermal noise, even with low level of energy for unitary operations, the resulting error probability levels are below 10^{-40} [5] at room temperature. However, it has been shown that at the energy of cosmic particles, the error rate (SEU) is increasing with technology scaling to critical levels. It has been shown recently [6] that SEU upsets can be significant, if not higher, for logic circuits than memory cells. Moreover, transient errors are not only due to thermal activity or cosmic particle strikes but also due to internal electromagnetic noise. If the levels of energy corresponding to the internal noise caused by switching activity, ground bounce and crosstalk are taken into consideration, it can be shown that the noise energy level is two or three levels higher than that of thermal noise at room temperature. As a a result, modern circuits are getting increasingly sensitive to internal noise and this trend will get aggravated in future technologies.

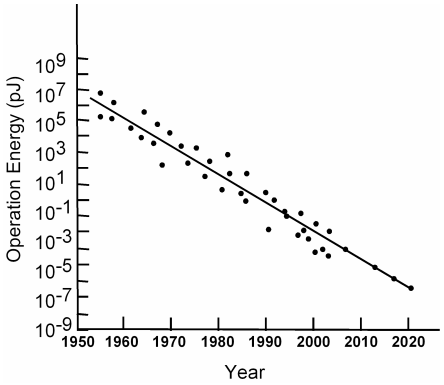


Figure 2: Evolution of the required energy to perform a single digital unitary operation.

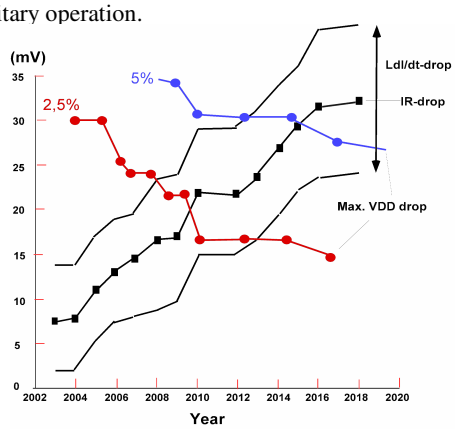


Figure 3: Predicted evolution of IR-drop [9] and LdI/dt noise versus trend of the maximum tolerable drop in V_{DD}.

Local impact of internal noise: The noise caused by the parasitic coupling of electromagnetic activity [7] inside an integrated circuit is a key issue in modern and future circuits [8]. ITRS mentions the imperative use of CAD tools for capacitive and inductive crosstalk evaluation in next generation circuits. Figure 3 shows the evolution of IR-drop for current and future technologies taken from the work of Meindl [9]. In the figure, the estimated levels bands when considering LdI/dt noise are also added. It can be concluded that in one decade from now, the levels of switching noise will be of the same level or higher than that of the acceptable drop in power supply (even for a 5% of acceptance). Sensitive parts of the circuitry as mentioned in the previous section will be affected by the noise resulting

in functional failures in the chips. Moreover due to the unpredictability of noise characteristics, design and test techniques have a serious challenge in addressing the noise induced failures. This will necessitate the use of resilient mechanisms in the architecture and design to cope up with this problem.

III. DESIGN IN 2020 THAT IMPACTS TEST

All indications are that current scaling trends will continue into year 2020 and that the mainstream high volume designs will continue to use transistors as switching devices, albeit in a form different than the transistors of today (such as tri-gate devices). Transistor integration capacity is expected to be in 100s of billions of transistors as shown in Table 1, doubling every two years in a trend predicted by Moore's Law.

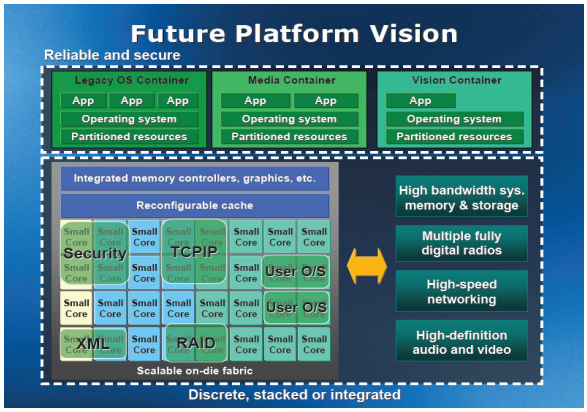
High Volume Manufacturing (Cross over year)	2004	2006	2008	2010	2012	2014	2016	2018
Technology Node (nm)	90	65	45	32	22	16	11	8
Integration Capacity (BT)	2	4	8	16	32	64	128	256

Source: Intel

Table 1: Projected Transistor Integration Capacity.

Availability of such large number of transistors enables several potentially new application paradigms. One such application is shown in **Source: Intel**

Figure 4 which depicts the future compute platform vision.



Source: Intel

Figure 4: Compute Platform of the Future

However, fully utilizing these available transistors requires addressing several challenges. In this section, we present key design challenges that impact manufacturing test and reliability in a significant way.

Power Management: Total power consumption which includes both active and standby (primarily due to leakage) power will be the primary constraint limiting full utilization of the available transistors. Overall product performance will be primarily realized through parallel computation by utilizing a large number of cores, which have been carefully optimized by trading off raw performance for lower power. As of today, the days of designs optimized purely for performance are a thing of the past, and optimizing designs for lower power dissipation plays an equally important role and this trend is expected to continue in the foreseeable future. The design will be optimized for power using a variety of techniques such as clock gating, transistor sizing, sleep transistors, multiple V_t devices, and low-leakage manufacturing processes and technologies. Apart from optimizing the cores themselves for power and performance, the entire design is optimized for power/performance/thermals using adaptive techniques such as voltage and frequency scaling, extensive clock gating and multiple standby modes designed to conserve power such as sleep states. As a side-effect of power and performance optimization, more paths become clustered in a narrow region around the cycle time, as shown in Figure 5, resulting in a larger population of paths which are sensitive to small delay perturbations [10]. As a result, high quality at speed coverage of each core will be necessary.

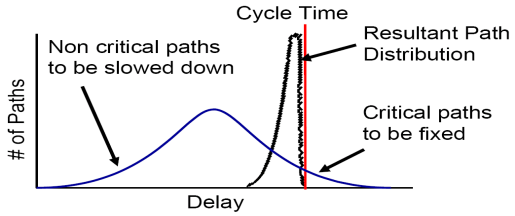
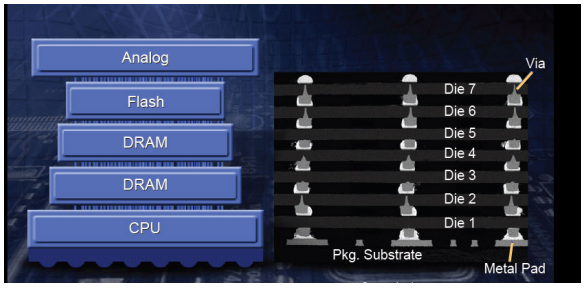


Figure 5: Effect of Power Optimizations on Timing

These techniques will pose a significant design and silicon validation challenge since many of the factors affecting silicon performance such as workloads that cause large current transients, simultaneous switching of multiple inputs, temperature changes, and interaction between various functional blocks that cannot be accurately accounted for during the design process.

An important ramification of increased number of computational cores is increased memory bandwidth requirements. Stacking of memory with the rest of the die is being actively researched and may become a reality in the future..



Source: Intel

Figure 6: 3D Stacking

A logical extension of memory stacking is stacking of other platform functionality as shown in Figure 6. The main advantage of stacking is not only increased bandwidth but

improved signal latencies since the signals travel shorter distances. 3D stacking with Through Silicon Vias (TSVs) provides several orders of magnitude more connections between the stacked die. On microprocessors, it is estimated that over half of the performance loss is from interconnect delay and over half of power consumption is due to wire capacitance. 3D stacking of partial die with TSVs is an attractive alternative to improving performance as well as active power consumption in future products.

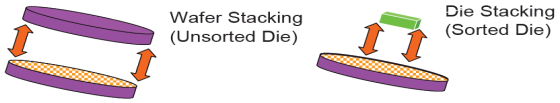


Figure 7: Stacking methods.

However, stacking of die poses significant test challenges. Blind wafer sorting may result in drastic yield reduction due to reasons such as no two wafer defect locations will be the same. Die stacking requires sorting of die before stacking. Sorting of die before pre-thinning the wafer is not possible since TSVs are buried in the wafer and are exposed only after thinning. Sorting of die after post-thinning poses mechanical problems since TSVs are very small (1um) and also the mechanical force applied during probing may not be sustainable on a thinned wafer. In summary, several test issues must be solved before 3D stacking can be realized but it appears to be a compelling technology to provide power/performance benefits.

Test Time/Test Data Volume: Another impact of massive integration of functionality is the increased test times/test data especially with integration of heterogeneous functional modules. Additionally, test content for more advanced fault models (such as circuit marginalities and parasitics) will place more demands on the tester memory. On the positive side, the presence of many cores of similar functionality provides new opportunities such as parallel testing or cores testing each other. To minimize test times while maintaining outgoing product quality, self-test and novel test data compression techniques which can take advantage of multi-core environments will be necessary in future products.

Wearout Effects: Wearout is due to many mechanisms as shown in Figure 8. Burn-in is the primary mechanism to screen infant mortality failures but current burn-in techniques will become prohibitively expensive in the future due to cooling efficiency (due to high static leakage currents) with massively parallel burn-in configuration. Additionally, the stress induced by the burn-in process itself can reduce the usable lifetime by introducing wearout. Also, other device degradation mechanisms such as NBTI and failures due to soft errors become much more rampant at smaller device geometries. These failures will manifest themselves only during field use and thus cannot be screened by any test performed during manufacturing prior to shipment. Efficient field testing methods coupled with spatial (redundant hardware) and temporal (retry) redundancy may become necessary to recover/repair the failing parts to meet product reliability requirements.

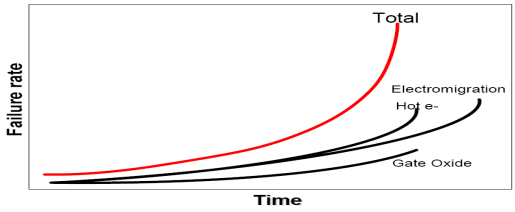
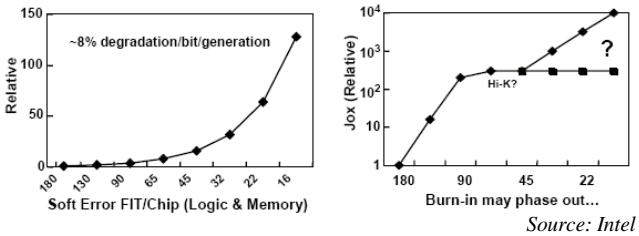


Figure 8: Wearout Failures as a Function of Time

The plot on the right hand side of Figure 9 shows how the chip level FIT rate increases form one generation to another. This increase is primarily due to the increased latch count and the array sizes in the designs. Soft errors cannot be tested during manufacturing testing – we need online and continuous test methods and recovery schemes.



Source: Intel

Figure 9: Reliability Issues with Scaling Technology

Process, Voltage and Temperature Variations: Circuit performance depends not only on the transport delays needed for holes/electrons to travel across a device or an interconnect, but also on environmental factors such as temperature, cross talk, power supply droop, V_{cc} variations, etc. However, accurate modeling for all these parameters during design is not practical due to the level of complexity and interactions with so many domains. If these cannot be designed out, many of them will manifest as speed failures in the silicon. One solution to this problem is to adequately

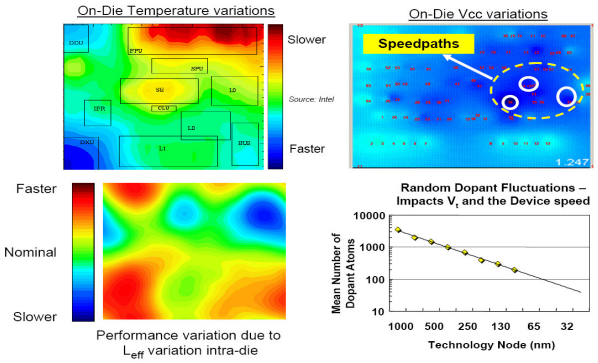


Figure 10: Sources of Speed Failures.

guardband against parameter variations and inaccuracy, but this leads to a sub-optimal and conservative design methodology which can leave a lot of performance on the table and lead to an unattractive product. Hence, the

common practice is to find these issues on silicon through post-silicon validation and make incremental design fixes through product steppings (if a large percentage of parts are affected) or screen them as a part of manufacturing test. Due to the complex interplay of workloads, wear-out mechanisms and process variations on the product performance, all failures may not be detected during manufacturing test (or even during post–silicon validation). Thus, screening such failures may have to be done in the field either through offline test techniques (where the part under test is taken out of doing useful work and subjected to test) or online/concurrent test techniques where the workload itself is continuously checked for correct operation. These techniques, though a common practice in today's high end mainframes, are not cost effective for high-volume mainstream products which are very cost sensitive. New low-cost self-test techniques will be required to guard against various sources of failures in the field that cannot be screened in the factory. In a way, such self-test techniques can also benefit manufacturing test by reducing the test data/time of future designs that are expected to be very large in size.

IV. TEST METHODS IN 2020

Now that test technology is being implemented across the board in every IC, the industry is beginning to leverage this fact for other uses. In that direction, there will be significant changes in the environments created around the test structures by the year 2020. In particular, the two areas that will blossom will be Debug and ATE environments.

By year 2020 testing will be done at different conditions for different die. Testing done on one die in a wafer will differ from the testing done on another. This automated learning environment will be done on the ATE itself. The timing tests data collected during testing of dies will be analyzed and the distributions of the responses will be used to enable different timing point selection on the ATE.

Another trend in test is the link between test and yield. Test provides the first sign that there is a yield problem. Providing the necessary insight into the problem is a responsibility of the test methods. As yield is projected to be a significant problem going forward, its reliance on test will increase.

In the future fault tolerance techniques at hardware and software levels will probably be mandatory to provide resilience/graceful degradation for soft errors, timing errors, defects, device degradation, and signal integrity issues. How these mechanisms will cooperate with BIST or off-line testing is another challenge for efficient test procedures and quality verification.

V. TEST AUTOMATION IN 2020

Automation for design and test is indispensable even in today's designs. It is important to ensure that automation technology that the EDA tools provide does not lag behind the needs of the various design styles to be implemented in the future. In this section we discuss test automation related methods that do not exist today but will be needed in the future.

Partition Aware ATPG: ATPG algorithms today operate on full-chip flattened design hierarchy. As the designs get much larger in future, we will see two trends. First is the new ATPG techniques that generate full-chip compatible tests by processing smaller partitions of the design. These partitions will be extracted exploiting the design hierarchy and by special using DFT techniques that enable tests generated at the partitions to be applied at the full chip level. Tests generated for one instantiation of a sub-design will be leveraged for testing other instantiations of the same sub-design. .Such techniques, which are already in use in some design houses, are expected to become main stream in the future. The required processing power for the ATPG tools will come from exploiting parallel processing techniques that use multiple CPU cores in a computer system. Second is the new ATPG techniques that generate tests at various levels of hierarchy partitions

Testing for Confidence: Test-time constraints for future designs is expected to overwhelm the test data compression solutions going forward thus necessitating ATPG research and technology in methods that limit the test application time. Optimizations in ATPG for test application time will span across fault models in a more uniform manner. Today, fault models are addressed one at a time without any optimizations across them. Test patterns not only will span the fault models more uniformly, the inherent knowledge of the effectiveness of each pattern will be used to apply tests in such a way that optimizations can be made in determining the goodness of a product. As test patterns are applied, the confidence of the quality of the IC keeps changing. Today the amount of testing done is not as tightly related to cost models and confidence calculations. In 2020, this will be the only way test will be done. Enabling this will need innovations in modeling the relationship between pattern count, fault coverage and confidence in the quality of the product.

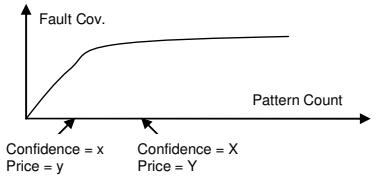


Figure 11: Testing for confidence.

Figure 11 shows test patterns and the relationship to confidence and price. The more the confidence the more the price paid for the testing services because testing for a higher confidence level will require applying more tests than testing for a lower confidence level.

Non-Determinism in Algorithms: ATPG algorithm technology today is not far from the papers we read on the

subject when we learnt test. Performance improvements of computers that the ATPG software runs on has been sufficient for ATPG to keep up with the exploding fault counts. Compression technology has kept the focus away from the inherent compaction algorithms in ATPG. As 2020 approaches we will see off-shoots of N-detect or TARO (Transition to All Reachable Outputs) fault model in the inherent ATPG search algorithm. Similar to branch prediction in the compute pipeline of a micro-processor, we are going to see more speculation in ATPG algorithms. ATPG will not deterministically complete the creation of a test for a fault. Instead, it will begin speculating the complete solution after creating partial solutions. Figure 12 shows the impact of such algorithms on the stimulus determined by ATPG. Fault simulation will still be used to determine the detection status and expected values of the faults.

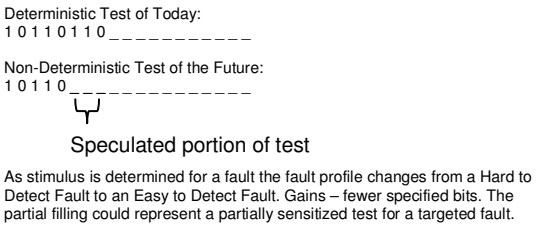


Figure 12: Non-determinism in ATPG

Benefits of such algorithms will be leveraged by the needs of compact test patterns and the fact that faults are not as deterministic when it comes to mapping to the actual defect mechanisms.

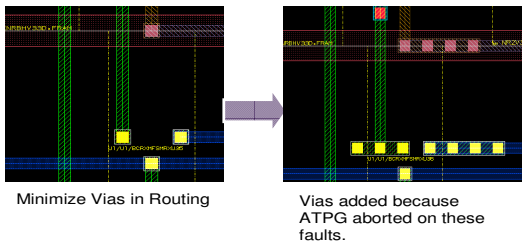


Figure 13: Layout strengthening its solution to counter ATPG's weakness.

Strength-Weakness Tool Interoperability: EDA as a whole is moving from a collection of point tools in a design automation solution to a suite of highly integrated tools that interoperate with one another. The general trend of tools to work closely together will continue to solve the complex problems that the technology brings forth. Test automation is not an outlier in this regard. Test automation today already is aware of layout, and recent developments have made it comprehend is power and timing issues in the design. As we go forward, signal integrity issues due to cross talk, power supply droop will need to be factored into the fault models. By the year 2020 the design automation system will be very sophisticated where the need to optimize across different tools will be in place. The strengths of the different tools will be used to mitigate the weaknesses of others. Figure 13 shows a situation where ATPG has a tough time detecting a fault, and the layout tool changes its solution to reduce the probability of that defect occurring. Similarly, synthesis can add redundancy in the design to mitigate problems in test.

Design Aware Test: Use of redundancy in the future designs to provide fault tolerance will make it important to understand the location of the defect to really classify the device as good or defective. If the failure is in a location that is tolerable then the testing process needs to understand it and categorize the chip accordingly. If adaptive designs are prevalent at that time, test will need to understand if the design can adapt around the failure that is discovered.

VI. CONCLUSIONS

In this paper a number of changes have been predicted for test of circuits in 2020. While a number of details have been given, the overall theme will be that test itself is moving into the mainstream design and manufacturing flow. Test will play a significant role ensuring higher yield manufacturing of ICs but not without solving a set of challenges.

VII. REFERENCES

[1] F. Ferhani and E. J. McCluskey, "Classifying bad chips and ordering test sets," IEEE International Test Conference, 2006, Lecture 1.2.

[2] International Technology Roadmap for Semiconductors, http://www.sematech.org, 2005.

[3] Emerging Research Devices, ITRS, http://www.sematech.org, 2005

[4] R. Landauer, "Dissipation and noise immunity in computation and ommunication", Nature 335, 779-784.

[5] K.U. Stein, "Noise-induced error rate as limiting factor for energy por operation in digital IC's", IEEE Journal of Solid State Circuits, vol. 15, issue 5, Oct. 1977, 527-530.

[6] P. Shivakumar et al., "Modeling the effect of technology trends on the soft error rate of combinational logic", IEEE Proc. Int. Conf. on Dependable Syst. and Net., June 2002, 389-398.

[7] X. Aragones et al., "Noise generation and coupling mechanisms in deep submicron ICs", IEEE Design and Test, vol. 19, issue 5, 2002, 27-35.

[8] X. Aragones and A. Rubio, "Challenges for signal integrity in the next decade", Pergamon Material Science in Semiconductors Processing", 6(2003) 107-117.

[9] K. Shakeri and J.D. Meindl, "Compact physical IR-drop models for chip/package co-design of gigascale integration (GSI)", IEEE Tr. on Electron Devices, vol. 52, no. 6, June 2005, 1087-1095.

[10] T.W. Williams et. al., "The interdependence between delay-optimization of synthesized networks and testing," IEEE Design Automation Conference, 1991, pp. 87-92.