
# INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

## 2011 EDITION

## TEST AND TEST EQUIPMENT

# LIST OF FIGURES



# LIST OF TABLES

# 1 INTRODUCTION AND SCOPE OF THE 2011 EDITION

The 2011 edition of the Test Roadmap contains some significant changes to many of the tables; includes a new section on 3D/TSV testing. A survey on Cost of Test was completed in 2009 and again this year and the results are included in the Cost of Test Focus topic section. The SOC/DFT and Logic sections also have been significantly reworked.

The Cost of Test section shares the result of our survey which unfortunately shows that test cost is an ever increasing concern for the test community. While there have been many efforts to confront this, the reality is that increasing complexity (device size, power domains, high-speed pins, and probe cards and interface boards to name a few) are having an opposite effect on the cost of test challenge.

With roll-out of 3D devices (stacked die with TSV and the older SiP devices) the challenges for the test community are rising to a whole new level. As we look to the future we have to remain focused on an efficient test approach which will simultaneously confirm the proper functioning for multiple interconnected devices within an environment which will likely include multiple technologies and vendors.

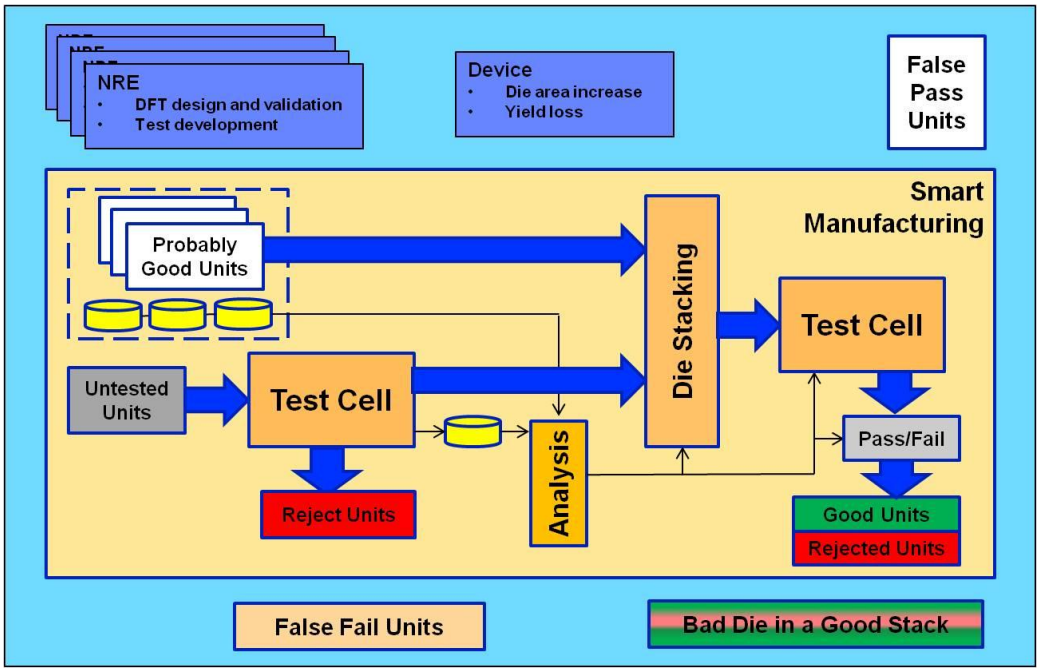



*Figure TST1 - Smart Manufacturing Test Flow for Stacked Die Devices*

As shown in the figure above, this change to stacked die devices has the prospect to significantly complicate the testing sequence with multiple test insertions (each with adaptive data driving the test focus). In addition, this flow will likely generate a new class of device failure – one where most, but not all, the devices in a given stack are good.

The logic and SOC tables have been refocused with this update. In the past we calculated the required pattern compression ratio needed from the EDA community in order to keep the test time flat. While this release we have sought out the assistance from multiple EDA companies in order to directly calculate realistic test data volumes and the resulting test times with these more accurate volume and compression ratio forecasts. This approach has a number of key benefits including 1) Showing clearly the relative impact of various pattern generation techniques thereby helping informed tradeoffs to be done, 2) A realistic estimate of the requirement pattern memory needed in ATE systems moving forward, and 3) An estimate of likely device test times.

The specialty devices chapter shows the influence of the mobile phone market with solutions in the digital camera and MEMS sensor areas.

The remaining portions of this document have been thoroughly reviewed and only minor changes were done.

This document represents significant contributions from a large number of participants representing a wide cross-section of the industry as noted in the acknowledgements.

# 2 DRIVERS, CHALLENGES, AND FUTURE OPPORTUNITIES

Starting with the 2005 edition of the ITRS Test Chapter, this important section of the test chapter was refined by splitting the Difficult Challenges section into Key Drivers and Difficult Challenges and adding a Future Opportunities section. This split distinguishes the drivers, which are primary boundary conditions that define the scope of solutions for upcoming manufacturing test for semiconductor components, from key technical and business challenges. At a high level, these boundary conditions actually represent expectations or even requirements of the test process, while the challenges represent current and upcoming key roadblocks, strategic inflection points, or opportunities for the future. The key drivers determine many of the concepts and table values in this chapter. The difficult challenges define areas where more development or understanding is needed in order to cost effectively meet the semiconductor roadmap.

For many years, the mission of semiconductor manufacturing test has been described as "screening defects" and to a lesser extent or within certain business segments "speed binning" or "speed classification." It is interesting to note that some of the most important test challenges are now actually centered on some of the more subtle historical missions of manufacturing test—reliability and yield learning. It is also important to note that the impact of these challenges affect not only on the manufacturing test process itself, but are essential to the entire semiconductor business, both in terms of enabling the cadence or timely delivery of future processes and cost effective products, but also in terms of meeting customer expectations for reliability. In addition, the demand to decrease device Time-to-Market has fused the realms of device debug and production test, increasing the scope of "manufacturing" equipment to include many functions previously addressed in engineering design laboratories.

Within the Difficult Challenges section, the challenges are listed in order of perceived importance or priority. For example, test for yield learning is followed by screening for reliability and is then followed by increasing systemic defects. In contrast, there is no specific intent in the ordering of the Key Drivers, as they are all boundary conditions or requirements that the semiconductor test solutions must meet. *Table TST1* summarizes all of the key test drivers, challenges, and opportunities.

## Table TST1 - Summary of Key Test Drivers, Challenges and Opportunities

| **Key Drivers (not in any particular order)** | |
|---|---|
| *Device trends* | *Increasing device interface bandwidth (# of signals and data rates)* |
| | *Increasing device integration (SoC, SiP, MCP, 3D packaging)* |
| | *Integration of emerging and non-digital CMOS technologies* |
| | *Complex package electrical and mechanical characteristics* |
| | *Device characteristics beyond one sided stimulus/response model* |
| | *3 Dimensional silicon - multi-die and Multi-layer* |
| | *Multiple I/O types and power supplies on same device* |
| | *Fault Tolerant Architectures and Protocols* |
| *Increasing test process complexity* | *Device customization during the test process* |
| | *Feedback data for tuning manufacturing* |
| | *Dynamic test flows via "Adaptive Test"* |
| | *Higher order dimensionality of test conditions* |
| | *Concurrent Test* |
| | *Maintaining Unit level Traceability* |
| *Continued economic scaling of test* | *Physical and economic limits of test parallelism* |
| | *Managing (logic) test data and feedback data volume* |
| | *Managing interface hardware and (test) socket costs* |
| | *Balancing Tool Capability, Multiple Insertions, System Test and BIST* |
| **Difficult Challenges (in order of priority)** | |
| *Cost of Test and Overall Equipment Efficiency* | *Continues to be the primary driver for innovation. Traditional drivers for COT are started to be limited by OEE* |
| *Test Development as a gate to volume production (Time to Market)* | *Increasing device complexity driving more complex test development* |
| *Detecting Systemic Defects* | *Testing for local non-uniformities, not just hard defects* |
| | *Detecting symptoms and effects of line width variations, finite dopant distributions, systemic process defects* |
| *Screening for reliability* | *Implementation challenges and effectiveness of burn-in, IDDQ, and Vstress* |
| | *Erratic, non deterministic, and intermittent device behavior* |
| | *Mechanical damage during the testing process* |
| | *Multi-die stacks/TSV* |
| | *Power Management Issues* |
| **Future Opportunities (not in any order)** | |
| *Test program automation (not ATPG)* | *Automation of generation of entire test programs for ATE* |
| *Scan diagnosis in the presence of compression* | *Collect better yield improvement and scan debug information* |
| *Simulation and modeling* | *Seamless Integration of simulation and modeling of test interface hardware and instrumentation into the device design process* |
| *Convergence of test and system reliability solutions* | *Re-use and fungibility of solutions between test (DFT), device, and system reliability (error detection, reporting, correction)* |

*ATE—automatic test equipment ATPG—automatic test pattern generation BIST—built-in self test HVM—high volume manufacturing MCP—multi-chip packaging MEMS—micro-electromechanical systems*

# 2.1 KEY DRIVERS

As previously mentioned, the key drivers in the Test chapter are considered *boundary conditions* within which the semiconductor test function must operate and still deliver acceptable component quality and reliability at continued test cost scaling.

## 2.1.1 DEVICE TRENDS

### 2.1.1.1 DEVICE INTERFACE BANDWIDTH

In the 2001 edition of the ITRS Roadmap, a change in direction was denoted to deliver increasing system performance with rapidly increasing (compared to previous trends) component I/O speeds. Beginning with the 2003 edition, the Test working group noted that while the clock speeds are not climbing as rapidly as the 2001 ITRS expected, the penetration of Gb/s interfaces is permeating throughout a wider spectrum of the semiconductor component market segments (memories, CPUs, chipsets, etc.). The design clock speed roadmap was changed in 2008 and 2011roadmap editions to reflect the lower GHz due to architecture and transistor scaling. High speed serial and differential I/O protocols will continue to rise in speed and continue to force innovation in the DFT and HVM test processes to maintain continued cost scaling.

### 2.1.1.2 INCREASING DEVICE INTEGRATION

SoC and SiP integration continue to rise in prevalence throughout a number of business segments. Increased device integration forces a re-integration of test solutions to maintain scaling of test costs and product quality. The optimized test solutions for stand-alone RAMs, cores, and other blocks typically do not scale linearly without modification, additional DFT, or new test partitioning solutions. In particular, incremental DFT in-die or even in-package may be required to provide access to and testing of embedded blocks and cores. Additional distribution or re-integration of the HVM test instrumentation may also be required. Techniques for known good die (KGD) that provide high quality dice for multi-die packaging also become very important and an essential part of the test techniques and cost trade-offs.

### 2.1.1.3 INTEGRATION OF EMERGING AND NON-DIGITAL CMOS TECHNOLOGIES

In the recent past, mixed-signal device circuits have been increasingly integrated on the same die with digital CMOS, posing key challenges for ATE, instrumentation, and test manufacturing flows. It also poses new challenges and opportunities for DFT innovation, where mixed signal and digital had not been previously prevalent (as it had in digital logic and memories). Integration of RF circuits is already starting to experience a similar rise in innovation. More radically different types of semiconductor devices, such as MEMS (which are already showing up in integrated forms) and optical are being integrated on a single piece of silicon with CMOS logic . As in previous types of integration, the test mission for these devices revolve around access to potentially embedded blocks and integration of radically different test methods into a cost effective manufacturing process.

### 2.1.1.4 PACKAGE FORM FACTOR AND ELECTRO-MECHANICAL AND THERMAL CHARACTERISTICS

The envelope of package form factors is pushing both outward and inward, with higher complexity form factors in MCP and SiP for multi-die on the high end, but also push chip-scale packaging towards the small end for systems targeted for the smallest platform form factors (e.g., handhelds). In addition, multi-functionality of die packaging with inclusion of heat sinks and spreaders, and perhaps even voltage regulation and power management functionalities, appears to be on the increase. With increasing power a major concern, form factor thermal transfer characteristics and uniformities (thermal gradient coefficient-junction to ambient, junction to case [theta-JA, theta-JC], etc.) become even more critical to the test process, where power requirements are typically higher than they are for the device's end-application. These expansions of the form factor technology envelopes necessitate improvements and delivery of key test sub-systems: handling for packaging test, Ohmic contacting technologies, and test sockets (the latter described later in the chapter for the first time).

### 2.1.1.5 DEVICE CHARACTERISTICS BEYOND DETERMINISTIC STIMULUS/RESPONSE MODEL

The history of semiconductor testing and ATE architecture and usage is built upon a foundation of deterministic device behavior. In addition, digital CMOS testing has historically been based on simplifying the effects of Vcc, temperature, and frequency, so that worst case (one-sided), one iteration tests guaranteed the performance over a wider and continuous range of environment topologies. When these topologies were driven by only two or even three variables, they were manageable and fairly easy to characterize and to optimize for HVM test. However, a number of device characteristics threaten the dimensionality of this one-sided test paradigm, threatening a non-linear effect on non-recurring engineering (NRE) development costs for test as well as potentially limiting continued HVM test cost scaling. Device features such as extended self-repair and correction and built-in variable multi-power management modes potentially add tremendous complexities to the device test conditions contour. In addition, from the device design, architecture, and behavior front,

there now looms the problem of non-deterministic device behaviors (e.g., asynchronous logic architectures and fault tolerant devices) where correct die to die device behavior (from end user system standpoint) under similar conditions is not deterministic within the time/vector synchronization standpoint. These behaviors, while correct from an end use customer and system standpoint would break the historical HVM test stimulus-response model, where one set of digital logic 1s and 0s delivered on the same synchronization, would suffice to test such behaviors. These device architectures and behaviors pose either a major paradigm shift or innovation challenge for the semiconductor test process.

#### 2.1.1.6 3D SILICON DEVICES

Developments in TSVs have enabled the ability to connect die together using tens, hundreds, thousands, or even millions of connections that could be as small as 2 $\mu m^2$ per connection. Because of this, it is now possible to design a system that spans multiple die domains and is partitioned across multiple die to optimize product cost and performance. A typical monolithic integration put many board components on a single die. 3D integration is simply taking the many components that are on a board, re-architecting the system to exploit the performance improvements that can be achieved from using 3D interconnects, properly partitioning the entire system for cost and manufacturability; and stacking the die vertically. As an example, an MPU can have the logic cores on one die, the RAM cache on a second die, and NVM storage on a third. The MPU example would require three dimensional design capabilities so the thousands of cache connections (TSVs) to the cores are defined to be in the same X and Y locations. Feed thru vias to the NVM memory must go from the logic die thru the cache die to the NVM memory die. In addition to connecting the die, since each die must be able to be tested on its own at wafer probe test, test insertion must be complete for a single die domain. As die are stacked on top of each other, connections are made with the TSVs and the multi-die stack must be able to be tested as a semi-completed system using the test insertion. Vectors and paths are likely to be different from those used for testing of the individual die. A further complication arises as the TSVs will likely not have any ESD protection because ESD structures will greatly increase the die area cost of 1000's of TSVs. Physically probing non-ESD protected pads or TSV nodes is likely to damage the die. For this reason, it is recommended that the test solution NOT contact the TSVs during test, but to use an alternative method such as boundary-scan or other non-contact means to access the TSVs pads.

#### 2.1.1.7 MULTIPLE I/O TYPES AND POWER SUPPLIES ON SAME DEVICE

Integration of devices in SoC and SiP will drive and increased number of power supplies in order to optimize the power consumption per core. Logic, analog, memory, mixed signal, and high speed I/O each optimally operate under different power supply conditions. Using multiple supplies versus attempting to generate or regulate voltages on chip saves die area and power and helps in achieving greater noise isolation across the cores. Single ended and differential I/O will be used to optimize power while not limiting the data transfer .performance.

#### 2.1.1.8 FAULT TOLERANT DEVICES

Device architectures are being developed to ensure that a device or system will never fail, or if it needs to fail, "gently" fail. These architectures implement various forms of error detection and correction to constantly monitor operation. If a fault is detected, the system can adjust to the fault by: correcting the error based on a correction algorithm; performing the operation again in hopes it was a onetime fault; changing the operational conditions of the functional block that experienced the error such that the functional block begins to operate correctly and yield a correct result; or disable the malfunctioning functional block and routing the information elsewhere for processing. Only when all possible options for recovery have been evaluated and failed, will the device or system fail. Fault tolerant devices represent a challenge to testing as they are architected to rarely if ever fail. In order to adequately determine if the device is functioning properly as it leaves the factory, test must be able to disable or reduce the capabilities of recovery mechanisms in order to evaluate the device or functional block for correctness.

### 2.1.2 INCREASING TEST PROCESS COMPLEXITIES

#### 2.1.2.1 INCREASED DEVICE CUSTOMIZATION AND LINE ITEM COMPLEXITY (DURING THE TEST PROCESS)

Increasingly, the test process is expanding to include not only the metrologies surrounding the question "is this a good die?" but also actual process steps that modify, differentiate, or customize specific die. Examples of these include memory block (and other) redundancy/repair, burning of on die configuration fuses, programming of read-only memories (ROMs), or other programming of product features. In addition, some business segments are seeing substantial increases in product line items derived from the same die base, with differentiation and physical segregation as part of the manufacturing test flow. All of this places increasing demand on the manufacturing test process and expands equipment (e.g., handling, fuse blowing, etc.) and factory automation and integration infrastructures.

#### 2.1.2.2 INCREASED TEST DATA FEEDBACK TO TUNE MANUFACTURING

Test data usage for purposes beyond identifying whether a given die is good or defective has become essential for several reasons and drives a need for expanded, revamped, and better integrated test data systems and infrastructure. In one sense the output of the component test process is data, i.e., the results of the various applied metrologies of the test process across the product dice and manufacturing lot populations. The need for better integrated usage of this test output (data) for fab process yield learning, maverick material identification, and feedback within a more distributed manufacturing test are all becoming more critical and even essential applications moving forward, not just nice to haves.

#### 2.1.2.3 DYNAMIC TEST FLOWS VIA "ADAPTIVE TEST"

An emerging method to reduce overall test cost is to break the paradigm of running a constant test flow and instead, instituting a method whereby results of recently tested die or units can be used to determine the probability of need for a particular test. If the "importance" of a test or group of tests is deemed statistically low, then the test(s) can either be temporarily dropped or have other tests substituted. Probability triggers can also be used to increase the tests that are executed.

#### 2.1.2.4 HIGHER ORDER DIMENSIONALITY OF TEST CONDITIONS

Historically, component manufacturing test has used testing at a simple matrix of points or corner conditions of two or three environmental variables (typically Voltage, Temperature, and Frequency) to guarantee the wider multi-dimension contour and interior space of product specifications beyond the contour outlined by the test points. Components in a broad range of markets from battery application platforms to client computing and servers are adding numerous schemes of power management that quickly add exponential complexity to the worst-case test conditions contour. For example, the complexities of component schemes may include multiple or variable power modes (sleep, hibernation, etc) or even multiple and *in situ* responsive variable Vcc and frequency control systems to achieve optimized power-performance in the application. The challenge of determining, characterizing, and optimizing (reducing) the larger set of environmental test points for component manufacturing test is daunting with the additional dimensions of these variables. Keeping pace with this complexity requires additional validation efforts and innovative methods to validate the quality of the more complex environmental test points. This is needed to ensure the lack of holes and predictability of more complex worst case conditions in the product test validation and test development phase. This increased complexity of test environment set points will also challenge the continued economic scaling of manufacturing test cost by the product of the applied test content and the number of iterations of its application at the various set points in the manufacturing flow.

#### 2.1.2.5 CONCURRENT TEST WITHIN A DUT

The "next level" of parallel test is to currently test multiple functional blocks within a single DUT, in conjunction with multisite test, where concurrent test is replicated per site. This capability is mostly enabled by Design-For-Test capability of the DUT itself. Concurrent Test of SoCs has been gradually introduced. The increased use of multi-die System-In-Package however accelerated this trend since multiple die naturally have the ability to be tested in isolation (as indeed is the case during the wafer probe insertion). The two major requirements driven in the test arena are: The ability of tester hardware to be separable into functionally independent blocks, and the ability of test software to independently manage independent "flow domains" and combine them to construct a complete test program. This later requirement is complicated by specific DUT capabilities in terms of signal access, crosstalk issues and power management.

#### 2.1.2.6 MAINTAINING UNIT LEVEL TEST TRACEABILITY

As a part of a manufacturing data feedback process, test data that is traceable back to individual die must be collected and managed. This capability will become more imperative with the increase in volume of Multi-die packages in order to enable root-cause analysis of failures of complex die stacks after assembly.

### 2.1.3 CONTINUED ECONOMIC SCALING OF TEST

#### 2.1.3.1 PHYSICAL AND ECONOMIC LIMITS OF FURTHER PARALLELISM

Over recent generations, continued increases in parallelism (in number of DUTs tested in parallel at a test insertion), particularly for commodity memories, but also for digital logic, have been a primary means of continuing the economic scaling of test in the context of devices with more transistors, increased functionality, and higher I/O and core speeds. In the current test tool and interface hardware integration paradigms, further increases in DUT test parallelism are reaching non-linear limits and the impact will be further seen in future generations. These are driven by the practical limits of how many electrical channels can be squeezed into the physical space between the parallel DUTs and the test instrumentation while yet maintaining acceptable physical and electrical proximities between the two. These approaching physical limits will require alternate means be used or expanded in order to maintain continued economic scaling of test, or alternately,

new paradigms of DUT, handling, contacting, and test instrumentation integration must be developed that enable further increases in DUT parallelism beyond what is currently envisioned.

Additionally, the economic limits of multisite test, especially for lower volume SOC devices are being reached in that Cost of Test no longer scales well – or at all – with the number of sites. Test cost reductions achieved by faster test times or more sites are being asymptotically limited by "background" issues such as Overall Equipment Efficiency, dominated by factors such as factory efficiency, lot sizes and equipment downtime.

#### 2.1.3.2 MANAGING (DIGITAL LOGIC) TEST DATA VOLUME

Increased digital logic die complexity and content drives proportional increases on the test data volume (number and width of vectors). Unconstrained, this additional test data volume drives increases in test capital and operational costs by requiring additional vector memory depth per digital channel of the test tools (ATE) and by increasing test time per DUT. Currently, a number of logic test vector compression schemes are being developed and applied in a variety of ways on the test databases themselves (for scan based tests) or via compression hardware (DFT) on the product die itself. Moving forward, compression will become more ubiquitous across component business segments, driven by the increasing product complexities and higher levels of product integration (e.g., SoC, SiP) and may ultimately require increases in the rate of compression (i.e. the compression ratio of the test database versus uncompressed).

#### 2.1.3.3 MANAGING INTERFACE HARDWARE AND (TEST) SOCKET COSTS

The portion of costs based on test and probe interface hardware and test sockets is an increasing proportion of the overall test cost. There are a number of factors driving this, such as higher speed (Gb/s), more complex DUT I/O protocols, increased DUT parallelism, higher signal and power pin counts, and increased power delivery and signal channel fidelity requirements. While this might be an alarming trend by itself, this trend needs to be considered in the context of the overall test process cost, and whether or not it enables a continued overall economic scaling of manufacturing test. Managing interface hardware costs within an acceptable range may also be dependent on technology drivers and boundary conditions, e.g., the continued extension of the use of FR4 materials within an acceptable numbers of layers as the primary materials platform for such hardware.

#### 2.1.3.4 BALANCING TOOL CAPABILITY, MULTIPLE INSERTIONS, SYSTEM TEST AND BIST

The "dynamic range" of test equipment is being stretched to include testing that differs from traditional structural or functional test. The balance being struck here is cost per second of general-purpose (GP) test equipment versus the additional costs of multiple insertions on more focused equipment. On the "lower end", tests which can be accomplished purely with BIST require that test costs not suffer the overhead associated with traditional test equipment. On the "higher end", devices which are constructed from multiple die will require some sort of "System Level" test in order to verify that a device is functional in its end application as opposed to simply repeating the structural test list which is typical done during SOC package test. This requirement will drive a need for new instrumentation and capabilities

## 2.2 DIFFICULT CHALLENGES (IN PRIORITY ORDER)

### 2.2.1 COST OF TEST AND OVERALL EQUIPMENT EFFICIENCY

A great deal of progress has been made over time to reduce cost of test, evidenced by the fact the percentage of semiconductor revenue used to pay for test has steadily decreased over the last 10 years. Economics will continue to be the main driver for equipment innovation in areas such as test time, multisite efficiency, concurrent and adaptive test and facility resources (power and floor space). It is clear, however, that test cost – especially in "high mix low volume" test environments - is now limited by factors outside the test equipment per se. Overall Equipment Efficiency (OEE) is generally defined as the percentage of time that equipment is used to do produce shippable product. In other words, good OEE results from minimizing "idle" time of equipment. For lower volume devices, reducing test time or increasing site count does not have a tangible effect on Cost of Test. Factors such as beginning-of-lot and end-of-lot delays (where the test cell has no parts to test) begin to dominate. Other significant causes of idle time are issues such as contactor failure jam rate, power interruptions, and device soak time. To make progress in this area, it will be necessary to expand the scope of work to include processes both inside and outside the test cell itself

### 2.2.2 TEST DEVELOPMENT AS A GATE TO VOLUME PRODUCTION (TIME TO MARKET)

The additional complexity of devices to tested detailed above drive a corresponding complexity of test development, particularly in the SOC space where analog components, non-deterministic behavior, and multi-die packaging drive a corresponding increase in test development complexity. This increased complexity drives up the cost of test development

and, more importantly opportunity costs that are the result of a delayed market entry due to the unavailability or ineffectiveness of tests developed in time to meet market windows.

### 2.2.3 DETECTING SYSTEMIC DEFECTS

The industry faces new manufacturing-imperfection-related test and yield-learning challenges that result from changing processing technology, changing circuit sensitivities and design modeling limitations.

- Increasing design-process interactions are increasing the prevalence of systematic defects. Such defects may occur only in certain circuit/layout configurations, due to, for example, pattern density, pattern proximity, and imperfections in optical proximity correction (OPC) algorithms. Although they are systematic, they may appear random because of their rarity and the complexity of the conditions required for their occurrence.
- Process-technology advancements change the population of physical defects. For example, changes in materials, such as those needed to support high-k dielectrics and metal gates, change defect populations.
- Changing circuit sensitivities and increased variability are likely to make defects that were benign in the past become killer defects in the future. For example, shorter clock cycles mean defects that cause picoseconds of delay are more likely to cause circuit failures. Furthermore, power-optimized and/or synthesized designs will result in fewer paths with significant timing margin, which implies that random delay-causing defects will more likely cause failure. Similarly, increasing noise effects, such as crosstalk and power/GND bounce, decrease noise and timing margins and again increase circuit susceptibility to defects.
- Modeling complexity and aggressive design threaten the ability of EDA/design to ensure that a circuit meets power and performance specifications under all process conditions. That decreased ability, combined with increasing process variability, may increasingly result in circuits with parametric failure modes that reach the test floor.
- Smaller transistors may increase the importance of degradation mechanisms, such as Negative (Positive) Bias Temperature Instability.
- Radiation-induced soft error rates are increasing to where, in addition to SRAMs, latches and flip-flops are likely to need protecting, at least for chips targeting enterprise applications. Other problems now increasingly observed include multiple-adjacent-cell SRAM upsets due to single radiation events and erratic shifts in minimum-operating-voltage, problematic for low power applications.

All aspects of the test process, including defect modeling, test generation, test-coverage evaluation, DFT solutions, test application and diagnosis, must handle these realistic and changing populations of manufacturing/operating imperfections. Promising strategies include out-of-spec testing such as low-VDD or temperature, statistical methods, adaptive test, and realistic defect-based modeling/targeting.

### 2.2.4 SCREENING FOR RELIABILITY

A lesser-publicized mission for semiconductor test beyond the primary "screening defects" or telling a good unit from a bad one has been to screen out infant mortality of the product population to acceptable levels. Another way of describing it more aligned to the "defect" mission of Test is to call this essential function "screening $t > 0$ defects" where "0" on the timeline represents the date product moves to the customer from the component provider. Historically, different semiconductor business segments have used a variety of techniques from burn-in to IDDQ, to voltage stressing during the manufacturing process to identify and screen sufficient numbers of the less reliable sub-populations or product dice to meet customer quality expectations. From a similar set of causes, notably increasing background leakage currents, and reducing product operational margin (collapsing Vdd/Vcc with process scaling), all of these techniques are becoming both less effective and more expensive to varying degrees. Where burn-in equipment and techniques had remained essentially unchanged and re-used over many process generations from the early 80s to the mid 90s, beginning in the late 90s burn-in production systems (in product segments that leverage this technique) have been one of the areas of the largest increases in test capital and interface hardware costs. At the burn-in elevated voltage and temperatures needed to accelerate latent defects, leakage levels are much higher than under normal application conditions. In addition, reducing product Vcc and temperature margins limit the range these conditions can be used to accelerate the latent defect populations. IDDQ, which has been widely used in product segments from ASIC to SoC to commodity memories, has been greatly challenged by the very same trend in proportional increases in background leakage, which greatly reduces the signal to noise ratio of "normal" static current levels versus DUTs that contain latent defects. In fact, this has been a challenge at least since the 250 nm DRAM half pitch, where many companies, began using more advanced techniques, for example, IDDQ "delta" and other Boolean static Icc/Idd combinations and comparisons for results of multiple tests to maintain efficacy against latent defects and reduce invalid yield "overkill" cost effects. Commodity memories, which

have tended to have somewhat lower intrinsic leakage levels per technology generation, have managed to extend and depend on the various IDDQ techniques a little farther than have other product families, but even they are now forecasting a real drop off in effectiveness fir IDDQ in the next one to two technology generations. Similarly, voltage stress, or applying Vcc/Vdd and patterns well above nominal range to accelerate latent defects, also has been losing effectiveness as the differential between *Vstress* and *Vnominal* has continued to get smaller at each generation. In the long term, new techniques for providing this reliability screening function will likely be needed in some business segments and product types sometime in the next few generations. Some of the newer concepts under exploration are improved Boolean and distribution algorithms among various test results both intra-die test (results from the same die) as well as inter-die (wafer neighborhood analysis, analyses within lots, adaptive test limits, analyses across lots, etc.). Another vector that will likely play a larger role moving forward is correction, whether by self-test and self-correction, which is being implemented on some embedded RAMs today, or by error coding detection and correction (ECC) techniques. The challenge here will not be on memories, but rather how and when similar capabilities or alternate approaches to self-correction might be practically applicable and affordable for logic.

## 2.3 FUTURE OPPORTUNITIES

### 2.3.1 TEST PROGRAM AUTOMATION

While the EDA industry has been providing a range of capabilities around scan DFT and more recently including scan compression and even some DFT synthesis capabilities, for example, for embedded memories, there is an area where the semiconductor suppliers software NRE costs present a productivity and automation opportunity: generation of actual test programs, especially for SOC devices with a large amount of analog content or non-deterministic interfaces. Across the industry today, the product or test engineering communities spend much of their time creating new customer software code bases, that is, unique and custom individual production test programs. There are a lot of innovative software products for different pieces of this, provided primarily by ATE suppliers along with the ATE, as well as a lot of additional software tooling done in house by the semiconductor companies themselves. There are even historical or new test program data structure standards (test data format (TDF) or standard test interface language (STIL)) that provide some common structural opportunities. In addition, new opportunities are becoming available with the ability of test equipment to allow abstraction of device IO functions to a High-Definition-Language format that relies on real-time adaption to Device IOs. By the same token, new complexity has been added with the added equipment capability to perform concurrent testing of separate device functions as permitted by DFT. Taking advantage of this functionality would add another axis to Test Program Automation problem. However, the basic fact remains is there are no widely available turn-key software products that produce and validate whole main code test programs and yet most of the 1000s of these in use across the industry today do basically the same kinds of things on similar tools across companies.

### 2.3.2 SCAN DIAGNOSIS IN THE PRESENCE OF COMPRESSION

There are several tools now available for debug of scan patterns and device faults using scan. A major impediment with the proliferation of scan compression is the ability to quickly and easily selectively disable compression to gain better visibility to individual faults. Currently, this requires time-consuming regeneration of patterns offline, thereby making the entire process far less efficient than other debug operations and it also disallows any ability to gather non-compressed failure data during production.

### 2.3.3 SIMULATION AND MODELING

Extending the design phase simulation systems and modeling to cover not only die packaging, but also test interface hardware and perhaps even the ATE instrumentation circuits themselves (both parametric and logical) will enable shorter product and manufacturing test validation cycles. This extension will also enable higher reliability pre-silicon verification of the DUT DFT scheme integrated to the manufacturing test instrumentation. Providing such a seamlessly integrated simulation and modeling environment in the pre-silicon design phase would better enable not only accuracy of product performance, but also help tune the test process, and reduce design turns due to minor discrepancies in DFT and test interface modes and hardware.

Additionally, the ability to simulate at a functional level for multi-die packages would facilitate System-Level testing for complex SoCs. This level of simulation can be achieved at a coarser granularity than traditional device simulation, needing only to generate the sequence and boundary conditions of the high-level transactions occurring at the Device Under Test.

### 2.3.4 CONVERGENCE OF TEST AND SYSTEM RELIABILITY SOLUTIONS

Some of the most widely proliferated DFT techniques (such as scan) were actually originally motivated by the need to provide better system reliability for high-end enterprise computing systems. The same scan hardware was re-usable for manufacturing test screening purposes and has proliferated throughout the industry. In the future, there are a number of physical and circuit phenomena that are challenging lower level intra-die reliability such as alpha particle and cosmic ray induced soft errors, erratic circuit behaviors, and other increasingly intermittent non-predictable lower level behaviors. In response, there is likely to be further extensions of device functionalities to provide additional detection and correction. Like scan, the development and deployment of lower level intrinsic mitigation schemes designed to provide improved system reliability, such as error correction used on RAMs today, and other capabilities of the future, are likely re-usable for test purposes as well (detecting, correcting defects, even systemic ones, on top of the more infrequent or non-predictable intermittent behaviors they would be originally designed to provide protection against). At the very least there will be integration issues with the test process (for example, for redundancy/repair for memories) but more optimistically they could likely be aligned to provide more effective, more efficient, or reduced overhead to future semiconductor test processes.

# 3 TEST AND YIELD LEARNING

In addition to the normal sorting function, test provides the essential feedback loop for understanding the characteristics of manufactured chips. Test must continue development of its capabilities to support failure root cause determination, cost-effective defect isolation, process measurements and design-process sensitivities.

## 3.1 ELECTRICAL TEST-BASED DIAGNOSIS

Test-based-learning is needed for both (1) defects and (2) parametrics and variability. Defect-learning is needed for both random and systematic defects (see "Defects and Failure Mechanisms"). Test-structure-based defect learning methods suffer from both traditional area-related limitations and by the number of physical design configurations they are able to cover. As susceptibility to localized failures becomes a function of complexities that include OPC algorithms, nearby shapes and neighborhood densities, it is increasingly necessary to base learning on product-test diagnosis, which inherently reflects the physical design configurations of importance to the product.

Parametric-related feedback is needed for (1) device and interconnects parameters and (2) design-process interactions. Measurement of device and interconnect parameters have traditionally relied upon test structures, especially scribe line FETs and interconnect resistance and capacitance monitors. Increasing across-chip variation (intra-die variability) enhances the negative impact of scribe-line-to-chip offsets. Moreover, test structures are limited by the number of physical and electrical configurations they are able to cover. As circuit parametrics are increasingly affected by such configurations, it becomes necessary to base learning on product test. Embedded, distributed monitor circuits such as thermal and VDD sensors and process-monitoring ring oscillators are now standard on microprocessor-class ICs and can be used to help diagnose parametric fails and understand variability. Understanding variability includes unraveling the structure of variations into spatial and cross-parameter components (variation in transistor length, Vt, source-drain resistance, etc.) The spatial component includes both die-to-die and within-die components. The resulting understanding can be used to improve yield in at least three ways. One is feedback to the manufacturing process for corrective action. Another is input to design via modeling. A third is input to post-manufacture adaptive operational control. Note that product test is a critical component of this process in that correlation between embedded structures and product power/performance is necessary to prioritize corrective actions. Cross-parameter variations, potentially including a spatial component, are important to analog/RF circuits, as well as digital. Methods for understanding/characterizing the manufacturing process and operating environment that are sensitive enough for analog/RF are needed. Embedded monitors are also valuable in that they can provide parametric data when scribe line data is not available or available on only limited sites, owing to test cost or because of limited data access for fabless chip producers. Note that embedded monitors may increase in importance as modeling capabilities relied upon during design fail to keep pace with technology changes and more parametric problems reach the test floor. Product test is uniquely well-suited to provide feedback on design-process interactions, including those leading to noise-related fails, such as power-grid-droop and crosstalk fails. Adaptive circuit operation that tailors operating conditions (frequency, voltage, possibly body bias) to operational load is synergistic in that it too leads chip manufacturers to embed distributed sensors on chip. At the same time, it makes test's job more difficult because it leads to decreased guard bands during operation. That in turn makes it all the more important for test to detect subtle failure mechanisms and to provide yield-learning feedback based on those detections.

Electrical-test-based diagnosis methods are needed for circuit types including memories, logic and scan chains. Memory techniques, the most mature, are no longer adequate alone. RAMs have been superseded by microprocessors as

technology leaders, they fail to utilize all metal layers and they represent only limited layout configurations. Logic diagnosis has been an active area of recent research and development, but still is challenged in terms of accuracy, resolution and large data requirements. Integration of layout information is commonly employed. Successful techniques for integrating in-line inspection results have been developed, but suffer from in-line data availability and ever-smaller, potentially non-visible, killer defects. Volume diagnostics, which combines results from many failing die, is a promising strategy for addressing remaining challenges. Multi-die statistical analyses, such as those employed in volume diagnostics, hold promise also for distinguishing systematic from random failure mechanisms. Identifying systematic failure mechanisms is important for process improvement, DFM guidance (including hot spot identification) and ensuring test quality. Challenges in making the distinction include setting expectations for random failures against which systematic are compared. Expectations based on design analysis, including traditional critical area analysis, suffer from the need to make failure-likelihood assumptions or depend upon defect models that may not be accurate given complicated design-process interactions. Note that *product*-based diagnostics are increasingly vital, as stated above. The need to base diagnostics on actual product hardware is driven by systematic defect mechanisms now being increasingly complex functions of neighboring shapes, local pattern densities, etc. As a result many failure mechanisms may be visible only on product. In addition, product-based diagnostics automatically put focus on key yield-limiting failure mechanisms. *Volume*-based diagnostics are critical since individual occurrence of any given systematic defect mechanism may be rare. The pooling of data across many failing die are needed to identify true systematic mechanisms.

Optical defect diagnosis/failure analysis methods (such as backside emission, modulation mapping, etc.) must complement electrical diagnosis. Such techniques can improve accuracy and also catch many electrical-diagnosis misses, e.g., on hard-to-detect defects. These methods can cover a larger area of the chip in a shorter time than traditional PFA. Another emerging role of tools is to provide guidance as to whether or not PFA is appropriate for a given failing part or group of parts, especially given the increasing prevalence of non-visible failure mechanisms (non-visible defects or parametric problems) and the related fact that PFA success rate and cost varies by failure mechanism. The yield impact of broken scan chains has made their diagnosis a recent focus, but more widely applicable and accurate methods are needed. Moreover, defects can affect clock and other infrastructural nets, which are difficult to diagnose with today's methods. The tools must handle all realistic physical defects, including resistive bridges, resistive contacts/vias and opens. Methods are also needed to handle diagnostics for fails detected by all major test methodologies, including scan-based and BIST-based test; functional; IDDQ and delay test.

The tools and methodologies should support several levels of software-based diagnosis:

- Production-worthy data collection, trading off resolution against test-cost overhead. Concerns may include test-data compression and BIST approaches. An absolute minimum requirement is failing-block, e.g., core or IP, identification. Average test time overhead should be less than 1%.
- Extensive data gathering on selected engineering or monitor wafers or lots. Granularity must be sufficient to build an accurate defect-type Pareto and support tool-commonality analysis. Throughput time must be short enough to provide timely feedback to the fabrication process on sufficient volume and must support both time-zero and reliability failures. Tools should identify not just failing nets, but failing layers. Such analyses may involve integrating layout information and/or in-line test results into fault localization. Typical test time should be on the order of seconds. During early production, a more significant part of the material may be exposed to such extensive logging.
- Individual die analysis that identifies defects to a single transistor or section of conductor no longer than a few µm and identifies the failing layer. Such analysis may involve special-purpose diagnostic-resolution-enhancing ATPG and fail-data collection and/or analog re-simulation and may be followed-up by failure analysis. Analysis time here may be considerably longer than in the previous two cases.

Data-gathering infrastructure must support increasing diagnostic needs. Specifically, ATE should allow for unlimited collecting of scan data at the model scan vector rates predicted in Table TST4 for DFT testers. DFT techniques such as BIST and test compression must be designed with special consideration to support the necessary data gathering. IDDQ measurement devices need to support the accuracy levels required by diagnostics. Diagnostic data collection allowing localization to a single or few failing net candidates should not add significantly to overall test time. In addition, factory integration issues must be addressed. Data capture and management capabilities must support increasing reliance on statistical analysis and data mining of volume production data for yield-learning. Secure mechanisms for yield-data flow for distributed design, manufacture and test, including fabless/foundry and 3rd party IP, are needed. Standard test data formats, such as STDF-V4-2007 scan fail data, and infrastructure to support their transmittal, are needed to support automation and sharing of data. Distributed design, manufacture and test also creates an emerging role for methodologies and tools to help determine in which area problems lie, e.g., design house, foundry or test house.

## 3.2 FAILURE ANALYSIS

While electrical-test-based learning is increasingly important, Failure Analysis [FA] is still required, especially during technology learning when defect types are not well-known, and for understanding important fails, such as test escapes and system/field fails. CMOS technology migration is severely challenging the traditional FA process. Fault isolation, de-processing, and physical characterization will increasingly be too slow and difficult for routine analysis. To keep pace, improvements and breakthroughs to existing tools/techniques are required. FA capability gaps are detailed in the following prioritized list.

### 3.2.1 CIRCUIT NODE PROBING

In-chamber (SEM) and atomic-force-microscope probing at the nanoscale is now required to characterize minimum sized transistors and SRAM cells at first contact level. This is driven by the need to measure increasingly subtle and non-visible defects and individual transistor parameters. Major limitations include adequate SEM image quality at low accelerating voltages required to avoid device damage, probe drift, probe contact resistance.

### 3.2.2 FAULT ISOLATION PRECISION

Present tools for isolation of electrical faults and defects in die are severely limited by sensitivity and spatial resolution. For example, a majority of techniques (e.g., TIVA, PEM) rely on infrared light that is by definition constrained to about one micron spatial resolution – already more than an order of magnitude worse than minimum feature sizes. SEM or FIB based methods (e.g., voltage contrast, EBIC) have much better resolution but only on exposed conductors and thus are not effective for global isolation on fully processed devices. Near-field solutions are available/emerging (e.g. photonic, magnetic, AFM) but techniques are limited to samples with surface or near-surface features exposed. As such in-situ high-precision, non-invasive probing capability is required to provide electrical stimulus to multiple nano-scale conductors for tracing signals back to root-cause faults.

### 3.2.3 PACKAGING ANALYSIS

Fault isolation and imaging of package level defects is impacted by new and emerging technologies like organic laminates, chip-scale packaging, stacked die, and package on package configurations. Improvements in key methods like magnetic current imaging, X-ray tomography, and CSAM are critical, as are continued emphasis on test and fixtures.

### 3.2.4 CAD/EDA TOOLS

Failure analysis is heavily reliant on scan methodology to lead directly to a fault location or to complement other localization tools. Continued improvements in accuracy, applicability to ‘soft’ defects and AC defects, and seamless integration are required. CAD navigation must be both spatial and time-based, i.e., linked to simulation waveforms.

### 3.2.5 ON-CHIP TIMING MEASUREMENT

Lower supply voltages are causing hot electron based photon emission to exponentially drop in intensity and shift to longer wavelengths. Improvements in time resolved emission (TRE) technology and solid immersion lens (SIL) optics are required to maintain capability. Emerging laser-based probes like laser voltage probing (LVP) are promising. Rapidly increasing active power is also challenging cooling solutions to not compromise optical access to the chip backside or induce vibration. Unless improvements are made in these areas, a radically different technology will be required.

### 3.2.6 SAMPLE PREP AND FIXTURES

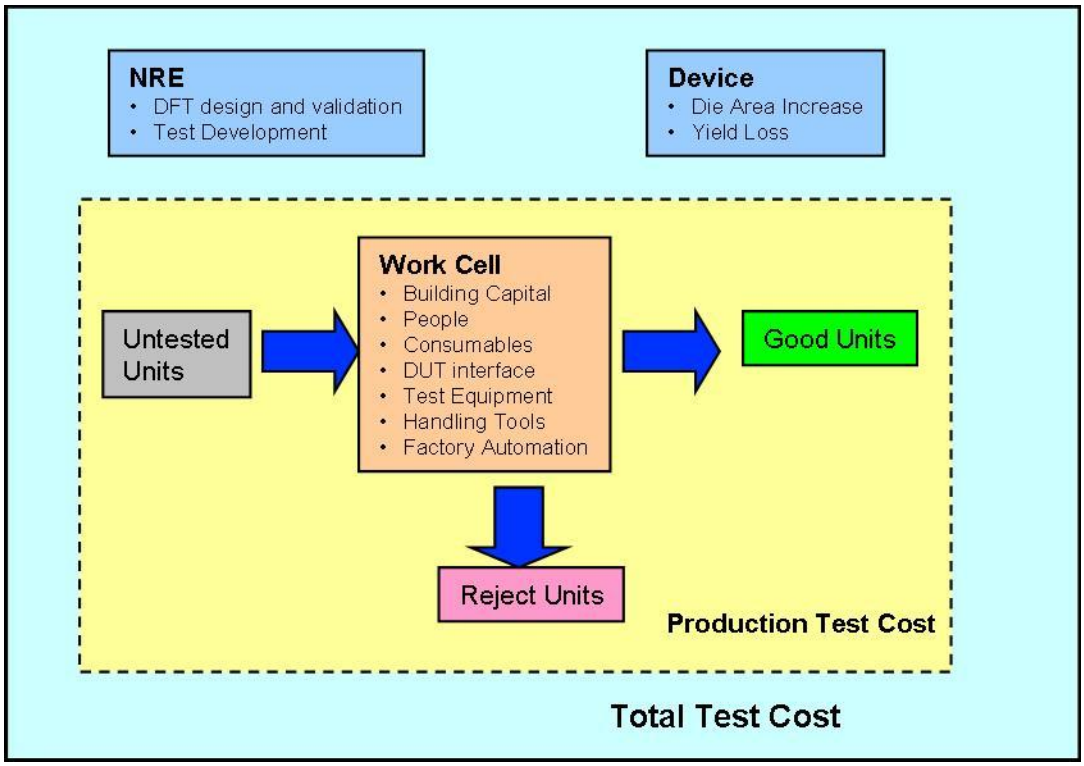
A number of developments in packaging have made backside access for timing and fault isolation difficult to impossible. For example, control of stress-relief induced warping and cracking is required to enable package and die preparation and fixtures for tools like TRE and OBIRCH. Stacked die packages and 3D chips will further challenge sample preparation methods.

The development of new capabilities for failure analysis has become increasingly expensive and high-risk, especially as throughput decreases and Failure Analysis value-proposition goes down. As a result, tool installation is more centralized, reducing total market potential and attractiveness, especially for smaller suppliers and start-ups. The existing model of incidental (or accidental) tool development and R&D investment driven by one or two companies may need to be replaced by a shared consortium whereby cost and risk are spread across a number of stakeholders.

# 4 TEST COST FOCUS TOPIC

Significant progress continues in the reduction of manufacturing test cost, however much work remains ahead. Semiconductor test technology continues the trend towards higher channel integration and higher degrees of multi-site testing, enabled by advanced probe card technologies, new handling technologies, and design for testability techniques. Evidence of this trend is clear in the growing number of low performance ATE systems and (BOST) load-board solutions in standalone memory test and logic test. Even though a lot of effort has gone into reducing the cost of test, 40% of the respondents to an extensive ITRS conducted survey consider the cost of test as one of their biggest concern (compared to 30% last ITRS roadmap) and 85% expect cost of test to become their biggest concern going forward. Significant work remains to translate similar improvements to the broad market of analog and RF. Current DFT techniques are point solutions in certain areas but not generic solutions. The continuing focus on cost of test will result in a better understanding of cost trade-offs between test methodologies, ATE architectures, and distributed test across multiple insertions among other considerations, resulting in overall test cost reduction.

*Figure TST2 – Test Cost Components*




The cost of semiconductor test to the organization has many drivers as shown in Figure TST2. To better understand various aspects and trends related to cost of test, ITRS conducted an extensive survey in 2009 and repeated this survey in 2011. Most participants consider their quality best in class (57%). Moreover, more than a third of the participants feel that test is important to the competitiveness of their products (i.e., higher quality products or lower cost products). The survey demonstrates that the main value of test is to obtain quality and support the yield and yield ramp-up. Only 5% of the participants don't emphasize the potential value of test but declare it a necessary evil (compared to 6% in 2009). Few market participants deploy very elaborate cost models beyond the major metrics cost per unit, cost as percentage of total product cost, and the cost per second. The metrics such as total capital expenditures or the test cost per transistor or bit are not used to evaluate the cost effectiveness of test. The importance of the specific drivers varies substantially from device to device. Test development costs are more important for products with a lower volume. The cost of DFT area depends on whether the product is pad limited or core limited. The acceptable cost of test is very market specific and must be determined by balancing the value of test with its cost.

The participants of the ITRS survey identify the ATE capital expenditures and interface expenditures as the two major cost drivers, followed by test program development and silicon debug. Moreover, going forward, the group identifies new

defects and reliability problems, device performance metrics, Known-Good-Die requirements, and the test requirements due to 3D stacking as the major cost drivers

## 4.1 CURRENT TOP COST DRIVERS

- ATE Capital & Interface Expenditures
- Increased concern over ATE utilization (especially for low volume devices)
- Cost of Test Program Development
- Test Time and Test Coverage for KGD (and 3DTSV)

## 4.2 FUTURE COST DRIVERS

- New Defects and Reliability Problems
- Test Requirements of Packaging
- Interfacing cost
- Data volume, diagnostics, yield learning, and traceability.

A new item that made the list of current test cost drivers is downwards pressure on utilization (especially for low volume devices). In addition, test time and test coverage specifically for KGD in 3DTSV applications become one of the current cost drivers (this item was previously in the list of future cost drivers). This is a clear indication that 3D TSV is taking off and having it impact on the cost of test equations. In terms of future cost drivers, data volume, diagnostics, yield learning and specifically traceability are becoming more significant cost drivers (compared to 2009). ATE capital cost has traditionally been measured using a simple cost-per-digital pin approach. Although this is a convenient metric, it is misleading because it ignores base system costs associated with equipment infrastructure and central instruments as well as the scaling that occurs with reducing pin-counts and number of sites. Moreover, it is not aligned with the current trend in ATE platforms, where the same base infrastructure can be used for very diverse sets of test channels. The following equation expresses test cell capital cost in terms of the relevant cost drivers of future test technology:

$$C_{CELL} = C_{BASE} + C_{INTERFACE} + C_{POWER-SUPPLIES} + C_{TEST-CHANNELS} + C_{OTHER}$$

In this equation, $C_{BASE}$ equals the base cost of a test system with zero pins/channels (e.g., it includes the cost of the mechanical infrastructure, back-plane, tester operating system software, and central instruments). $C_{INTERFACE}$ includes all costs required for interfacing with the device, e.g., the cost of interface electronics, sockets, and probe-cards (including spare probe-cards). $C_{POWER}$ equals the cost of the power supplies. $C_{TEST\text{-}CHANNELS}$ is equal to the cost of the instruments (such as digital, analog, RF, memory test instruments). $C_{OTHER}$ includes the remaining costs (e.g., floor-space). Practical considerations may limit the overall performance breadth that can be cost-effectively achieved by a given $C_{BASE}$ infrastructure and should be taken into consideration in the overall test cell planning. For example, a low-end system may have an air-cooled infrastructure, whereas the high-end system will use liquid cooling. Test scenarios are evaluated by dividing the capital cost and performance metrics. For example, an important figure of merit is the *Units per Hour per Cost (UPH/$M)*, that is, the number of shipped devices per hour (*throughput*) over the total cost.

Figure TST3 indicates the rapidly rising interface cost—which must be contained over time to avoid dominating the overall test cell cost. The ITRS cost of test section committee membership pointed out a number of examples. For example for one example SOC DUT high volume manufacturing test floor, the ATE cost is 3 times the interface & prober cost. In some low temperature setups, the ATE cost is 2 times the interface & prober cost. Sometimes the strip handler is as expensive as the ATE. In one example Memory DUT high-volume manufacturing test floor, the probecard cost is 60% of the total cost in wafer test.

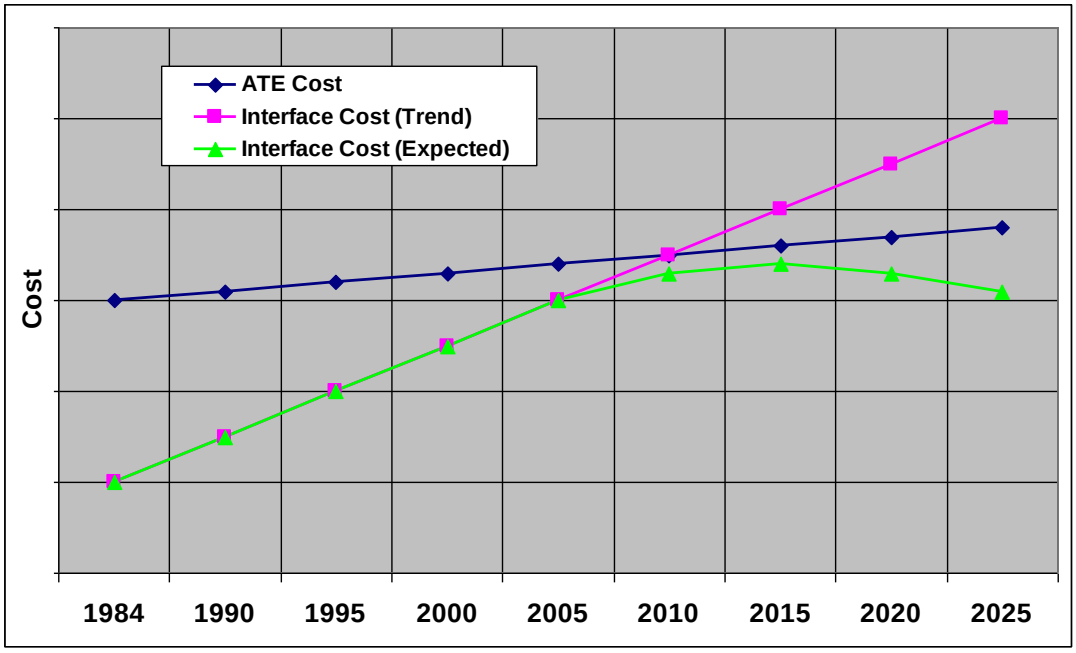



*Figure TST3 - Test Cell Cost / Unit versus Interface Cost Trend*

Optimizing cost of test is limited because of the supply chain fragmentation. For example, probe cards and handler interfaces may need to be prepared for both the design house and subcontract manufacturer(s). The survey pinpointed the top test cost reduction techniques that are currently used and those that may be deployed in the future. The results were:

## 4.3 CURRENTLY DEPLOYED COST REDUCTION TECHNIQUES

- Multi-site & reduced pin-count
- Structural Test & Scan
- Compression/BIST/DFT and BOST
- Yield Learning & Adaptive Test
- Concurrent Test
- Wafer-level at-speed testing

## 4.4 COST REDUCTION TECHNIQUES THAT MAY BE DEPLOYED IN THE FUTURE

- Advanced embedded instruments
- New contacting technologies
- In-system level testing to detect latent defects and potentially repair
- Built in fault-tolerance
- Testing multiple wafers with the same ATE
- Data-processing on centralized servers

Between 2009 and 2011, wafer-level at-speed testing became pervasive as one of the techniques to obtain quality targets within cost pressures. In 2009 the approach was declared as one of the future techniques. This year it's declared as one of the current approaches. Similarly, adaptive test moved from being one of the techniques of the future to being deployed today. Finally, the committee sees a trend of more BOST to extend the life time of existing techniques. For example, techniques include a simple oscillator to generate high speed clocks that the tester can't handle or use a golden device (known good die) to do comparison on the load board. Note that these techniques are mainly used for extending the lifetime of existing testers and not to allow buying cheaper testers in new setups.

There are three new techniques being explored as one of the techniques to control test cost in the future. In-system level testing to detect latent defects and potentially repair the defects is becoming popular in terms of one of the potential future

techniques to control test cost. Moreover, setups that are able to test multiple wafers are being investigated today (especially in applications of NAND flash or other low pin count devices (but not in segments such as SOC or high-end microprocessors). Finally, there is increased interest in centralizing server resources to control overall test floor cost (especially for image sensors and redundancy analysis)

## 4.5 BASE COST TREND

The total base cost is expected to decrease slightly over time. Platform strategies will extend the lifetime of the base infrastructure. Moreover, cost may move from the base infrastructure to the instruments. Multi-site test increases throughput and distributes the base cost across multiple dies, thereby reducing the base cost per site (and making the base cost less of a concern). For successful cost scaling using multi-site test it is important that the ATE infrastructure allows dedicated resources because shared resources may limit throughput. The trend of massive parallel test in memory will continue. Moreover, new probe card technologies and handler technologies, will enable massive parallel test in other segments (for both wafer and package test).

## 4.6 CHANNEL COST TREND

Continuing reduction in channel cost is essential for successful cost scaling using multi-site test: A dominating channel cost per site reduces the advantage of distributing the base-cost among many sites, whereas sharing expensive channels across multiple sites limits throughput. The channel cost is expected to decrease through continued integration within the tester electronics, and also by increased DFT adoption that reduces the ATE pin's performance requirements. Additionally, reduced pin-count test strategies utilizing small test ports can reduce the channel cost per site.

The relatively high cost of analog and RF test instruments, and the long test times associated with testing of these circuits, remain key challenges. DFT methodologies for analog and mixed-signal test are required.

The cost of testing high-speed I/O is becoming significant. In telecom applications, SONET data rates will increase from 2.5 Gb/s to the range of 10–40 Gb/s. Moreover, techniques likes simultaneous bi-directional signaling may become significant, adding to the test cost. High-speed I/O DFT techniques and new test methodologies are becoming a more important part of the solution to control the cost of test.

## 4.7 POWER COST TREND

With increasing degrees of multi-site, the cost of power supplies will increase. Especially if reduced pin-count techniques are deployed, the power cost per site may dominate the channel cost per site. The cost increase in power supplies may be contained by innovations in power supply and power delivery technology. Note that some DFT techniques increase demand on power supplies to achieve shorter test times.

## 4.8 INTERFACE COST TREND

Controlling the interface cost is essential for successful cost scaling using multi-site test: A dominating interface cost that increases exponential with the number of sites may defeat the purpose of increasing the number of sites. The interface cost becomes very challenging with high bandwidth (2 Gbit/s) and/or high multi-sites (128 sites). There is a need to develop consistent cost models that cover the wide range of probe card technologies in the market place. Long probe card lead times cause significant cost problems, especially for the advanced technologies. Probe card lead times will be reduced by a factor of 2 within this roadmap's horizon. For certain products, it may become economical to skip wafer test or only do a simple low performance test. Looking forward, high-speed I/O DFT techniques will become a more pervasive part of the solution to control interface cost.

## 4.9 MULTI-SITE TREND

As discussed in the previous sections, the most important way to reduce cost of test is increasing the number of sites. The effectiveness of increasing the number of sites is limited by (1) a high interface cost, (2) a high channel and/or power cost, and (3) a low multi-site efficiency M:

$$M = 1 - \frac{(T_N - T_1)}{(N-1)T_1}$$

Where $N$ is the number of devices tested in parallel (N>1), $T_1$ is the test-time for testing one device, and $T_N$ is the test time for testing $N$ devices in parallel. For example, a device with a test time $T_1$ of 10 seconds tested using $N$=32 sites in $T_N$=16

seconds has a multi-site efficiency of 98.06%. Hence, for each additional device tested in parallel there is an overhead of *(1-M)* = 1.94%.

*Table TST2 - Test Parallelism*

Shared ATE resources that reduce the channel cost may cause a low multi-site efficiency (for example, this is apparent in mixed-signal / RF test). Moreover, as one continues to increase the number of sites, a low multi-site efficiency has a larger impact on the cost of test. For example, 98% efficiency is adequate for testing two and four sites. However, much higher efficiency is needed for testing 32 sites. At 98% efficiency going from testing a single site to testing four sites will increase a 10s test time to 10.8s. However, going from testing a single site to testing 32 sites will increase a 10s test time to 16.4s, that is, significantly reducing the potential advantage of multi-site test Figure TST4. More often than in 2009, companies experience more efficient ways to reduce overall cost of test than going to the next setup with more sites. Especially for high mix, low volume applications, there are many tester utilization challenges. In these setups, frequently, lower degrees of multi-site is acceptable because test time improvement of techniques to improve utilization have a higher impact on the overall cost of test

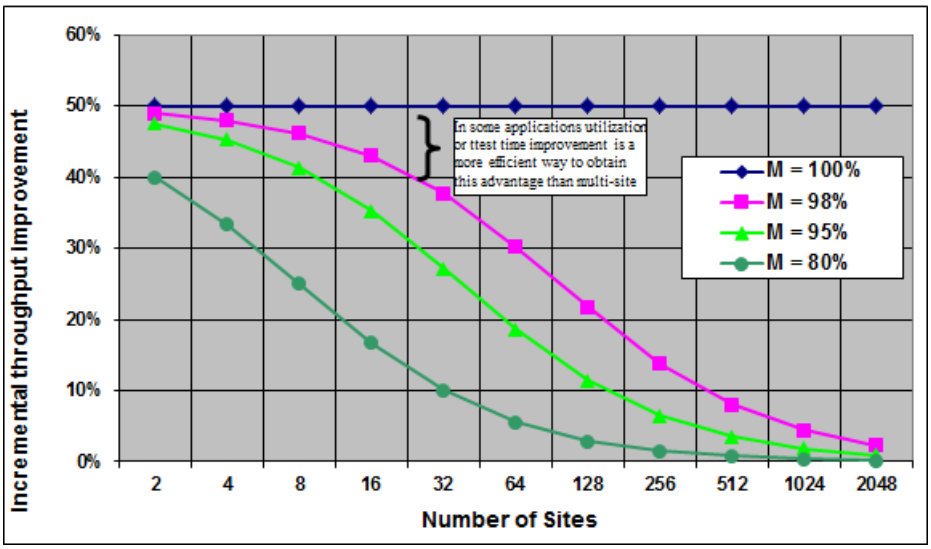



*Figure TST4 – Importance of Multi-Site Efficiency in Massive Parallel Test*

**Error! Reference source not found.** presents the expected trend in the number of sites for an arbitrary device in each product segment. A custom economic model should be deployed to identify the optimized roadmap to scale test cost of a custom device. Note that there are multiple trajectories/approaches that achieve the test cost targets.

## 4.10 OTHER COST TRENDS

The dramatic increase in SiP solutions that integrate memory, logic, and RF have further increased pressure to reduce the cost of test for mixed technology designs as well as improve the wafer test quality of KGD.

Test development time and cost will be reduced further by DFT techniques; test standards (to support test content reuse, test program inter-operability, compression, and manufacturing agility); automatic generation of test patterns (such as structural test approaches), and the programs that use them.

Stand-alone memory DFT, like BIST and built-in self-repair (BISR), will become pervasive as DFT will be essential to control the cost of test. For certain segments, new manufacturing process flows may become economically justified. For example, BIST can be used to test low performance followed by a second insertion on a high performance tester.

Based upon the test survey, the top 5 techniques to reduce cost of test are reduced-pin count and multi-site testing, structural test & scan, compression/BIST/DFT, yield learning and concurrent test. Survey respondents did not rank adaptive test, embedded instruments, nor standardization efforts (e.g., CAST) in the top 5 cost reduction techniques currently deployed. However, adaptive test, new contacting technologies (e.g., MEMS, contactless probing), wafer level at speed testing, build-in fault tolerance, and embedded instruments were identified as techniques that might offer test cost savings in the future.

## 4.11 IMPORTANT AREAS OF CONCERN

- The overall test flow for 3D devices requires making many tradeoffs. There is an increased need for new economic models that allow trading off wafer test with package test and potentially even additional test insertions targeting partial stacks. The expectation is that not all cost metrics and defect coverage can be predicted during test design time, requiring new adaptive test techniques to deal with this optimization.
- The relatively high cost of analog and RF test instruments and the long test times associated with testing these circuits remain key challenges. To enable parallel test, multiple instruments are required with fast execution of DSP test algorithms like fast Fourier transform (FFT) or other correlation tests. A secondary consideration for mixed-signal multi-site test is the load board circuits required for package test, especially for complex packages. DFT techniques for mixed-signal and RF devices remain development needs. Because of the high costs, mixed-signal resources (and post-processing steps) are frequently shared, significantly reducing multi-site efficiency.
- Increases in the number of sites place severe demands on ATE architectures and probe card technologies. Research and development must continue to bring to market cost-effective probe technologies directed at product offerings and multi-site test trends. Lead times for advanced probe cards are an issue. RF radiation may become an issue for massive RF parallel test.
- High-speed serial interfaces are penetrating ASIC and SoC markets. Jitter testing results in high test-times and equipment capital costs. As the number of interfaces increase, the cost problem will increase linearly. New test methods need to be developed to manage the cost scaling.
- New DFT techniques are required to test pins that are not connected during massive parallel test. Moreover, in setups where reduced pin-count techniques are deployed, power supply cost may dominate channel cost. The power supply cost may need to be contained by innovations in power supply and power delivery technology.
- Although a low multi-site efficiency is not a problem for low site counts, for massive parallel test the impact can defeat the purpose of multi-site test. To continue multi-site scaling, ATE architectures may require dedicated channels/instruments per site, because shared channels/instruments may limit multi-site efficiency. To allow dedicated channels/instruments in a cost effective way, the channel cost needs to be reduced for certain segments.

# 5 3-DIMENSIONAL DEVICE TESTING

3D/TSV is the next evolution beyond SiP. In the short term, incremental evolution will present similar challenges to those already presented for SiP, excepting potential changes in die to die interconnects. In the medium to long term, as die stacking becomes more prevalent and more complex/exotic die stacks appear, test challenges will become increasingly more difficult. It is certain that new and additional Design-For-Test features will be needed to mitigate increased tester resource and time requirements as well as increased test complexity due to large numbers of different die in the same package. This section will look at six key test challenges based on the evolution of 3D from incremental additions to SiP through complex die stacks: (1) test flows, cost and resources (2) test access (3) heterogeneous die in a single stack/package (4) debug and diagnosis of failing stacks/die (5) DfX (Design for Test, Yield and Cost) and (6) power. There will be no additional entries to the 2011 test tables based on 3D/TSV. Additions to the 2012 table will include: test access/coverage, cost of test, test data volume, test data storage requirements and test power requirements.

## 5.1 TEST FLOWS, COST, RESOURCES

At a high level, the 3D test flow is comprised of four test process steps: (1) pre-bond test – testing the individual die prior to integration into the stack (2) mid-bond test – testing the partially constructed die stack (3) post-bond test – testing the complete die stack assembly (4) final test – testing the packaged assembly. This test flow adds potentially 2 new test steps: mid-bond test and post-bond test. Addition of these test steps will need to consider cost, process complexity and potential for damage with respect to the At this point, however, it is questionable as to whether mid-bond testing will be considered, based on cost and complexity impact to the die stack assembly process.

Modeling of test flows for stacked die is already being developed in both academia and industry. Much of this modeling is related to optimizing overall test time, however, additional modeling to optimize resource utilization, cost and yield need to be looked at as well. A “cost-weighted yield” model (i.e. which process/test steps have the highest impact on product cost) can incorporate both assembly and test flows to determine the optimal test flow with respect to the stacking process as well as die level test and yield requirements. More work also needs to be done in the “Design for Stack Yield” area. Redundancy (for die, logic, memory and TSVs) is certainly more attainable in a 3D configuration. Such redundancy may be able to increase pre, mid and post-bond yields by “tolerating” a certain level of defects in either the

die or the stack. New defect/fault models will be needed to account for new process steps related to 3D, including wafer thinning and die stack assembly. Defect/fault models for TSVs will also be critical to help define TSV test requirements and process.

Responsibility for some test process steps may also change in a die stack relationship. Reliability testing is one example where responsibility (die provider or stack integrator) is not yet clear. pre-bond, Die-level, active burn-in (versus passive burn-in, i.e. simply baking) will be difficult given very limited access to the die. However, if burn-in is done post bond there is a risk of significant yield loss. At this point there is no clear direction on where and how burn-in is to be done. A significant technology advancement will be required to make burn-in more practical at either the die or stack level, or to eliminate the requirement for burn-in altogether.

Test costs can be influenced by several factors: test resource requirements, overall test time, and cost-weighted yield. mid and post-bond testing can add costs beyond the obvious test time increase. Most notably: failure of a single die in a stack will most likely compromise the entire stack; inter-die testing may add significant time and complexity; stacked die test resource requirements will be driven by the aggregation of all of the individual die requirements, increasing tester resource requirements and cost; huge test data volume due as a result of testing multiple die in parallel; data transfer/storage/security/integrity issues as information sharing between stack integrators and die providers becomes more relevant and important. For mid and post bond testing, the proposed IEEE P1687 (IJTAG) standard may simplify die to die functional test generation on ATE. In addition, boundary scan testing of die-to-die interconnects should be relatively straightforward. However, tools will need to be developed in order to simplify test program generation for partial and/or full die stacks.

## 5.2 TEST ACCESS

This section addresses two significant test challenges for 3D/TSV configurations: pre-bond access - probe access to the die, including die I/O (TSVs) and mid/post-bond access - access to the die-in-stack. The next paragraph addresses test challenges with respect to pre-bond testing. The subsequent two paragraphs address test challenges with respect to mid/post-bond access.

The bottom dies of a stack (those that hold the external I/Os) will typically have probe-able pads for wire-bonding or flip-chip bumps. All other dies (middle and top dies) will typically be connected (power/ground, clocks, control, data) through TSVs. Most TSVs are/will be too small to be probed with conventional probe technology. A "typical" TSV configuration may be 5μm diameter at 10μm minimum pitch. However, in many cases, TSVs are not directly bonded, but equipped with micro-bumps. Micro-bumps may be 25μm diameter at 40μm pitch, which is still small for conventional probe cards. The challenge will be to develop probe technology that can reliably probe on these micro-bumps (not only individually but in "arrays" of probe/contact points). This would require advances in probe contacting, configuration of "contact arrays", metallurgies, tip cleaning recipes, minimizing damage from the probes. Contactless probing also might play a role in this domain. Its main benefit is that it does not inflict probe damage. However, it still does not deliver the small sizes/pitches required, and power/ground still needs to go through traditional needles. As long as the probe-technology is not up-to-par with the micro-bump/TSV sizes/pitches, we will need additional dedicated probe pads to enable sufficient probing. Probing on 40μm pitch micro-bumps seems feasible, but for limited array sizes only. Also, micro-bump technology scales down pretty quickly, so it will be an uphill battle for the probe industry.

Without standardization of a test access protocol for die-in-stack, routing of test signals, vertically through the stack may be difficult due to test signal density, and programming test features may be confusing due to potentially different test protocols for individual dies in the stack. Test data routing through the stack may be defined by one of two scenarios: 1) the development organization drives the design of all die and they are designed with all TSVs physically lined up and the test, verification, debug and other requirements are designed and implemented at the stack level as one design effort; and 2) off the shelf die will require either standardized "test access areas", or interposers to support re-routing of test signals to neighboring die in the stack.

Four basic access functions are required for test access: 1) the ability to provide access to on-die DFx features; 2) the ability to provide a bypass function for skipping-over the die; 3) the ability to provide a turn-around function for terminating the access function at the die; and 4) the ability to provide access to the next die above the current die. JEDEC has already defined some test/access capability in the Wide I/O specification for stackable mobile memories. In addition, the IEEE P1838 Working Group is currently trying to define a standardized access (both physical and electrical) and test protocol to address these requirements. It is contingent on the Working Group to try to release the standard before the stack complexity increases to a point that test access to the die will be a significant challenge. There will obviously be tester implications based both on the number and location of the test signals, and the protocol to "address" individual die in the stack and to access specific test features on the die. Ideally, the access mechanism should assemble

itself as the die are stacked – even if the die come from different fabs and are made in different processes. An access mechanism should include a port on the base die, physical TSV definitions, and a communication protocol or control structure to talk to all of the per die DfX in a stack. The access mechanism must allow test of the die before they are stacked (pre-bond test) and must allow test after the die are in a complete or partial stack (mid/post-bond test).

## 5.3 HETEROGENEOUS DIE

The evolution to complex, heterogeneous die stacks will have significant ramifications on test. Some of these ramifications have already been discussed in the previous sections. However, the implications of testing die-to-die interactions go well beyond what was described earlier. A stacked die can be analogous to a Printed Circuit Assembly (PCA) from a test perspective. Testing of the 3D stack must account for potential die level, pre-bond test escapes based on: untested, functional die to die interactions; power and signal integrity in the stack (compared to die level testing); yet to be discovered defects/faults based on assembly and interconnect processes (wafer thinning may be a good example) which could expose and/or exacerbate die level defects. One area that will be improved over the PCA environment will be die-to-die latencies. Chip-to-chip latencies were a significant contributor to the test escapes described above. Still, test escapes to the stack will be prevalent. Irrespective of the test time/cost perspective, generating a comprehensive, full stack, functional test can be anywhere between impractical to unperceivable. Depending on when testing occurs (mid-bond "stack and test" versus post-bond "assemble and test") it is possible that there may need to be several versions of functional tests to account for the different variations of die on the stack.

Similar to SiP, comprehensive testing can be accomplished by a combination of Built-in Self Test (BIST), judicious use of existing and upcoming test standards, such as: IEEE 1149.1, IEEE 1500, IEEE P1687, IEEE P1838 and limited functional testing. However, this will require a significant amount of coordination between die providers, stack integrators and the architect/designer. It is imperative that Built-in Self Test be used extensively, both to test logic on the die and to test die-to-die interactions.

Traceability at the die level will be significant as the number of heterogeneous die on the stack increases. Data sharing between the die providers and the stack integrator will be critical to maintain quality levels of the stack as well as the individual die. Access to die IDs needs to be standardized in some way – either through a standard access protocol or through a standardized description language. In addition, data sharing and analysis tools will need to evolve in order to accommodate data driven process control that extends beyond the die provider to the stack integrator. It should be noted that test data storage requirements will increase dramatically in such a situation.

Ultimately, the testing of heterogeneous die will be a mixture of "component test", where a die is tested (in the stack) against fault models, parametric requirements and yield criteria to ensure the die is "known good"; and "board test" where a die is tested as a part of an integrated stack, covering interconnect, and interaction properties.

## 5.4 DEBUG-DIAGNOSIS

Similar to a PCA test environment, the major challenges with debug and diagnosis will be the ability to correlate failures at the stack level to defects/faults on the die (at the ATE level). Functional tests at the stack level (mid-bond, post-bond and final test) will be very difficult to debug, especially when a die is in the middle of the stack (allowing for little to no "debug" access). The problem is compounded by the implication that it will be virtually impossible to remove a die from the stack without significant damage. Failure analysis (both electrical and physical) of "systemic" defects will be costly, time consuming and ineffective unless adequate test/debug/FA resources are available at stack level test. Diagnosis may also be impaired by "environment factors" (thermal and power integrity) and an inability to identify potential TSV defects pre and post assembly. Note it is imperative that stack level testing be able to discriminate die level defects/failures from failures due to the assembly process.

Significant integration of built-in test and debug features will be required at both the die level and the stack level. These features could include: built in logic analyzers/state capture, oscilloscopes, temperature and power monitors, droop detectors. Data from these debug features should be logged by the stack integrator and provided to the die provider along with the failing, stack level test. Moreover, significant advancements are also needed in Built-in Test and Debug technologies themselves. Areas such as Analog BIST and Functional BIST, as well as current logic and memory BIST require significant breakthroughs in order to facilitate test and debug capabilities in a limited access environment. Use of Built-in Test and Debug features may allow the die provider to recreate both the test environment and the failure at the stack level. Socketing/fixturing technologies will need to be significantly enhanced to allow for partial/full stack configuration on ATE. "Test interposers" may also be required to help identify/debug systemic, assembly-induced defects. Data-driven debug may be another alternative over time.

### 5.4.1 DFX

Design-for-Test/Debug/Yield (DfX) has been a prominent part of each of the previous sections. These embedded resources may enhance controllability, observability and defect/fault tolerance. DfX at the die level will include: a standardized access protocol to the die; built-in test features to enable die level (ATE) test capability for die in the stack (comprehensive logic BIST or compressed and stored ATPG vectors plus BIST for all memories) in addition to interconnect test capability for all I/Os (boundary-scan and or at speed loopback testing); built-in debug and monitoring features to isolate defects in the stack (including potential capability to measure/monitor TSV continuity and performance); some level of fault tolerance/correction to enable higher yields in the stack (this may include the ability to "partition" given die in the stack). Built-in test and debug features mentioned above will become more prevalent for die in a stack. Partitioning logic or even the die itself can facilitate parallel testing at the die level and may be used to "decommission" logic on the die or a die on the stack (given some level of fault tolerance designed into the die).

DfX at the stack level will evolve over time as knowledge is acquired about defect opportunities during the die stacking process (i.e. back-grinding, wafer-thinning, and laser-drilling) and fault models are developed to accommodate those defects. Stack level integrators may be able to utilize the interposer for stack level DfX features. Interposer based DfX features may assist with die to die testing and parallel testing of multiple die. Built-in interposer based test features may also help to reduce the requirement for tester resources driven by heterogeneous die on the stack.

During the test/debug/yield-analysis process it is imperative that tests may be conducted on one die alone, with multiple die simultaneously, and even with multiple die interacting with each other. Therefore, the access mechanism must provide two connectivity schemes on a per-die basis. The access mechanism connectivity schemes are 1) probe pads for bare-die test; and 2) interconnected and scalable TSV connections after stacking. These schemes must accommodate the four access requirements mentioned in the Test Access section.

## 5.5 POWER

Given that test power requirements can be greater than operational power requirements, power could be one of the most significant test challenges. Power issues may occur at the power domain level, the die level or the stack level. Die and stack level power distribution requirement must be aware of test power consumption. Judicious power monitoring and droop detection throughout the stack will be imperative to guarantee test integrity. Selective logic/die power shutdown may be employed to reduce localized or overall power. Test scheduling may help this. Test scheduling could be handled by the test program or potentially by an adaptive test controller either built into the stack or on to an interposer. Power-aware test can be employed at the die level (in the stack). Power-aware testing at the stack level is something that will need to be considered and refined as test power requirements for the stack grow in the future.

High power levels may also contribute to thermal issues. Thermal issues could impact fixture design and performance. In addition, thermal variations may impact not only performance but also the integrity of the die stack itself. Given this, there will be a need to develop a "thermally induced, inter-die fault model" to describe the impact of thermal variations across the die and the stack. Validation of individual die and die in the stack should also consider guard-banding for potential thermal variation.

## 5.6 3-D TEST CONCLUSION

Significant evolution is required in test equipment, tools, EDA and methodology to address the challenges posed by 3D die integration. As is typical with test, the pace and scope of technology development in 3D die integration will drive test requirements and technology development. Challenges imposed by the manufacturing, assembly and packaging processes will also drive test technology and processes.

# 6 ADAPTIVE TEST

"Adaptive Test" is being used increasingly because of its ability to continually optimize IC testing (within constraints of pre-defined rules) – beyond what is possible using the static optimization which has historically been applied. The benefits include lower testing cost, higher yield, better quality & reliability and improved data collection for yield learning. In this section, we'll provide a short introduction to Adaptive Test – more information is available at http://icdt.ece.pdx.edu/~icdt/cgi-bin/adaptive.cgi/AdaptiveTest

## 6.1 ADAPTIVE TEST DEFINITION

Adaptive Test is a broad term used to describe methods that change test conditions, test flow, test content and test limits (potentially at the die/unit or sub-die level) based on manufacturing data and statistical data analysis. This includes feed-

forward data from inline and early test steps to later test steps and feedback of data from post-test statistical analysis that is used to optimize testing of future products. Adaptive Test also includes real-time data analysis that can perform Statistical Process Control (SPC) and adjust test limits and content during product testing on-the-fly. (e.g., Parts Average Testing algorithms) Although some simple applications have been applied for some time, Adaptive Test will increasingly be applied and will require updated software algorithms and complex statistical analysis methods and database infrastructure.

## 6.2 EXAMPLE APPLICATIONS

Below is a short list of Adaptive Test example applications. These applications are shifting from manual and static methods to automatic adaptation at any level (e.g., lot, wafer, die) during test execution without requiring human intervention.

- **Real Time monitoring of test results that can dynamically change test flows** -- adding or removing tests and selectively performing more yield learning characterization and diagnostic data collection. This application would also perform real-time Statistical Process Control.
- **Post-Test statistical analysis of test results** (e.g., for entire wafer or lot) to better identify outliers or maverick events that could cause DPM or reliability failures. This post-test analysis can also be used to route target dies through unique test flows.
- **Feed-forward of test results from one step to another** to optimize testing or to enable more focused screening. (e.g., inline testing to wafer test, wafer test to package-level testing, burn-in results to package-level testing, package-level testing to card/system testing)
- **Off-line data analysis used to drive test changes for future devices.** (performed offline – but fully automated) For example, off-line analysis would be used to optimize test flows, test content and measurement routines. This analysis could take input from many sources including historic data collected, test capacity, required turn-around times, DPM requirements, expected yields and parametric data.
- **Production Test monitoring & alerts** including real-time statistical analysis of all available data to enable more complex production test operations than has historically been performed. (e.g., real-time parametric data analysis to identify subtle parametric shifts that may be due to marginal wafer probe contacting)
- **Card/System level configuration and test based on feed-forward of component or card-level test.** For example, component test results (such as parametric data, yield or partial good data) would be used to customize the card and/or system test flow. Another example is customizing the voltage of board test based on wafer/component-level test measurement data (process corner). Also, card and/or system test steps could be modified for boards which have gone through multiple rework steps.

## 6.3 DIRECTIONS FOR ADAPTIVE TEST IN THE NEXT 5-10 YEARS

A general form of Adaptive Test is needed to gauge its advancement. Fundamentally Adaptive Test is algorithmically making decisions to change the acceptance limit, test flow, or test content. A decision algorithm is based on a model of the behavior of the device under test. This decision involves the steps that are the rows of the table below. These steps are executed, explicitly or implicitly, from bottom to top. Modeling starts with choosing a model form, an equation, from a family of candidate models, then determining the values of the parameters of the model form to create a model of the specific behavior, computing the decision limit for the unit in question from this model, and finally comparing the measured value to the limit to make the decision.

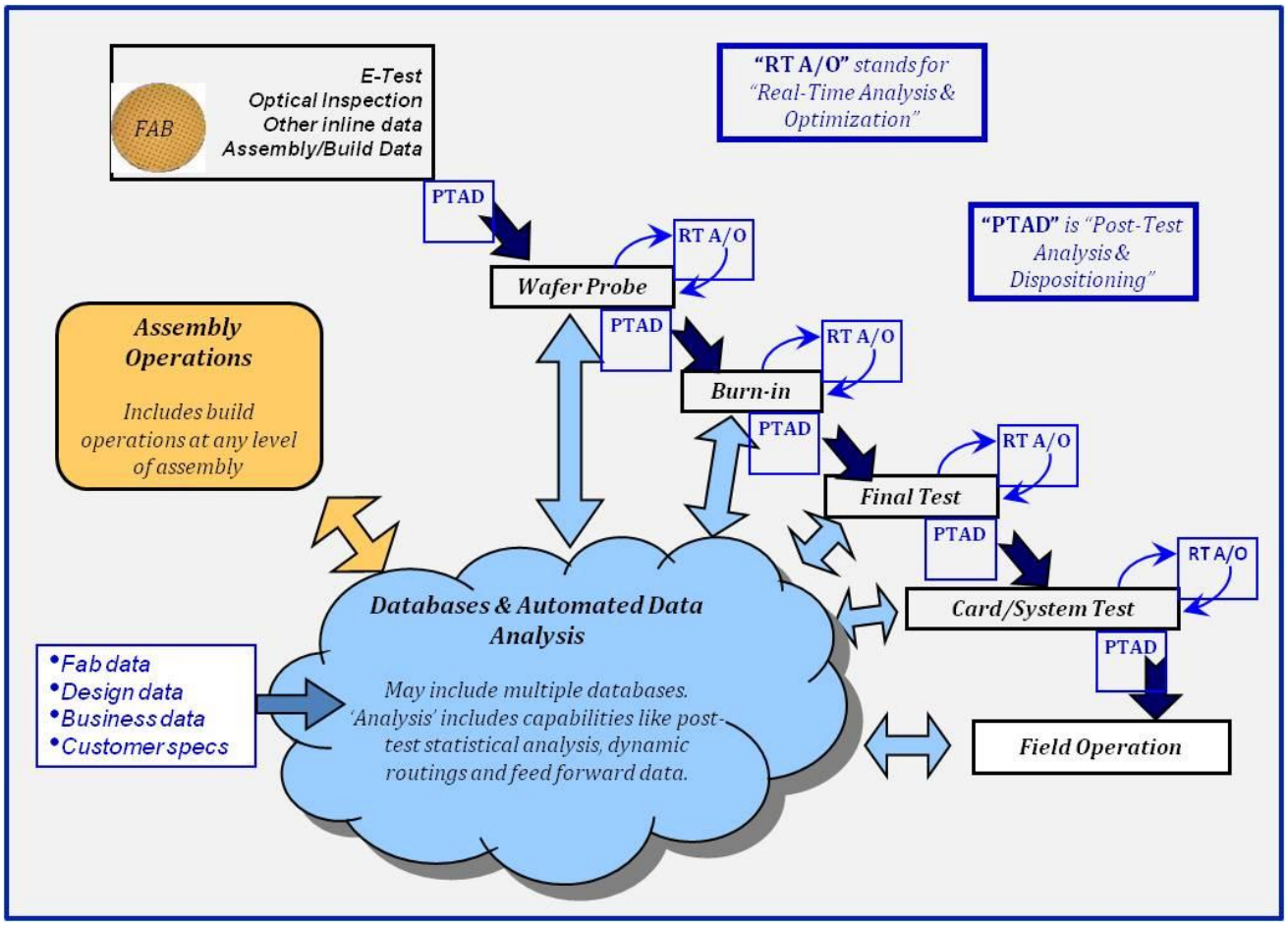



*Figure TST5 - Adaptive Test Architecture / Flow*

The degree of adaptability in making the decision depends on the phase of test development in which these steps are performed, which progresses across the columns of the table from left to right. **Static** means the step is assumed known and is not updated within the adaptive test flow. **Static** elements are either fixed with the algorithms or automatic procedures (e.g. VIL/VIH) or are further removed from the test flow when they are used to generate inputs into the test flow (e.g. ATPG). **Characterization** means measurements are required from devices, process or other data collection steps outside the test flow. **Test Execution** as it implies is the shortest time constant as data is collected per part, per wafer, per lot.

The degree of adaptability trades test effectiveness for the complexity of the test and result processing flow and the **scope** of the data put into it. The data scope (e.g., test, die, wafer, or lot) can be viewed as the minimum sample that must be acquired before a decision can be made on a test or collection of tests.

| Model Elements | Static Limit | Model With Static Parameters | Model Parameters Extracted from Data |
|---|---|---|---|
| Decision | T | T | T |
| Limit Setting | C | T | T |
| Parameter Value | S | C | T |
| Model Form | S | S | C |
| Model Family | S | S | S |

| S=Static | C=Characterization | T=Production Test Execution |
|---|---|---|

Presently these concepts are most highly developed for outlier detection which is adapting the acceptance limit according to a model of the variance of some measurement. For example, IDDQ originally was compared to a static limit but, as its background value increased and became more variable as FETs shrank, the industry moved to adapting the defect threshold to the background value. One such variance reduction model is to predict the maximum current over a set of circuit states from the minimum current over those states.

Here, the model form is usually a line, chosen from the family of polynomials. Its parameters are slope, intercept, and a margin of error. The limit on the maximum current for each part is adapted to the minimum current by plugging the latter into this linear equation. The parameters are found by performing linear regression on a suitable sample. If the fit is not good, another polynomial, i.e. another model form from the chosen family, might be tried. The simplest test program to use this model would use constant parameters determined by linear regression on a characterization sample off-line. This "model with static parameters" can be applied to each chip as it is tested (chip scope) but cannot adapt to a process drift in the parameters. That drift can be accommodated by "model parameters extracted from data" but, not only does this make the test program more complex by adding the linear regression routine, it requires more data be collected before the accept/reject decision can be made, e.g., moving to wafer scope. Similar considerations affect test flow branching decisions as in sampling plans.

## 6.4 TEST RESULTS DRIVING "ADAPTIVE DESIGNS"

More and more designs are being reconfigured during Test. Examples include partial goods, VDD/frequency adjustment and local clock tuning. In most cases this product personalization will be based on either test measurements or data feed-forward from other operations. In some cases, this reconfiguration will be based on "application demand".

## 6.5 ADAPTIVE MANUFACTURING

An emerging direction is using test results to drive other IC production steps such as packaging. For example, the card/board assembly operation may require that specific dies or types of dies to be used based on test results. Given the emergence of 3DICs and power constraints, specific bare dies may need to be selected for assembly based on parametric data collected at test such as IDD or power/performance measurements.

This document doesn't describe all implications of "Adaptive Manufacturing" – but clearly some of the data requirements described in a later section would also apply if data is required for other manufacturing operations such as assembly.

## 6.6 ADAPTIVE TEST BUILDING BLOCKS

Adaptive Test depends on a set of infrastructure building blocks to enable dynamic optimization of test content and/or flow. The key elements are:

- Data - various sources -- may be real-time, historical or feed forward, structured to feed decision algorithms.
- Data Structure & Control Path - deliver targeted data to the analysis engine and decisions to the action point.
- Decision Algorithms - models to automatically adjust test limit, flow, or content based on input data.
- Analysis Engine - Executes model/algorithm on the data structure to return control signals based on specific product parameters. The engine may be separate or embedded in test program or station controller software.
- Station controller/test program hooks to communicate with or run the analysis engine and take action.

Adaptive Test will drive additional infrastructure and business processes that must enable the following:

- Full traceability - details of how Adaptive Test was applied for each product/part tested.
- Standard models & algorithms - Parts Average Testing, Linear models, Machine learning, sampling, …
- Recipe management - model form & parameters, target products/lots.
- Analysis functions - to quantify trade-offs in quality, reliability, test time, yield loss, risk assessment.

## 6.7 IMPLICATIONS FOR ATE AND THE PRODUCTION TEST CELL

- Define/create a (standard) interface at the ATE for a data path transfer of the real-time test inputs/results
- Create a (standard) real-time interface at the ATE for controlling the test flow and hardware configuration for an adapted test flow.
- Define the interface and rules for adjusting test program content real-time (between units or lots).
- Automatic adjustments to hardware configuration for optimized test flow

## 6.8 IMPLICATIONS FOR TEST DATA & ANALYSIS REQUIREMENTS

Adaptive Test involves making decisions throughout the manufacturing process with algorithms of varying complexity using data of varying detail from multiple sources. Therefore it requires structuring more data in new ways for new uses compared to traditional testing. In addition to the usual test result stream, Adaptive Test may also require access to information on test content, product design composition, manufacturing flow and disposition, and business information such as volumes and costs.

This data must be carefully structured to make it available when it is needed to make a decision. The temporal availability of test results suggests a hierarchical framework for organizing test results as well as data from the other sources.

Test data sources will include**:**

- **Device memory** – Data contained in the device under test such as scan registers and DFT memory (volatile) and electronic ID and memory repair fuses (nonvolatile).
- **Test cell memory** – Stream of test results and conditions written to the tester or cell controller's memory as they are measured or set. This data will be required for "real-time" decision-making on the tester.
- **Tester log file** – Stream of test results, test conditions, and on-tester decisions written to a file with some delay and consolidation.
- **Operational database** – Structured database combining the results from all the test insertions of a manufacturing lot. Access may require low latency access for making decisions while the parts are being manufactured (including testing).
- **Historical database** – Database (such as a data warehouse) combining manufacturing lots and products. This database would support traditional yield engineering and periodic test refinement.

Adaptive Test places new requirements, or new emphasis, on this data and structure which can be specified in the following attributes:

- **Identification** – Keys for identifying data records for insertion and retrieval including connecting different sources.
- **Traceability** – "Long-term" recording of the algorithms used, data inputs and the model parameters used in decision-making to enable future analysis (e.g., evaluation of the efficiency and effectiveness of the applied methods)
- **Size** – Precision and level of detail affect the economics of both data storage and transport as well as data latency and persistence.
- **Consolidation** – Data abstraction, e.g. bin for each die to summary bin counts for entire wafer. This will define the size and level of detail retained at each level of hierarchy and therefore the scope of the decision algorithms that can be applied there.
- **Latency** – Time between the request and the delivery of data. This latency will limit where in the manufacturing flow decisions can be made in time to disposition parts for the next stage.
- **Persistence** – The length of time that data is available measured in terms of manufacturing stages or time. This will determines what data is available to make decisions at a manufacturing stage.
- **Integrity** – Coherence and consistency across data sources and time. This will affect when and from where data can be accessed to make a decision.
- **Security** – Who, what, and when of data access.
- **Storage Architectures** – Structure and medium in which data is stored -- such as scan register on chip, log file on tester disk, or relational database on server storage. The storage medium will determine latency and persistence.
- **Transport** – Medium and protocol by which data is transported between source or storage and a decision-making processor over geography, organizations, and systems. This will determine the integrity, security, latency, and persistence.

## 6.9 IMPLICATIONS OF ADAPTIVE TEST FOR CARD/SYSTEM/FIELD

"Adaptive Test" methodologies are expanding beyond component test to include card, system and field applications. Similar to component test, there are various levels of sophistication including feedback, feed forward, static and dynamic techniques. A number of challenges exist and solutions must be provided before system-level adaptive test can be broadly deployed.

The requirements described in the earlier sections must be applicable up through system test and then field applications. (such as data requirements, reconfiguration and adaptive manufacturing) Results from earlier test operations will be fed-forward to enable system-level adaptive testing. These applications will include modification of card/system test flow based on rework history and the selection of enhanced stressed testing based on data from incoming material.

*Table TST3 - Implications of Adaptive Test*

| ***Challenge*** | ***Required Direction*** |
|---|---|
| *IT Infrastructure* | • *Infrastructure to enable the Adaptive Test flow shown in Figure TST5.*<br>• *End-to-end supply chain data integration – including data from Fabs, Test Houses and other Subcons.*<br>• *Develop supply chain data integration and processes which automatically detect supply chain issues and implement corrective actions in near real-time.*<br>• *Integrate multiple databases, flexible logistics system, and full part tracking at each test step, and feed-forward/feed-backward data flows.* |
| *Traceability* | • *Enable full traceability of Adaptive Test parameters (limits, content, flows, and rules) for each die. (accessible anytime in future)* |
| *Real-Time Communications* | • *Develop tester-to/from-data analysis engine communication – without significantly impacting test time.* |
| *Development of Improved Models & Algorithms* | • *Development of methods where the models are not fixed – instead the models are dynamically adjusted based on DUT responses.*<br>• *Develop peripheral coverage metrics and associated quality impact of dropped or modified tests.*<br>• *More encompassing fault coverage metrics are required – particularly for analog circuits.* |

# 7 TEST TECHNOLOGY REQUIREMENTS

## 7.1 INTRODUCTION

Over the past 25 years, semiconductor test technology requirements have been driven primarily by relentlessly increasing performance and transistor counts. A fundamental shift is underway driven by the emergence of new market demands (for example, mobility, security, ease of use, ease of system management, low power, etc.). This in turn is fueling the integration of different semiconductor technologies in more ways and in a greater set of applications than ever before. This in itself is a huge challenge to test as it is ultimately the application requirements or specifications that determine test technology requirements, but it would be impossible to capture a comprehensive set of applications and their associated test requirements trends within this chapter. Therefore, core semiconductor technology building blocks have been identified to provide a framework for describing the test challenges and trends associated with each core technology as well as for describing the test challenges associated with integrating these core technologies together either as a SoC or a SiP.

Each core semiconductor technology has certain applications associated with it and some of these will be used as a basis for extracting long-term trends. In particular, the ITRS publishes key technology attribute trends for CPU, ASIC, DRAM, and Flash memory. These will be referenced where appropriate in the core technology sections. Figure TST6 shows the core semiconductor technologies addressed in this chapter as well as examples of associated applications. The application mapping is intentionally loose as many of the examples listed may contain multiple core technologies. The core technologies are differentiated mainly by their inherent functional differences and thus their different test requirements. Two emerging core technologies that are not included in this revision are MEMS and optical.

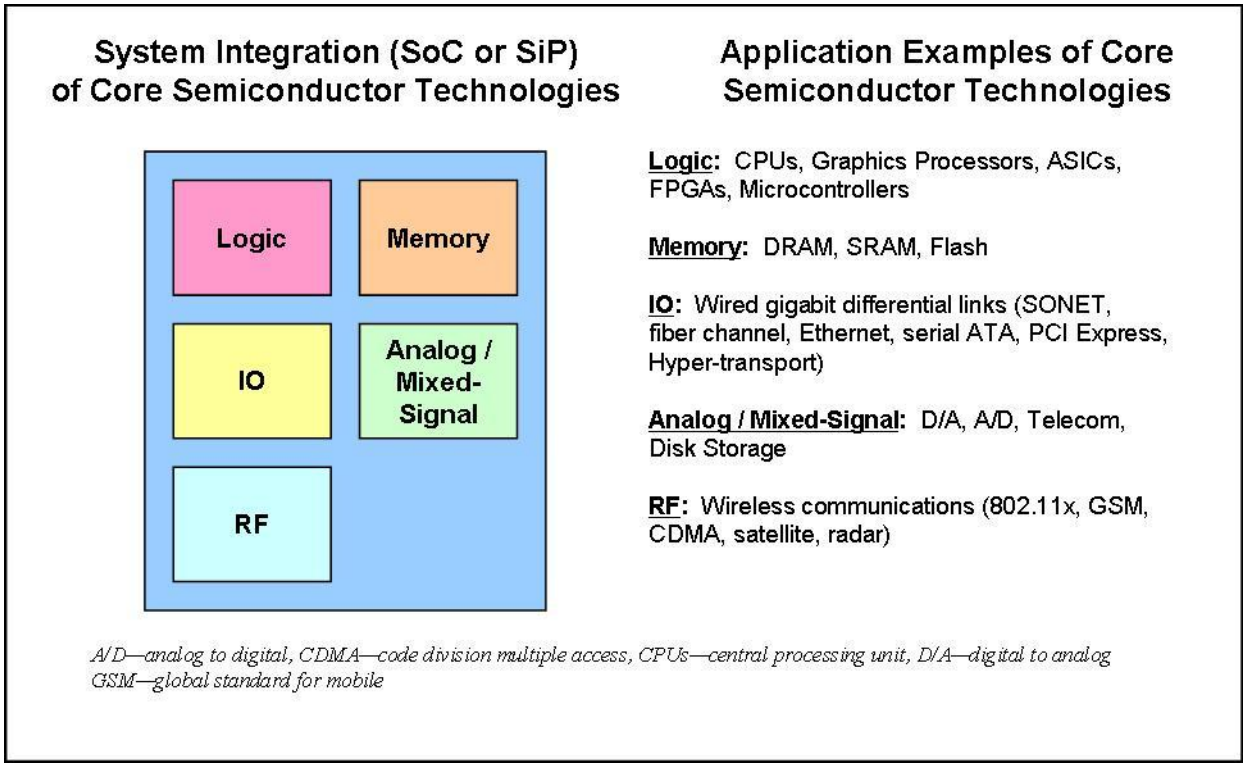



*Figure TST6 - Organization of Cores for System Integration and Applications*

In the recent past, these core semiconductor technologies and applications have demanded distinctly different test solutions, each having specific test equipment and interface tooling markets. Increasing integration is blurring these boundaries. It has also raised the stakes for DFT as successful integration is determined not just by "can it be done?" but also "can the integrated whole be tested economically?" The remainder of the Test Technology Requirements section will address the test challenges associated with increasing integration followed by the test requirements of each constituent core technology.

## 7.2 System Integration—SoC and SiP Test Challenges and Implications

While possibly equivalent in theory, SoC and SiP are very different from each other in terms of which technologies tend to be more easily integrated in package versus on chip and each has very different test implications. Recent advancements in assembly and packaging technologies coupled with the difficulty of optimizing the same wafer fabrication process for different core semiconductor technologies have provided a lot of momentum for SiP, causing some to forecast that SiP will be dominant. It may be that wafer fabrication process improvements and design/DFT needs could push SoC to the fore front or there could be hybrids of the two. One thing is clear: integration is a trend that will continue. The only questions are how fast and in what forms. The next two sections will discuss the test challenges and implications associated with SoC and SiP respectively.

### 7.2.1 System on a Chip

A SoC design consists of multiple IP cores, each of which is an individual design block and its design, its embedded test solution, and its interface to other IP cores are encapsulated in a design database. There are various types of IP cores (logic, memory, analog, high speed IO interfaces, RF, etc.) using different technologies. This assortment requires a diversity of solutions to test dies of the specific technologies corresponding to these embedded cores. Thus SoC test implies a highly structured DFT infrastructure to observe and control individual core test solutions. SoC test must include the appropriate combination of these solutions associated with individual cores, core test access, and full-chip testing that targets the interfaces between the cores and the glue logic (i.e. a non-cored logic). Effective hierarchical or parallel approaches and scan pattern compression techniques will be required to evaluate and adjust the overall quality and cost of the SoC to an acceptable level for customers.

On the other hand, the SoC test technology improvement to handle a progression of design technologies accelerated by the evolving applications is indispensable. The technologies and the requirement to the DFT design (*the design intent*) are addressed in the Design Chapter. The well-organized roadmap and the potential solutions that reflect these *design intents* should be reviewed by the readers. For example, the low power design methodologies, which improve the chip performance, are widely adopted in various current SoCs. However, it is not easy to test the SoC without deeply understanding its functional behaviors and physical structures. As a result, the conventional DFT that focuses only on the static logic structure is not enough anymore, and the evolution to tackle with this issue is strongly required.

The quantitative trends and requirements as a consumer logic chip is shown in Table TST5 compared with a MPU chip. Table TST5 introduces the guideline for DFT design and the requirements for EDA tools.

*Table TST4 - DFT Requirements*

### 7.2.1.1 REQUIREMENTS FOR LOGIC CORES

Sophisticated DFT methods such as random pattern logic BIST or compressed deterministic pattern test are required to reduce large amount of test data for logic cores. The adopted method should consider the pros and cons regarding DFT area investment, design rule restrictions, and associated ATE cost. DFT area mainly consists of the test controllers and test points, which can be kept constant over time by using a hierarchical design approach

Both SoC and MPU devices have an increasing amount of digital logic on the devices. Table TST5 shows a common view of the DFT techniques which are expected to be used moving forward in an effort to cover the most likely faults (as modeled by the EDA systems) while attempting to keep the test costs low by effectively managing the test data volume.

There are four basic approaches in use for scan test generation:

- The EDA tools can consider the circuit in its entirety and generate what's called a "flat" test without leveraging the hierarchal design elements nor including pattern compression techniques.
- The EDA tools can consider the hierarchal design elements to achieve an on-die parallel test setup.
- The EDA tools can imbed compression circuitry inside the scan chain resulting in less data being required from the ATE.
- The EDA can implement a combination of 2 and 3 for a compressed hierarchal approach.

The approach used to apply tests to embedded cores will have a large impact to test time and perhaps also test data volume. One approach that has been used in the past is to test a core in isolation and route its stimulus and expected responses up to the SoC pins so as to avoid having to do ATPG for the core at the SoC level. This saves CPU time for running ATPG, but fails to help reduce test time for the SoC. A more effective approach that can be applied when using test compression is to test multiple cores in parallel and not put them into complete isolation from other cores. Thus, while test compression may be used inside of cores, it may also be used above the cores to allow the scan stimulus to be sent to multiple cores in parallel and to compact the output from several cores before sending it off chip. When this approach is applied to multiple instances of the same core, as long as the cores are configured to be in an in-test mode so they cannot be influenced by the logic feeding their functional inputs, it becomes possible to test all copies of the same core simultaneously and for the cost of testing one copy of that core. This clearly helps improve the overall test compression effectiveness when more copies of cores are used even though the compression within the core does not change. Testing instances of difference cores in parallel is also possible, but it is not possible to generate such combined tests by looking only at the cores in isolation. It requires being able to create merged test patterns that target faults in any core actively participating in the testing. Combining the tests so that they target multiple cores is expected to inflate the test pattern size compared to testing one core type at a time since the care-bits will conflict and require more patterns to get all faults covered. The data in table TST6 assumes that when using hierarchical compression, testing of unique cores in parallel inflates the pattern counts by 10% for each unique core over the first one up to the largest 6 cores. It assumes that smaller cores will have their tests merge into those for a larger core more readily. So, parallel test of identical cores is assumed to be essentially free while parallel test of difference cores is assumed to be a 10% overhead per unique core up to 6.

A tradeoff between test quality and test cost is a great concern. ATPG should support not only stuck-at and transition faults but also small delay and other defect-based faults to achieve a high-level of test quality. Test pattern count will increase over the roadmap as logic transistor count increases. To avoid rising test cost, the test application time per gate should be kept constant. Therefore various approaches, such as test pattern reduction, scan chain length reduction and scalable speed-up of scan shift frequency, should be investigated. However the acceleration of scan shift speed may increase the power consumption during scan shift cycles and so it may make the test power problem more serious. Some DFT and ATPG approaches to reduce the power consumption during scan shift cycles are required. There is also the

important issue of excessive power consumption at scan capture cycle. Several approaches to relax this issue have been proposed, but most of them cause an increase of test pattern counts and consequently make its impact on test application time intolerable. Some low capture power test approaches to minimize the increase of test pattern counts are also required. The impact on test data volume from these low-power scan sequences is shown with a 20% test data volume premium in the low-power rows.

Another problem caused by the increase of test patterns is the volume of test data. Even assuming tester memory size will be doubled every three years, a high test data compression ratio will be required in near future. Therefore test data reduction will remain a serious issue that must be tackled. One possible solution for simultaneous reduction of test application time and test data volume is the repeated use of the same test segmentation IP blocks (TSIBs) in a design. Though this can relax the requirement for keeping the pattern counts per gate constant to a certain extent, it will likely not enough to solve the problem completely. Therefore, other approaches, such as the combination of BIST and compressed deterministic pattern test or an innovative DFT scheme should be also targeted.

The increase of power domains may require some additional test patterns. However, the increase of test patterns will be linear to the number of power domains, so it won't have severe impact on overall test pattern counts. Nevertheless the increase of power domains may deteriorate the effectiveness of simultaneous test of identical TSIBs. The impact of this effect will be investigated for future editions of the roadmap.

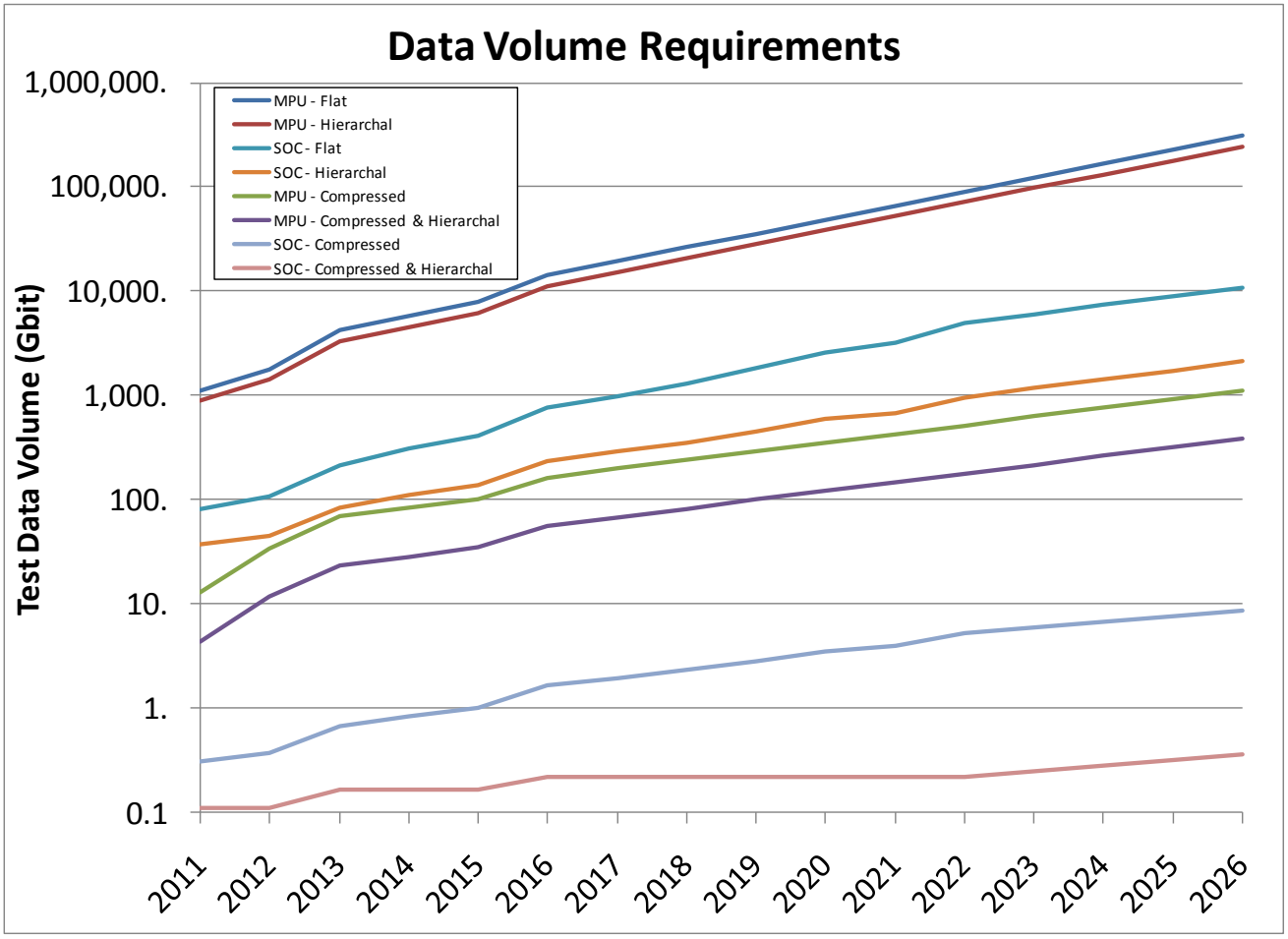



*Figure TST7 - Test Data Volume Requirements as a function of DFT technique utilized*

| Year of Production | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MPU | | | | | | | | | | | | | | | | |
| MPU: With hierarchal compression only | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 |
| MPU: With scan compression only | 87 | 52 | 62 | 69 | 78 | 87 | 98 | 110 | 123 | 138 | 155 | 174 | 196 | 220 | 246 | 277 |
| MPU: With hierarchal amd scan compression | 253 | 151 | 179 | 201 | 226 | 253 | 284 | 319 | 358 | 402 | 452 | 507 | 569 | 639 | 717 | 805 |
| | | | | | | | | | | | | | | | | |
| SOC | | | | | | | | | | | | | | | | |
| SOC: With hierarchal compression only | 2.2 | 2.3 | 2.5 | 2.8 | 3.0 | 3.2 | 3.5 | 3.7 | 4.0 | 4.4 | 4.6 | 5.2 | 5.2 | 5.2 | 5.2 | 5.2 |
| SOC: With scan compression only | 260 | 288 | 323 | 368 | 412 | 467 | 515 | 572 | 644 | 736 | 794 | 939 | 1011 | 1089 | 1173 | 1264 |
| SOC: With hierarchal amd scan compression | 740 | 970 | 1317 | 1865 | 2518 | 3527 | 4560 | 6042 | 8281 | 11844 | 14490 | 22669 | 24415 | 26296 | 28321 | 30503 |

*Figure TST8 – DFT Compression Factors (Flat with No Compression = 1)*

Figure TST7 shows for the impact of various DFT techniques to the problem of the test data increase. The current compression technologies mainly utilize the fact that each test vector has many 'X-values' (don't care bits that don't contribute to the increase of test coverage), and more than 100X compressions have been achieved. However, even a 100x compression won't be enough as shown in Table TST5-SoC, therefore, more sophisticated technologies will be required in future. Figure TST8 shows the level of compression anticipated. The similarity of test vectors applied on scan chains will allow a chance of achieving higher compression ratio. The similarity of test vectors applied in time space may also allow further compression. Thus, utilizing multi-dimensional similarity will be a potential solution.

In order to map this anticipated test data volume to tester and test time requirements one must take into account the number of externally available scan chains and the data rate used to clock the test data into and out of the device. Estimates for these important parameters are shown in the SOC and MPU sections of the tables in TST5. Since these parameters may vary on a part by part basis, the resulting data will need to be adjusted based on the approach taken on one part versus another:

- Designing more scan chains into a device results in more parallel test efficiency and a proportionally faster test time and less memory per pin in the test system.
- Clocking the scan chains at a faster speed also results in a faster test time but doesn't impact the pattern memory requirements of the ATE.

The other question when looking at the ATE memory requirements is which pattern compression technique is chosen for use on a given device. This question in impacted by many parameters including device size, personal preference and time to market constraints. As such, the analysis in TST6 shows the *minimum* patterns per pin necessary to test the most complex devices. Thanks to the usage of more elaborate pattern generation techniques the data suggests that the minimum pattern requirement will only grow by 2x to 3x over the roadmap period.

The test time needed to drive and receive this much test data volume is impacted by the data speed in use. As cost effective higher speed tester capabilities are deployed we feel that these data speeds will be used to speed the tests and reduce the test time per device. The analysis calculates this impact and suggests that test times will be dropping over time due to these faster scan shifting speeds. It should be noted that keeping the test application time per gate constant does not immediately mean a stable test application cost. Therefore some approaches to reduce the ATE cost, such as the increase of parallel sites, the use of low-cost ATE, or a speed up of test, are also required to establish the scalable reduction of test cost per transistor.

Concurrent parallel test in the core hierarchy has the potential of reducing test time. ATPG/DFT level reduction technologies should be developed in future. "Test per clock" means a test methodology that is quite different from the scan test (i.e. non-scan test). The test is done at each clock pulse and the scan shift operation is not needed. There are some researches regarding the methodology, however, more research is required for industrial use.

High-level design languages are being used to improve design efficiency, and it is preferable that DFT is applied at high-level design phase. DFT design rule checking is already available to some extent. Testability analysis, estimation of fault coverage, and DFT synthesis in high-level design that includes non-scan-design approaches are required in the next stage. Yield-loss is a concern. As test patterns excite all possible faults on DUT, it will lead to the excessive switching activity, which does not occur in normal functional operation. This will cause excessive power consumption making the functional operation unstable, and eventually may make the test fail, which will cause over-kill. In addition, signal integrity issues due to resistive drop or crosstalk can also occur which would make the functional operation unstable or marginal and eventually cause failures. Therefore, predictability and control of power consumption and noise during DFT design is required.

The discussion so far in this section has focused on the automatically generated scan based testing requirements. Functional test techniques continue to be broadly deployed in order to enhance the scan based testing techniques in an attempt to confirm the device's suitability for the desired end-use application. Additionally, more and more memory arrays are getting embedded inside of both MPU and SOC devices.

## 7.3 REQUIREMENTS FOR EMBEDDED MEMORY CORES

As process technology advances, and due to some special application needs, the total capacity of memory bits increases and will cause an increase in area investment for BIST, repair and diagnostic circuitry. As the density and operating frequency of memory cores grow, memory DFT technology will require functionality and performance innovations as follows:

- To cover new types of defects that appear in a nanometer process, testing algorithms should evolve from a generic fixed one to either a selective combination of generic algorithms and test conditions, or a dedicated optimal algorithm for a given memory design and defect set. Furthermore, a highly programmable BIST that enables flexible composition of the testing algorithms should be developed.
- Practical embedded repair technologies, such as built-in redundancy allocation (BIRA) and built-in self-repair (BISR) technologies should be developed. BIRA analyzes the BIST results and allocate redundancy elements for yield improvement. BISR performs the actual reconfiguration (hard-repairing) on-chip.
- On-line acquisition of failure information is becoming essential for yield learning. Failure types, such as bit, row, and column failures or combinations of those failure types, need to be distinguished on-chip without dumping a large quantity of test results. A built-in self-diagnostic (BISD) feature could enable this functionality by analyzing memory output and pass the results to ATE to utilize them for the yield learning. The testing algorithm programmability mentioned above has to be more sophisticated to contribute the diagnostics resolution enhancement. It must have a flexible capability to combine algorithm and test data/condition, and a memory diagnostic-only test pattern generation capability which is not used in the volume production testing.
- All the above features need to be implemented in a compact size, and operate at the system frequency.

Based on Table TST4, the embedded memory test, repair and diagnostic logic size will be up to 35 K gates per million bits in 2011. This contains BIST, BIRA, BISR, and BISD logic, but does not include the repair programming devices such as optical or electrical fuses. The ratio of area investment to the number of memory bits should not increase over the next decade. This requirement is not easily achievable. In particular, when the memory redundancy architecture becomes more complex, it will be difficult to implement the repair analysis with a small amount of logic. Therefore, a breakthrough in BIST, repair and diagnostic architecture is required. Dividing BIST, repair and diagnostic logic of memory cores into a high-speed and a low-speed portion might reduce the area investment and turn-around-time for timing closure work. A high-speed portion that consists of counters and data comparators can be embedded in the memory cores which will relax the restrictions for system speed operation in testing mode. A Low-speed portion that consists of the logic for scheduling, pattern programming, etc. can be either designed to operate at low-speed or shared by multiple memory cores, which will reduce area investment and ease logical and physical design work. A lot of small-size memory cores are very often seen in modern SoC; however, they require a larger amount of DFT gates than for a single memory core of the same total bit count. Therefore consolidating memory cores into to a smaller number of memory blocks can reduce memory DFT area investment drastically. Testability-aware high-level synthesis should realize this feature in the memory cell allocation process and consider the parallelism of memory access on system operation.

#### 7.3.1.1 REQUIREMENTS FOR INTEGRATION OF SOC

Reuse of IP core is the key issue for design efficiency. When an IP core is obtained from a third party provider, its predefined test solution must be adopted. Many EDA tools already leverage a standard format for logic cores (for example, IEEE1500[1]); and this format must be prevailed and extended to other core types, such as analog cores. DFT-ATE interface is going to be standardized (for example, IEEE1450[2]), and it should include not only test vectors but also parametric factors. Test quality of each core is now evaluated using various types of fault coverage such as stuck-at fault, transition delay fault, or small delay fault coverage. A unified method to obtain overall test quality that integrates the test coverage of each core should be developed. Conventionally, functional test has been used to compensate structural test's quality. However, automated test for inter-core or core interface should be developed near future. SoC level diagnosis requires a systematic hierarchical diagnosis platform that is available for learning the limiting factors in a design or

[1] *1500-2005 IEEE Standard Testability Method for Embedded Core-based Integrated Circuits.*
[2] *P1450.6, Draft Standard for Standard Test Interface Language (STIL) for Digital Test Vector Data—Core Test Language (CTL).*

process (such as, systemic defects). It should hierarchically locate the defective core, defective part in the core, and the defective X-Y coordinate in the part. A menu of supported defect types must be enhanced to meet with the growing population of physical defects in the latest process technology. Smooth standardized interfaces of design tools with ATE or FA machines are also required. Volume diagnosis is required to collect consistent data across multiple products containing the same design cores, which is stored in a data base and is analyzed statistically using data mining methods. The menu of data items is very crucial for efficient yield learning, but it is a know-how issue now.

#### 7.3.1.2 SYSTEM IN A PACKAGE

In contrast to SoC, SiP offers the option of testing components prior to integration. This is important since integrating one bad component could negate several good components in the SiP, severely limiting SiP yield. In addition, this component testing must typically be done at wafer probe test since integration occurs at assembly and packaging. A key challenge then is identifying good die prior to integration. The term "known good die," or KGD, was coined during the mid-1990s to designate bare die that could be relied upon to exhibit the same quality and reliability as the equivalent single chip packaged device.

In most instances, testing and screening the device in a single chip package format achieves the outgoing quality and reliability figures for IC products shipping today. Wafer probe test is not generally suitable for performance sorting, reliability screening, or effective parallel contacting, so it is generally more efficient to do these tests at the package level using test and burn-in sockets, burn-in chambers, and load boards. Consequently, KGD processing implies that die will be up-binned at probe or with a subsequent insertion of die level tests and screens to meet acceptable quality and reliability targets. The key short term challenges are to determine the quality and reliability targets required in different market segments, develop cost effective tests and reliability screens that can be applied at the wafer or die level, and to develop quality and reliability methods that provide high confidence regarding quality and reliability levels achieved. Longer-term challenges will be to move to a complete self-test strategy with error detection and correction available in the end application.

#### 7.3.1.3 STACKED DIE TESTING AND EQUIPMENT CHALLENGES

Stacked Die (SiP and TSV) products can present many unique challenges to backend manufacturing flows because these products can contain die from more than one supplier. This can create problems in the areas of:

- development of a package test strategy to realize both cost and DPM goals
- production flows to accommodate the necessary reliability screening methods (burn-in, voltage stress, etc) of diverse product/process technologies
- failure analysis methodologies for fault localization in order to resolve quality problems and systematic yield issues

Stacked Die test at the package level closely resembles the test problems of complex SoC products, that is, a variety of IP, each with specialized test requirements, which must be consolidated into a single consistent test flow. In the case of SoC, because everything is on one chip and designed together, various block test strategies can be consolidated via the use of test shell wrappers, test control blocks, etc. using strategies such as defined in the IEEE 1500 specifications. In the case of Stacked Die situation, die suppliers may be reluctant to provide information needed to access special test modes (sometimes considered confidential, especially for commodity memory products) and the individual die may not have the necessary test infrastructure overhead to implement test strategies commonly used for SoC.

Even in the case of SiPs that use only KGD, a certain amount of testing is necessary after final assembly to ensure that the die have been assembled properly. When final assembly may include die thinning and stacking, which can damage/change KGD die, additional testing may be necessary. For the case of fault localization, the ability to narrow the failure to a specific die, and further to a small region of that die, may require full understanding of the detailed test strategies for that die, even if not necessary in normal production..

In the case of reliability screens, some die may require burn-in while others may require only voltage stress. Stress conditions for one die may be inconsistent (or even detrimental) to other die in the same package. Resolution is more difficult since the different die in a SiP product often have totally different processes. One solution is to avoid reliability screens after final packaging but this can increase overall costs (for example, wafer level burn-in is typically more costly than package level burn-in).

When heterogeneous die are assembled into a multi-chip package, several test insertions on different platforms may be required to test the assembled module fully. The multiple test insertions may result in test escapes or yield fallout due to mechanical damage. New testing equipment will be required to accommodate contacting the top side of the package for package stacking. For wafer stacking technologies, better redundancy/repair technologies are needed so that the final

stack can be "fixed" to achieve yield/cost targets. Design and production of electronic systems that can detect failed components and invoke redundant elements while in service is a key challenge for SiP reliability.

#### 7.3.1.4 WAFER TESTING AND EQUIPMENT CHALLENGES/CONCERNS

The probe card technologies in common use today are less than ideal as a "final test" environment. Since much of the performance based speed critical, RF, delay and analog testing is presently performed at package level, a critical challenge for KGD processing is the development of cost-effective, production worthy, reliable and accurate methods of rapidly identifying devices that are defective or will fail early in an application before those devices are transferred to the next level assembly.

Test time for certain technologies, such as display drivers or state of the art DRAM is exceedingly large. Because of the limitations in the wafer probing process, the test throughput is much less than packaged components. The challenges for fully testing DRAM die in a cost effective manner at the wafer level include development of technology that can probe multiple die on a wafer without overlapping previously probed die or stepping off the wafer, and to avoid wasting test time and power on all previously rejected and obviously non-functional die.

#### 7.3.1.5 WAFER TEST FOR RF DEVICES

A key challenge for applying KGD processes to RF die is development of high performance, fine pitch probe cards. Because of the small size of RF die, the pad pitch is very small. As an example, the pad pitch in some products can go below 75 µm, which is the limit of the actual probe technology today.

In order to obtain good signal integrity during RF probing, a configuration of GND-Signal-GND for RF signals is required. A key challenge for KGD processing of RF devices is to ensure that the GND-Signal-GND configuration is designed into the die to maintain the RF path at controlled impedance, given proper probe card design and RF probing techniques.

#### 7.3.1.6 RELIABILITY SCREENING AT THE WAFER OR DIE LEVEL

Voltage and temperature over time are the known stresses for accelerating silicon latent defects to failure. These are more readily applied at package level than at wafer or die level. Applying these stresses prior to packaging the die is a key challenge for KGD.

Development of a cost-effective full-wafer contact technology with the process capability required for manufacturing is a key challenge for the industry. Contact process capability is a function of not only the contactor technology performance but also the burn-in stress requirements for a given product.

#### 7.3.1.7 STATISTICAL PROCESSING OF TEST DATA

Techniques using statistical data analysis to identify subtle and latent defects are gaining favor in the industry, especially for device types with low shipping volumes, part number profusion and short product lifetimes that make burn-in an untenable option and for products where intrinsic process variation makes separating good die from defective die impossible using traditional on-tester limits. The advantages of reliability screening at a test insertion instead of burn-in are savings in time, fixtures, equipment, and handling. The KGD implications are that screens can be performed at the wafer level with standard probes and testers, so every device can be considered fully conditioned in compliance with data sheet specifications and shipped quality and reliability targets for that process regardless of the final package in which the device is to be shipped. Using off-tester statistical methods the test measurements (for example, Idd, Vddmin, Fmax) of each die are recorded instead of being binned. These measurements can be recorded for different test conditions, pre and post stress testing, and at different temperatures. Pass or fail criteria are determined based on statistical analysis of the measurements recorded using off-tester post processing algorithms. Outliers to the statistical distribution are graded based on their statistical likelihood of being system failures or early life failing devices, and the inkless wafer maps are modified accordingly. The challenge for testing using statistical methods is to meet an acceptable trade-off between the potential failing population and the intrinsic yield loss.

#### 7.3.1.8 SUBSEQUENT PROCESSING AFFECTS THE QUALITY OF THE DIE

The processing that occurs during assembly can damage some technologies. Wafer thinning is one example: when DRAM wafers are thinned, a shift in the refresh characteristics has been observed. A die from a wafer that was fully tested at wafer level may fail the exact same test after being thinned and assembled into a SiP or MCP. The thermal processing steps that are present in the assembly process can also lead to a change in the refresh characteristics of individual bits. This phenomenon, known as variable retention time (VRT), is impossible to screen prior to the assembly process.

A key challenge is to re-establish the quality levels achieved by the die supplier. This can be accomplished through additional post assembly testing, invoking redundant elements in the individual failing die within the multi chip package, or using components that are specifically designed for multi chip applications.

# 7.4 LOGIC

The focus of this section is the testing of CMOS digital logic portions of highly complex digital logic devices such as microprocessors and more generally the testing of logic cores that could stand-alone or be integrated into more complex devices. Of primary concern will be the trends of key logic test attributes assuming the continuation of fundamental roadmap trends. The "high volume microprocessor" and "Consumer SoC" devices are chosen as the primary reference because the most trend data is available for them. Specific test requirements for embedded memory (such as cache), I/O, mixed-signal, or RF are addressed in their respective sections and must also be comprehended when considering complex logic devices that contain these technologies.

## 7.4.1 HIGH VOLUME MICROPROCESSOR TRENDS DRIVERS

The trends in the first part of the tables are extracted from other parts of the ITRS, and are reproduced here to form the foundation of the key assumptions used to forecast future logic testing requirements. The first two line items in *Table* TST5 show trends of functions per chip (number of transistors) and chip size at production. Chip size is held relatively constant aside from incremental yearly reductions within a process generation. The next line item reflects a trend toward multiple core designs to address, in part, what has become the primary microprocessor scaling constraint - power consumption. The ITRS currently assumes a doubling of cores with each process generation. The last two line items in this part of the logic tables are the nominal device Vdd range and off-chip data rate trends.

## 7.4.2 SYSTEM TRENDS DRIVERS

System trends drivers are very important to consider when projecting future test requirements. For example, one of the most critical system constraints is power consumption. The proliferation of mobile applications, lagging battery technology improvements, system power dissipation issues, and increased energy costs are all contributing to a practical cap on device power consumption. The era of device power increasing unconstrained with increasing performance is over. This does not necessarily mean that performance will be similarly capped, but this is one of the main challenges to be overcome if Moore's Law is to continue. Innovations in transistors, process technology, design architecture, and system technologies could all have a major impact.

One system technology innovation that could impact test would be the integration of voltage regulation on-chip/package. Increasing chip power and increasing number of cores make this ever more likely for at least two reasons. The first reason is that eventually the package limits power consumption by constraining the number of power/ground pins and the maximum current per pin. These constraints can be greatly eased with on-chip regulation since you can then deliver power to the chip at a significantly higher voltage. The second reason is that multi-core architectures may necessitate more sophisticated independent power delivery to each core in order to optimize fully the power consumption. Eventually it is likely that this will need to be done on-chip. Overall, this trend would simplify the problem of delivering power to devices under test, but it also could create many new test issues since precise voltage control and current measurement have always been an important aspect of testing high power devices.

Another important system trend is the continuous increase of chip-to-chip data bandwidth required over time. This translates into increasing chip I/O data rates and or an increasing numbers of I/O pins. In order to reliably achieve speeds much greater than one giga-transfers per second (GT/s), it is necessary to incorporate high-speed serial signal techniques such as differential signaling or embedding the clock together with the data. For example, this is already occurring with PCI Express and Serial ATA interfaces and will proliferate to all I/O ports (e.g., front side bus) for microprocessors over the ITRS timeframe. Refer to the High Speed Input/Output Interface section for more detailed discussion on the test requirements and challenges for testing these interfaces.

## 7.4.3 DFT TRENDS DRIVERS

In order for test cost not to increase proportionally with chip scale trends, continuing improvements to DFT coverage and effectiveness will be crucial. The general trend toward multiple reusable cores offers many exciting DFT possibilities. Could the cores be tested in parallel? Could the cores test each other? Could one or more cores be redundant so that "bad" cores could be disabled by test or even in end-use? Also, could there be opportunity for general purpose test cores—sort of an on-chip ATE? It is very difficult to determine exactly how this will evolve and how this will impact manufacturing test requirements. However, there are some clear trends:

- Structural, self test and test data compression techniques will continue and will be important for containing test data volume increases as well as constraining the device I/O interface during test.
- DFT will continue to be essential for localizing failures to accelerate yield learning.
- DFT is required to minimize the complexity of testing embedded technologies such as memories and I/O. For example, memory BIST engines are routinely integrated into the device to alleviate the need for external algorithmic pattern generator capability and, similarly, I/O DFT features are increasingly being employed to alleviate the need for sophisticated I/O testing capabilities in high volume manufacturing test.
- DFT will increasingly be required to ensure device deterministic behavior or accommodate device non-deterministic behavior. Power management features, I/O communication protocols, and device self repair mechanisms are a few examples of desirable non-deterministic behaviors.

*Table TST5 - Logic Test Requirements*

*Table TST6- Test Data Volume*

*Table TST7- Data Compression Requirements*

### 7.4.4 HIGH VOLUME MICROPROCESSOR TEST REQUIREMENTS

The second parts of the Logic Test Requirements tables enumerate the test requirements over the ITRS horizon. I/O data rates are bounded on the low end by the need to provide slow speed access for structural or DFT-based testing and on the high end by the native speed of the interface. There will be considerable overlap of I/O types, including on the same device, through at least the near term horizon. Power consumption has been effectively capped at a maximum of 300 W for client microprocessor applications and a maximum of 400 W for server microprocessor applications through the end of the roadmap. Similarly, equipment vector memory requirements are expected to grow modestly over time.

Important Areas of Concern

- **Power Consumption and Thermal Management** - While it is assumed that device power consumption will be significantly constrained through the roadmap horizon, new innovations in power delivery and thermal control at test may be required to stay ahead of system design optimization.
- **High Speed I/O Interface** - Test equipment solutions are needed for device characterization and DFT / test equipment solutions that scale economically are needed for high volume manufacturing test.
- **Multi-core Trends** - Create both test challenges and opportunities. Multi-core designs may exacerbate the previous two areas of concern and add test complexity, but also may offer exciting DFT opportunities such as redundancy that could significantly aid test.

## 7.5 HIGH SPEED INPUT / OUTPUT INTERFACE

High frequency multiple Gbps/GHz I/O technology continues to show significant growth in speed and port count in computing, networking, and as well as consumer applications, beyond its true serial communication origin. It is now a very typical IP block found in microprocessors, peripheral interface chips, and inside peripheral devices. Gbps interface standards such as PCIe, hyper-transport, QPI, GDDR, DisplayPort, DDR, USB, Infiniband, SATA, SAS, Fiber channel, Gigabit Ethernet, XAUI, SONET, OTU, and OIF/CEI have been gaining in application popularity. Because of the diversity of applications, it is difficult to outline the future trends in a simple manner. Devices pushing the technology envelope represent a very small percentage of the IC industry. There are very high volume devices that do NOT necessarily push the speed limits, but they are driven by very a different cost structure than the high-end low-volume devices and therefore a very different tolerance level on test cost. There is another category of devices that require large port counts in a single device, which deviates from the normal "serial" concept of this technology, and demanding different test considerations. Because of the above noted reasons, the technology trends are illustrated as a scattered plot in Figure TST9.

The booming telecomm industry of 1998~2001 drove the standard serial communication technology to 13Gbps, and towards 40Gbps using more exotic processes. Multi-Gbps/GHz interface speed experienced an exponential increase (as shown in **Error! Reference source not found.** plotted in logarithmic scale) in computing and network applications; partially as a result of leveraging early telecomm experience. However, because of the much different cost structure, processing, packaging and test technologies, used in high speed interfaces, the adoption was far from simple. For example, most of the Gbps/GHz interfaces are built in SOC type of devices with mainstream CMOS processes. The large scale integration not only presents a challenge in designing reliable high performance interface IP targeted for a very noisy SoC environment, but also on testing of such kind of high performance interfaces. The other challenge derived from

high integration level is the trend for lower Gbps/mW or pico Joule/bit, which is becoming a major combined performance metric measure in this area. This power reduction requirement challenged the design margins. This impacts the test area by an intricate balance of "guarantee by design" and "production test coverage". As an integrated IP block, Gbps/GHz interface testing is now tied to the whole SoC chip level testing. For most of these applications, it means high frequency/performance testing is tied together with large amount of CMOS digital pins/logics. Test cost tolerance, high frequency instrument availability, and signal integrity restriction on test hardware design are the new areas for development in the last few years.

As in any technology roadmap analysis, sustainability of the exponential growth trend depends on changes in technology requirements and foundation technology development. As forecasted in 2007, the telecomm, datacom, and storage industries reached a temporary Gbps/GHz plateau and it will require a serious undertake in technology to go beyond the somewhat established 13 Gbps range. As of 2009, shrinking process technology is starting to support growth on data rates beyond 13 Gbps; such as 14 Gbps FC and 16 Gbps PCIe. As of 2011, has emerging technology reached beyond 25Gbps. Unlike the long haul telecomm applications, some of the Gbps/GHz interface (especially the chip-to-chip interfaces) have chosen to grow in port count (like a bus) until the higher speed technology becomes more cost efficient for use (for example, CEI/OIF, and 40G/100G Ethernet (IEEE 802.3ba standard)).. Future data rate progression may be slower than that seen in recent history but still requires doubling every few years to absorb the performance scaling with better power efficiency in terms of mW/Gbps. However, this slower but still fast ramp in data rate increase could be disrupted by new technologies. There are two technologies in the horizon that can rekindle the data rate ramp and pose significantly new challenges for testing.

The first long term test challenges is on silicon photonics. Because of the higher cost associated with today's discrete optical components needed (e.g. VCSEL, DFB, and lasers), the optical interface was mainly used for communications with reach distances of > 10 m communication links (e.g., LAN, Metro, and WAN networks). The recent silicon photonics advancements could potentially offer an alternative to the electrical interface for distances down to < 1 m. This is largely driven by the fact that the loss for the commonly used FR4 copper material becomes significant (e.g., ~ 5dB/in at 25 GHz), and the maximum reach distance would be limited to < 10 inch (or ~ 25.4 cm), significantly shorter than backplane distance required (e.g., ~ 40 in or 1 m). One may argue to use other relatively low-loss copper materials such as Megatron6, but its higher cost (several times higher than FR4) diminishes the cost advantages for copper over fiber, even for reach distances in the range of 1m of shorter. Due to the loss, power, and cost characteristics between copper and fiber media, it is speculating that at 50 Gbps or higher, many chip-to-chip, chip-to-module interfaces will use optical signaling and optical fiber as the media. The integration of some optical components on to silicon can potentially lowered the cost barrier for using optical media instead of electrical. The challenge for integrated photonic devices in production ATE environment presents a new challenge in the future. Not only the optical instrument availably on the tester, but also the device interface socket/PCB and package handler will be significantly changed accordingly to the physics of electronics and photonics, to meet performance, power, and functionality requirements. . Test hardware design and signal integrity solutions will add a new dimension in the mixed optical and electrical environment.

The second long term test challenge is on multi-level encoded/modulated I/Os. With the CMOS technology advances to support further growth in data rate, new technology may be developed to maximize efficient usage of the electrical media bandwidth. First round of these technologies was focused on analog and digital equalization to offset media loss with the non-return to zero (NRZ) or pulse-amplitude modulation-2 levels (PAM-2). It is now a common feature to have adaptive equalization (e.g., decision feedback equalization (DFE)) for Gbps/GHz interface with data rate in the 6~25 Gbps range. Multilevel encoding scheme (e.g., PAM-M) was already used to reach Gbps data rate for media with much lower bandwidth for a few Ethernet standards (e.g., 1000Base-T, PAM-5). Because of the complexity in mixed signal design, multilevel/PAM-M encoding/modulation is still just an emerging technology under consideration and not yet adopted by the majority of the Gbps/GHz I/O standards. The test challenge is obviously high for multilevel encoding while most of the current tester instruments are binary in nature. New test solution will need to be developed to align with the Gbps/GHz interface technology if it choose to adopt PAM-M signaling.

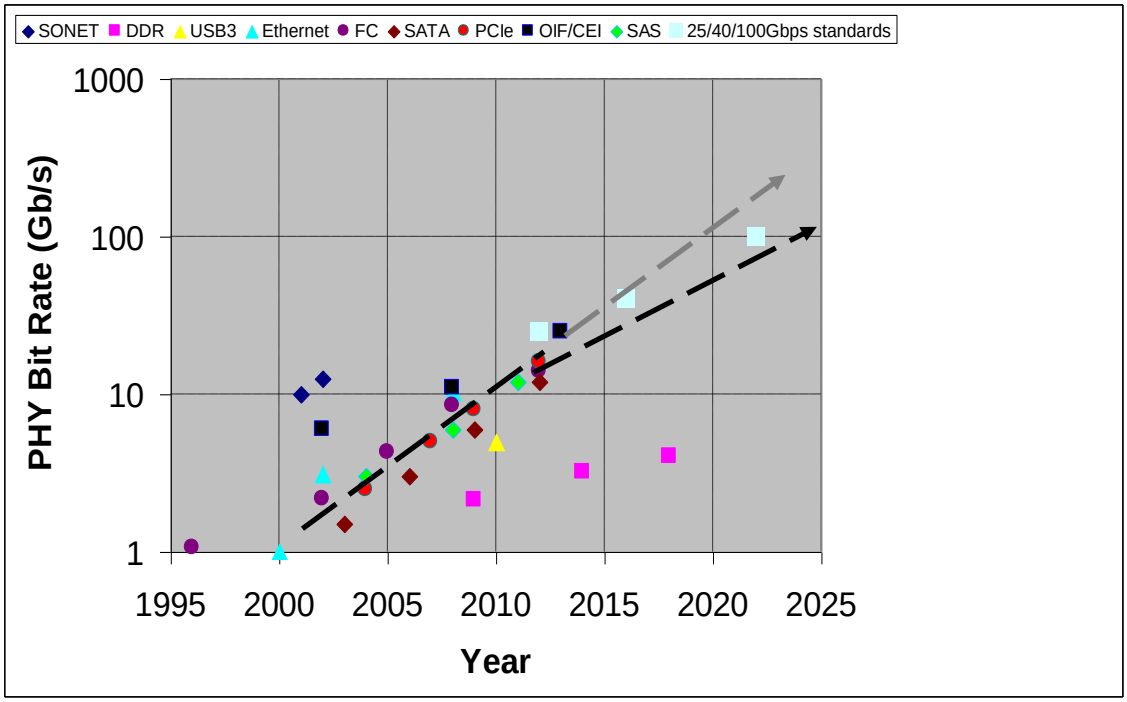



*Figure TST9 - High Speed Interface Trend*

In the past two years, the test and measurement industry has made significant progress in providing high-speed serial link test solutions. Several ATE suppliers provide pin-card-level integrated solutions beyond 6 Gbps, and others already have 12 Gbps solutions. Throughput for some ATE solutions has greatly improved from the previous generation. More flexible clocking schemes eliminated a lot of overhead for frequency and phase synchronization. Jitter measurement and jitter tolerance tests are no longer an after-thought in these instruments, even though true jitter compliance tests are still challenged by the test cost constraints of many applications. ATE instruments are also beginning to support alignment pattern type of synchronization and the typical PRBS based patterns used for BIST. From the basic at-speed testing instrument stand point, it is safe to say that ATE test solutions for 6~8 Gbps have been established. Of course, there will always be applications that push beyond the ATE speed limit (such as for high-end PHY) or port count limit (such as for network switching devices). There is also some special industry standard introduced some features enhancements needs. For example, the DC coupled I/O interfaces such as HDMI/GDDR need 50/40 ohm terminations with pull up voltage from 1.5 to 3.3v at ATE pin electronics. Display devices have to be tested with functional patterns with HDCP, GAMMA, and Floating to Fixed point conversion encryption and decryption. These imposed challenges other than high speed alone. Commercial high speed ATE test instruments are traditional set to target high volume industry standards at the ramping phase to provide necessary test coverage for early adoptions, but lagging behind the leading edge high speed interface serving enterprise needs. At the present time, high-speed production test solutions range from internal digital loopback, DFT assist digital loopback, external wired loopback, external active loopback, high end ATE pin cards, test modules, golden device on test board, to add-on external instrumentation. Each of these approaches has their own pros and cons, which each company must consider. The tradeoff is of course on silicon area used for DFT vs. implementation costs vs. test equipment and interface cost. The tolerance to defect rate of different products is also a major determining factor for the different test choices. A near term challenge to test instruments is the wider adoption of on-chip equalizations. In order to keep the overall system cost down, low cost PCB lamination materials such as FR4 will still be the material of choice for most of the telecomm backplane and computing applications. However, the preference for FR4 forms a bottleneck in spectral bandwidth. Several techniques are being developed to extend the transmission data-rate under this constraint, such as transmit-side equalization (single/multi-tap pre-emphasis/de-emphasis) and receiver-side adaptive equalization (analog equalization and digital feedback equalization). In order to test these types of features, test instruments need to go beyond generating and capturing binary data. For example HDMI and Display PHYs where the transceivers provide pre-emphasis and equalization on the drive and receive respectively for cable media at 3.4 Gbps today. The specification requires that the HDMI receivers implement “reference cable equalizer” to meet the SINK (i.e. receiver) eye mask. Apart from testing BER type testing with cable effect such as ISI, eye mask verification has to be performed post equalizer inside the silicon for compliance test. This requires DFT schemes to support testing device quality after equalization. In relatively lower speed applications, these types of equalization functions were handled with

analog instruments such as AWG and sampling oscilloscope. However for speeds beyond 3~6 Gbps, most of these equalization features are covered in overall closed-loop BER test, with bandwidth limiting media mimicking the final application. This limits the ability to isolate and spec individual components of the on-chip equalization circuit. In practice, it is also challenging to make the production test PCB match the end application. On-chip data eye analysis in the receiver is becoming more standard in order to facilitate this type of testing. More research is needed to properly specify these features, as well as on test, DFT and BIST.

Test equipment manufacturers are refining the instruments for GHz testing. It is becoming a norm to expect at-speed instruments with flexible clocking and device synchronization. Jitter measurement and tolerance testing capability is also becoming more readily available. There are other types of new challenges requiring further timing measurement capabilities such as "inter pair" or "inter lane" skew measurements. Most existing ATE instrument has only white band jitter testing. However, a lot of standards today spec jitter with ideal recovered clock with specific CDR loop bandwidth. Of course, viable solutions need to be cost effective. Currently, solutions to this problem range from early stage mid volume thorough testing for prove in; guarantee by design and thorough characterization; to proprietary outlier detection techniques.

In the long run, the existing DFT features need to be extended beyond the current functional PRBS BIST approach to provide more performance related parametric coverage. With the per transistor silicon cost scaling down, and high speed test instrument and interface hardware cost staying flat, this has encouraged more focus on DFT and BIST. Test hardware design in 25 to 100 Gbps/GHz needs a lot of innovation to keep the off-chip test solution to stay relevant in high volume productions. However, with the design margin coming with a price on power and silicon area, it is envisioned that on-chip instruments and built-in design verification techniques will evolve and coexist with off-chip test equipment. An economically ideal distribution of on-chip and off-chip test coverage can vary from application to application. The overall goal is to minimize manufacturing test cost and efficiently test high port count devices.

## 7.5.1 IMPORTANT AREAS OF CONCERN

- Data Rate Increase - In the high volume computing and networking applications, the Gbps/GHz interface is still keeping its exponential pace but with a reduced slope in logarithmic scale, as shown in **Error! Reference source not found.**. Of course, as discussed above, the 8 Gbps and below post other "feature" challenges beyond just binary speed.
- Port Count Increase - Low voltage CMOS technologies and low output voltage swings enable massive integration into large ASIC and SoC devices. In 2011, up to 200 pairs of 10Gbps backplane style SerDes are found in some applications. However, for a large percentage of applications the port count will be limited to 32 or less. The port count is still increasing and will be limited by chip level power consumption and the actual increasing single lane Gbps data rate.
- Cost Factor - Traditionally, most multi-Gbps transceivers were designed as high-performance, high-priced, and high-margin devices with a low level of integration and relatively low production volume. With the introduction of low cost, low power CMOS macro cells, Gbps/GHz interfaces have become valued additions to many high-volume and low priced commodity devices. In addition to high port count, a cost efficient ATE solution that can concurrently test all serial ports is essential for production. The constant trade-off between performance and integration level results in the separation of SerDes devices into two categories: High-performance-level serial transceivers and high-integration-level gigahertz link macro-cells. The test methods for each type are selected with cost in mind. The economics of high-performance, long haul, communication related products typically allow a more traditional, instrument based test approach or use of a hybrid tester as discussed earlier. Although reliable DFT features or other low cost test techniques are the ultimate solution for large port count SerDes, there is still a strong desire that the tester can provide at-speed characterization or early production. With accelerating technology improvements, the life cycles for most products are becoming shorter, therefore it may become increasingly difficult to verify and optimize DFT circuitry and yield learning.

### 7.5.1.1 JITTER DECOMPOSITION AND MEASUREMENT

The jitter generated by the transmitter is the key parameter to guarantee transmitter quality. Many serial link standards adopt the concept of separating jitter into deterministic jitter (DJ) and random jitter (RJ). The traditional concept of histogram based peak-to-peak jitter has been replaced by the concept of total jitter (TJ), which is associated with a certain BER for the serial link (typically 10-12 or smaller). As the data rate keeps increasing to ~ 10 Gbps and higher, jitter amplification (a new phenomenon during which high-frequency jitter caused by pulse width shrinkage (PWS) or duty-cycle distortion (DCD) gets amplified by the lossy channel (e.g., FR4 based)), becomes severe and needs to be bounded. In addition to the conventional DJ, RJ, and TJ components, Pulse width jitter (PWJ) and DCD have become the new jitter

components to test at 10 Gbps and higher. Furthermore, in some high-volume and high performance I/Os such as PCIe 3.0, a new criterion of whether the jitter is correlated or uncorrelated (to the data pattern) is established to account for the compensation effects of equalization to the correlated jitter. As such DJ and RJ are further separated to correlated DJ (including data–dependent jitter (DDJ)) and uncorrelated DJ and RJ. Correlated DJ (e.g., DDJ), uncorrelated DJ and RJ, along with PWJ and DCD need to be tested to ensure the interoperability and utilization of the jitter margin provided by equalization circuits. Ideal data edges are commonly used as the references to determine the correlated DDJ, and clock edges of the recovered clock are used as references for determining uncorrelated DJ and RJ, PWJ and DCD are determined from data edge to data edge differences. There are also other trends to measure jitter in terms of cycle-to-cycle jitter, peak-to-peak jitter or RMS jitter within a certain numbers of cycles, which could be more meaningful for clocking schemes only allowing very short-term jitter accumulation. Jitter measurement also imposed a very stringent signal integrity requirement. On the other hand, with the source synchronous bus entering the Gbps/GHz range, it introduces another challenge in jitter measurement – the uncorrelated jitter between the clock path and multiple data paths. This is certainly beyond the traditional serial PHY jitter definition.

To accurately measure the DJ and RJ, the jitter floor instrument or tester needs to be significantly below the DJ and RJ of the DUT to avoid the errors in the measured DJ and RJ. The instrument or tester for jitter measurement had also adopted the DJ and RJ merits for its performance. Most of the high-speed standards specify a ~ 0.3 UI TJ at BER =1e-12, with 0.15 UI from DJ and 0.15 UI from RJ. Following this jitter budget allocation distribution, and assuming that a 10% (a relatively loose accuracy target) TJ accuracy for an instrument or tester, then its DJ and RJ accuracy limits can be estimated via dual Dirac model (i.e., DJ distribution is a dual Dirac, and RJ distribution is a Gaussian), and the TJ, DJ, and RJ accuracy limits as a function of data rate are shown in **Error! Reference source not found.**. At 10 Gbps, the TJ, DJ, and RJ accuracy requirements are 3ps, 1.5ps, and 0.11ps respectively according to **Error! Reference source not found.**. These jitter accuracy targets can be met by most leading edge laboratory instruments. If large TJ error is allowed, then DJ and RJ accuracy precision will be relaxed proportionally. It is worth mentioning that this figure intends to give a math and physics based TJ, DJ, and RJ accuracy guideline for a jitter measurement instrument or tester, rather than a specification or hard requirement.

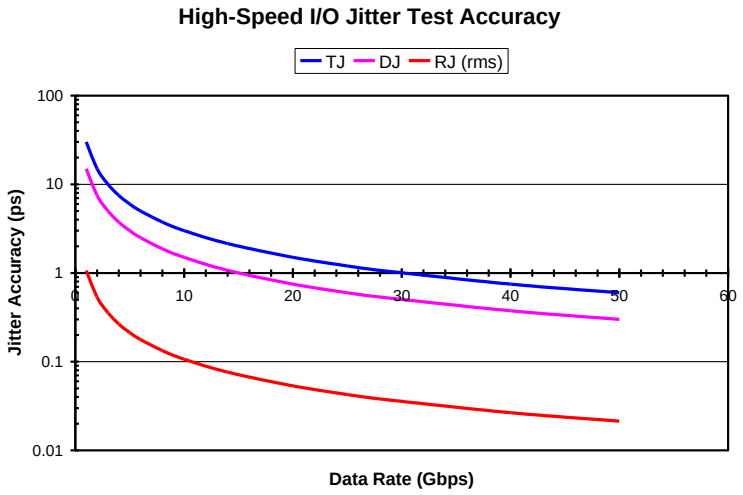



*Figure TST10 - High Speed I/O Jitter Test Accuracy Requirements Scaling with Frequency*

### 7.5.1.2 JITTER TOLERANCE TEST

Jitter Tolerance measures the level of jitter on the input signal that the receiver can tolerate before communication quality, in terms of BER, is degraded. This is a key specification for receiver (Rx) jitter and noise immunity. To conduct a jitter tolerance test, jitter must be deliberately injected into the data stream in a controlled fashion to mimic the worst case jitter and signaling conditions derived from the combinations of transmitter, channel, and reference clock in general. Depend on the architectures of the clocking schemes used; different type of jitter is needed to stress the receiver. Some applications need sinusoidal jitter (SJ) injection, some require a combination of spectrally resolved or frequency-band limited DJ, RJ, DDJ, periodic jitter (PJ) or SJ, and bounded uncorrelated jitter (BUJ) (to mimic the crosstalk induced jitter), and others demand jitter injection in terms of peak-to-peak or RMS jitter within a certain numbers of cycles. An emerging important topic is the receiver equalization test; especially feedback based adaptive equalizations such as DFE. The challenge comes from the fact that testing observability is intrinsically poor for Rx, with the BER as the only and

common observable merit. Without the assistant of on-die or on-chip instrument such as on-die scope, it is very difficult if not impossible to verify DFE tap adaptations to the targeting values. To perform diagnostic or debug test for DFE, on-die instrumentation appears to be a viable solution. Integrated instruments that can inject these kinds of jitter do not exist on previous ATE world, but start to become more available in the lab equipment world, as we have predicted 2 years ago. As of today, some ATE solutions does provide jitter injection capability to stress the receiver and/or clock and data recovery for manufacturing test, which is a big improvement from 2 years ago.

#### 7.5.1.3 TEST FIXTURE BANDWIDTH

The test fixture used to interface the device to the instruments/ATE includes a printed circuit board, cable, connector/pogo pin and etc… With frequencies growing and port counts increasing, the ability for delivering high frequency signals to instruments without much loss and distortion becomes a monumental task in test engineering. The bandwidth requirement is commonly set to be at 3-5th harmonic of the signal under test, depending on its rise/fall time, to have minimum ISI jitter from the test fixture. Bandwidth of 15-25 GHz for a 10 Gbps signal, and 37.5-62.5 GHz for 25 Gbps, are commonly required. In most cases, the bandwidth is NOT the only problem. The phase response has an even bigger impact on jitter measurements. Although there have been significant improvements in socket bandwidth recently, the new high bandwidth socket solutions are challenged for their limited mechanic specifications. Non-reliable insertions and inconsistent contact are still the most common problems causing test result variation. Because of these test fixture limitations, complicated de-embedding effort became a norm for GHz interfaces in performance tests. This therefore inspired another area of active development – tools to simulate and de-embed the signal path including socket, PCB, cable and connector. However, unlike the narrow band RF de-embedding, the wide band de-embedding challenge requires not only insertion loss but also return loss across a wide frequency spectrum. Some SerDes specs also try to alleviate the de-embedding effort with standardized test fixtures. Test hardware design for 20Gbps and above is still at its infancy for potential volume productions.

#### 7.5.1.4 DFT AND TFD

The basic Pseudo-Random-Binary-Stream (PRBS) generator and a BER checker functions have been a norm for years now. In order to cope with adaptive equalizations, more Gbps/GHz designs now include internal data eye analysis capability. As more analog content is introduced into Gbps/GHz I/O design, more research and innovation is needed in analog DFT.

Because of the power and area requirements for GHz I/O design, the analog portion of the circuits is under constant pressure to be scaled down. As a process variation removal technique, the calibration and trimming become widely adopted. That leads to more digitally-assisted analog design implementations, which means more on-chip self-calibrations or trimming based on test are required. This has created a new test requirement – testing the robustness of the calibration firmware. The performance related to trimming also positioned the testing charter from defect detection into a new dimension of performance guarantee process. In other words, it is not only “design for testability” but also “test for design performance”.

## 7.6 MEMORY

Memory density will continue to grow over the roadmap period according to Moore’s Law but the density trend of all memory devices has slowed compared to previous roadmaps. Continued scaling of memory will require further developments in new materials. Memory I/O data rates will increase over the course of the roadmap with DRAM on the leading edge as compared to other memory types. DRAM will continue to be the I/O performance leader throughout the roadmap while NAND will dominate the density.

Memory test solutions have recently targeted by device (DRAM, NAND, NOR) type, but the recent requirement to easily flex capacity between the various memory types is driving more common solutions, especially in the wafer test functional test environment. The great disparity in I/O data rate across the various memory types will make development of a single low cost solution difficult for performance test.

*Table TST8 - Memory Test Requirements*

### 7.6.1 DRAM

DRAM bit density and I/O performance is projected to increase over the roadmap. DRAM will reach a data rate of 8.4 Gb/s per I/O in 2022 based upon the DRAM historical model that forecasts data rates thru DDR6, but the possible addition of protocol interfaces on future generations may increase that data rate. Increasing memory size will drive higher device test time and decreasing manufacturing cell throughput unless offset by DFT or multi-insertion testing is to minimize cost. Failure detection, analysis, and repair are necessary for commodity DRAM as is the need to test in

reliability. To enhance test productivity, new test-oriented architectures will be required. Multi-bit testing, BIST, and built-in self repair will be essential to maintain the production throughput and yield.

Considerable test parallelism increase from ATE will be required over the roadmap in order to manage test cost. Commodity DRAM will lag the leading edge specialty DRAM in I/O bit rate. In the realm above 2 Gb/s DUT interface signal integrity issues exist with the test socket so probe card and Parallel test exceeding 128 devices per test head will be a challenge due to the interface routing complexity required. These challenges will ultimately drive the need for die self test. Performance DRAM I/O bit rate will be at least two times greater than the HVM production bit rate. High I/O bit rates create a challenge on how to define tester accuracy in the future properly. The traditional I/O bus width of 4, 8, and 16 bits has been supplemented by bus widths of 32bits that are being driven by mobile devices. The bus on high volume commodity DRAM may transition to 32 I/O near the end of the roadmap driven by potential differential bus requirements.

## 7.6.2 FLASH

NAND will double in density ever year in the short term and slow to a doubling every 2 years. The doubling every 2 years will be faster than the projected lithography node migration due to the further increase in the number of bits stored in a single memory cell from one and two to four on some cell types. NAND density has generally been 4× the NOR density at any given technology generation and the forecast shows that trend continuing. Use of error correction allows greater uncorrected error rates and enables increased bits per memory cell. NAND bus width has continued to be primarily 8 bit with some products at 16 bit.

NOR memory density is expected to increase slowly over the roadmap period and remain flat toward the end of the roadmap. NOR was projected in previous roadmaps have 32 I/O products, but 32 I/O for NOR has been removed starting with the 2007 roadmap. NOR products have been either 8 or 16 bit, but there has been a rapid increase in the number of designs with x1 to x4 serial I/O. I/O data rates are forecasted to increase to 533 Mb/s over the roadmap period to keep pace with faster ASIC devices used in embedded applications.

NAND and NOR generally have not had the same test solution due to architectural, usage, and bad block handling differences. However, bus differences between NAND and NOR have blurred over time and the Serial Peripheral Interface (SPI) bus and other serial bus definitions are targeted to support embedded applications. Further proliferation of bus types is expected due to the customization of flash for applications. Bus width is predominantly 8-bit and 16-bit on densities > 4 Mb. The 32 bit bus width NAND has been dropped from the roadmap starting in 2007.

The need for internal voltages that are 3–8 times the external supply requirements is expected to continue, driven by the hot electron and Fowler-Nordheim charge transport mechanisms. Increased absolute accuracy of supply voltages will be required in the future due to the trend toward lower voltages, but percentage accuracy relative to the supply voltage will remain a constant. I/O voltage decreases are pushing the operation limits of standard tester load circuits; new methods will be required in the future.

Wafer test generally does not require the performance of package test, but error detection, error analysis, and redundancy processing is required. Stacking of various types of Flash and other memory and/or logic components in a single package has become standard and is expected to continue. Multiple die within a package has complicated the package test requirements and has increased pin count and number of DUT power supplies required. Data and clock rates for flash will increase, but there is expected to be a wide variability in the requirements based upon the end application.

## 7.6.3 EMBEDDED MEMORY

Embedded DRAM bits will not match the density growth rate of commodity DRAM and NAND. The major concern for a merged logic-DRAM design in a dual-gate process will be array noise and sense-amp imbalance. For the 100 nm DRAM half pitch and below, DFT for inline defect detection will be necessary for product development.

Embedded Flash memory bits will grow exponentially in the near term and then double every two years in the later years of the roadmap. More devices will include both DRAM and Flash memory. Oxide reliability, sense-amp imbalance, and oxide-nitride-oxide (ONO) scaling will be the major concerns in flash memories in the future.

To enhance test productivity, new test-oriented architectures will be required. Built-in self-test and built-in self-repair will be essential to test embedded DRAM and embedded Flash memories and to maintain production throughput and yield. The primary test algorithms for Flash memories will continue to be Read-disturb, Program-disturb, and Erase-disturb while March tests with all data backgrounds will be essential for embedded DRAM.

Considerable parallelism in test will be required to maintain test throughput in the face of rising memory densities. In some cases, test is made cost-effective by double insertion of devices rather than testing both logic and embedded memories on the same tester. In double insertion, embedded Flash and DRAM could be tested and repaired on the

memory tester, while the logic blocks are tested on the logic tester. Embedded SRAM test requirements are captured in the High Performance Microprocessor section of the chapter.

## 7.7 ANALOG AND MIXED-SIGNAL

The economic benefit of monolithic integration (SoC) and single package integration (SiP) is well established. This integration has combined digital, analog, power management, mixed signal, and RF/microwave circuitry routinely in a single package and often on the same die. This trend has increased the breadth of interface types on a single part, and given rise to test equipment that mirrors this range with a corresponding breadth of instruments. Now this trend has again escalated with the emergence of through silicon via (TSV) packaging technology driving the challenge in a 3rd dimension.

Another important trend impacting mixed signal and analog testing is the compelling economics of multi-site testing for devices manufactured in extremely high volumes, also called parallel test. To support parallel test, many more instrument channels of each interface type are required to keep test cell throughput high.

The increasing number of interfaces per device and the increasing number of devices tested simultaneously raise the need to process an increasing amount of data in real time. The data from the mixed signal and analog circuitry is non-deterministic and must be processed to determine device quality. This processing must be done in real time or done in parallel with other testing operations to keep test cell throughput high. In fact, as site count increases, overall throughput can decrease if good Parallel Test Efficiency (PTE) is not maintained.

Looking forward, the breadth, performance, density, and data processing capability of ATE instrumentation will need to improve significantly to provide the needed economics. The area undergoing the most change is RF/microwave and so it is covered in its own separate section. The digital and high speed serial requirements for mixed signal devices are equivalent to logic and are covered in that section.

This section focuses on analog/mixed-signal test requirements. The Mixed-Signal Test Requirements table focuses on test instruments rather than specific chip applications. The test instrumentation must often cover more than one device market segment to provide sufficient utilization in a single tester configuration, so the requirements for multiple segments are aggregated into a few instrument categories. The analog waveform generation and capture requirements are set in two classes: low frequency—basic/minimum requirements for a mixed-signal ATE and very high frequency high-end requirements. Where appropriate, the mixed-signal instrument requirements are linked to other sections and tables in the roadmap.

There are two important trends. The first is to deliver adequate quality of test. Most analog/mixed-signal testing is done functionally. This requires instrumentation capable of accurately generating and analyzing signals in the bandwidths and resolutions of the device's end market application. Both of these parameters are trending upwards as more information is communicated between devices and/or devices and the physical environment.

The second key trend is to enable the economics of parallel test through instrumentation density and parallel test efficiency, a measure of the overhead in testing multiple parts. The level of parallelism shown in Table TST2 "Multi-site Test for Product Segments" indicates an increase in instrumentation density.

These trends of increasing ATE instrument channel count, complexity, and performance are expected to continue, but at the same time the cost of test must be driven lower (see the areas of concern listed below).

Analog/mixed-signal DFT and BIST techniques are lagging. No proven alternative to performance-based analog testing exists and more research in this area is needed. Analog BIST has been suggested as a possible solution and an area for more research. Fundamental research is needed to identify techniques that enable reduction of test instrument complexity or elimination of the need for external instrumentation.

*Table TST9 - Mixed-signal Test Requirements*

### 7.7.1 IMPORTANT AREAS OF CONCERN

- Time-to-market and time-to-revenue issues are driving test to be fully ready at first silicon. The analog/mixed-signal test environment seriously complicates the test fixtures and test methodologies. Noise, crosstalk on signal traces, added circuitry, load board design complexity, and ATE hardware/software issues currently dominate the test development process and schedule. The test development process must become shorter and more automated to keep up with design. In addition, the ability to re-use analog/mixed-signal test IP is needed.

- Increased use of multi-site parallel and concurrent test of all analog/mixed-signal chips is needed to reduce test time, in order to increase manufacturing cell throughput, and to reduce test cost. All ATE instrument types, including DC, will need multiple channels capable of concurrent/parallel operation and, where appropriate, fast parallel execution of DSP algorithms (FFTs, etc) to process results. In addition, the cost per channel must continue to drop on these instruments as the density continues to increase in support of parallel test drivers.
- Improvements in analog/mixed-signal DFT and BIST are needed to support item 1 and 2 above.

## 7.8 RADIO FREQUENCY

Four main RF frequency areas are distinguished. The low frequency range up to 3 GHz is dominated today by long distance cell phone communications technology. With wireless client communications protocol (WiMax) this will transfer over time to the higher frequency bands. The 3-6 GHz range is used by satellite TV and 802.11a networking. The 6–45 GHz band is focused on medium distance communications (Bluetooth moving to ultra wide band (UWB)). The high section, 45-94 GHz, is used primarily for short distance radar applications, particularly automotive, but 60GHz will also be used for congested area (short range) WLAN and satellite to satellite communication. CMOS devices are expected to be able to support 60GHz.

The most important movement is the increase in frequency: using higher frequency bands in line with the 802.11 and 802.16 communication standards. For the high frequency ranges (> 12 GHz) classical test techniques (non-modulated) are expected to fulfill the requirements.

An important requirement for test is full synchronization between the power/digital/AC baseband part of the tester and the RF instruments. Error vector magnitude measurements are a prerequisite. An important trend is the move of RF into SoC and SiP solutions. This requires not only test solutions for the RF parameters, but also one in combination with high-end digital and mixed signal requirements, with the potential to test these functions concurrently if practical. For SiP, the wafer probe capability becomes important. To cope with the economics of RF becoming a real commodity, multi-site will also be mandatory and will increase in this application area. The tooling (load boards, sockets, probe cards) is also critical to ensure signal integrity to and from the DUT. Looking at these challenges, the need for specific design-for-test for RF and finding lower cost alternatives to functionally testing RF devices is expected to increase heavily in the near term.

*Table TST10 - RF Test Requirements*

### 7.8.1 IMPORTANT AREAS OF CONCERN

- The increase in performance and frequency in combination with the economics of test will put a strong focus on novel design-for-test and alternative test techniques to be developed in the coming years.
- RF will much more frequently be embedded into products via SoC or SiP techniques. Combination of RF tests with (high-end) digital and mixed signal will be more common. RF test on wafer level will increase. Next to the test system, there will also be emphasis on the tooling (load boards, sockets, and probe cards) to cope with signal integrity.
- Source and measurement accuracy for phase noise and signal detectors are adequate today but must improve in the near term. Phase noise at 100 kHz needs to improve to better than -140 dBc/Hz over the next couple years.
- The visibility of trends in the high RF ranges (> 6 GHz) is evolving.
- Probe is a challenge over 6 GHz and the timeframe for modulated waveform testing over 6 GHz is uncertain.
- Impedance standards and calibration methods for high frequency measurement at probe need to be created.
- The EMI environment of the test development setup may be substantially different than the environment on the production test floor creating yield and correlation issues.
- Most SOC test requirements still trending from RF-to-BB and vice-versa. RF-to-Dig or line to line is limited but widely discussed.
- OEM is conceptualizing RF BIST/Loop back test methodology but functional and parametric tests still dominate the market.

## 7.9 RELIABILITY TECHNOLOGY REQUIREMENTS

Reliability solutions are an optimization of 1) reliability defect density (RDD), 2) learning, reliability screens, and test methods (RS&TM) applications, and 3) design for reliability (DFR). The goal of the reliability solution optimization is to

provide the best value for the reliability dollar spent, where value is defined as the ratio of customer satisfaction to customer cost.

In reliability circles, customer satisfaction is measured by the field failure rate or failures in time (FITs). The cost of reliability has two components: manufacturing operations costs and yield. As such, these two components of the reliability cost equation are the primary challenges facing every reliability solution provider. In turn, manufacturing operations costs are also driven by two fundamental components—burn in duration and equipment sophistication. The industry is still searching for a means to accelerate latent defects outside of the traditional elevated voltage and temperature methods. It follows that much progress has been made in detection techniques, but acceleration remains all about applying elevated voltage and temperature.

Applying voltages and temperatures in excess of the application specs most often demand providing solutions to leakage induced power (electrical power delivery and thermal/heat dissipation). The component of reliability cost reduction associated with yield is severely biased towards elimination of "overkill"/"false rejects," which in many ways are tied to derivatives of the power solution. However, the primary source of false rejects stems back to the stress methodology, through the modeling assumptions, and ultimately finds its root in what one believes about "escapes" from the manufacturing stress process. The argument proceeds like this: the majority of market applications are most concerned with the early life component of the failure rate. Most latent defects that "escape" acceleration will fail early in the product life. In order to guarantee a part received stimulus—and therefore did not "escape" stress—is simply to measure the outputs during stress. Defining terms, measuring outputs is called in situ stress, while measuring no outputs is dynamic stress. Obviously the "escapes" component is less for in situ, and hence the early life failure rate is lower. As anticipated however, this lower failure rate does not come without cost. In situ stress requires functionality at stress conditions, which in turn shrinks the performance distribution. Completing the original thread, measuring outputs during stress also introduces a component of yield loss. Due to process variation, some portion of the distribution does not have sufficient margin to function at stress voltages or temperatures, however these same parts operate fine at application conditions. Although these parts contain no reliability defects, in situ stress will fail these perfectly functional parts—hence "over-kill." These same parts with "marginal margin" are the target of the advances in detection techniques mentioned earlier. Achieving reliability requires trade-offs. In most instances performance and yield hang in the balance.

Reliability defect density learning rate is the most cost effective means of achieving the reliability demands of the marketplace. In itself, it is the by-product of the fundamental core practice in achieving profitability in microelectronics, yield learning rate. Defect learning is addressed in the Defect Modeling and Physical Defects section—and although historical data has overwhelmingly supported the premise that the component of defects that are "reliability unique" has been small—recent advances in technology may be changing the picture. The section on defect learning will always be directly applicable to RDD learning; however the high voltages and temperatures of defect acceleration are causing us to peer over the edge of device physics and materials science. Stress conditions are no longer dictated by "technology nominal" specs but by system application conditions. Technology's recent inability to meet marketplace performance demands at reasonable power has forced systems designers to increase system application conditions (voltage and temperature) to compensate. Shifts in array Vmin operating range, NBTI-driven performance margin, and gate oxide integrity (time dependent dielectric breakdown (TDDB)) as a result of the application of stress conditions still remain largely unexplained. As such, they dictate compensatory actions and/or reliability failure rate modifications. Even the standard thinking of metal electro migration for C4 and BEOL wiring requires careful scrutiny when confronted with the radical currents and powers conjured up by stress conditions. The industry's ride on the "Performance Juggernaut" isn't over quite yet.

DFR also has three key components: 1) technology design, 2) chip design (logical and physical), and 3) system design. In each of the three, the DFR work must strive for defect tolerance. In the case of technology design, leakage induced power mitigation maintains an edge in importance over defect tolerance. Regarding chip design and DFR, power mitigation and fault tolerance are at par in design priority. Redundant element analysis and power dissipation analysis burn considerable design engineering horsepower. At the system level, defect tolerance exists in the forms of error detection/correction and redundant elements.

In the arena of reliability screens and test methods, the literature is rich with techniques and methodologies with champions and supporting/compelling/biased data. Debates vary, depending upon the technology generation, chip/circuit type, design style, performance target, reliability requirements, and defect type. As long as excessive voltage and temperature retain the throne of defect acceleration, RS&TM will challenge the best and brightest minds in power delivery and thermal solutions. One must be able to accelerate defects while avoiding destroying the device—which is a change in precedence. In years past, stress conditions or actions that invoked or even hinted upon wear-out were to be avoided. The adage in the past was "one must be able to accelerate defects while avoiding the onset of wear-out."

However this is becoming increasingly more difficult in the face of stretched system applications conditions; sub-10 nm oxides; NBTI; marginal margin (that is, array Vmin); hundreds of amps and Watts, miles of copper wire, and billions of interconnects.

RS&TM are best categorized by separating them into wafer applications and package (or module) applications and then further segregation into "detection" and "acceleration" techniques. This tiered structure will help to dilute the perennial argument between test and reliability regarding whether a field return is a test escape or an early life reliability failure.

Regardless of operational process step (wafer or package), acceleration techniques invariably must deal with potent power implications simply because acceleration requires temperature and/or voltage far in excess of application conditions—and leakage varies exponentially with both. The same is not true for detection techniques. In many instances, detection techniques employ conditions that reduce leakage (that is, VLV (very low voltage) or VLT (very low temperature), and in instances where detection requires application conditions that exacerbate leakage, those conditions typically do not approach the level of acceleration conditions.

## 7.9.1 BURN-IN REQUIREMENTS

Technical challenges for the burn-in process are driven by increasing device pin count, decreasing package pitch, increasing device functionality and operating frequencies, dramatically increasing leakage current, and eroding voltage/thermal acceleration. Several alternate techniques such as IDDQ, high voltage stress, and wafer mapping are being used to try to improve device reliability, since many reliability failure modes are proving to be resistant to burn-in.

Burn-in system technology must continue to drive down costs, in particular for high power devices. The minimum device core voltage continues to decrease. Scan requires very deep vectors for large memories, while high power requires individual device thermal and power management. The burn-in process (system/driver/burn-in board/socket) will be challenged to meet speeds of the newest technology devices without some form of internally generated clock. Devices without DFT are requiring increasing I/O. The growing need for KGD continues to drive efforts for wafer level burn-in, KGD carriers, or additional stress during probe.

Device power and signal requirements are driving burn-in boards toward higher board layer counts, smaller traces, less space for routing, more complex processes and materials, higher test costs, and board reliability issues. Tight pitch on future devices will require new cost-effective, innovative interfaces between the burn-in sockets and the burn-in boards.

Burn-in sockets are undergoing major design challenges as they must accommodate increasing contact count, decreasing pitch, higher currents, and higher frequencies. At the same time, sockets are a key component of an overall thermal solution designed to prevent high power devices from self-destructing. A major challenge for socket manufacturers is to maintain low costs and short lead times while providing the technology to meet these new demands. Horizontally actuated contact design will be displaced below 0.5 mm pitch ball grid array (BGA) by vertically actuated contacts as pin count increases and existing socket materials fall short of increased mechanical stress requirements. New designs and new materials will be required for higher current carrying capabilities. Socket design will need to accommodate looser packaging specs in areas such as warpage and package dimensions, while coping with increased package size, thinner/more fragile packages, and reduced/non-standard/mixed pitches. Contact design will need to provide greater strength without a loss of electrical/mechanical performance.

Approaches to burn-in include traditional unit level burn-in, system level burn-in, wafer level burn-in, and strip/array burn-in (Figure TST11). On high reliability applications, system level burn-in complements or replaces traditional device level burn-in. Wafer level burn-in technology continues to be developed, but has not been able yet to make significant inroads against traditional packaged level burn-in. The challenge here is to eliminate socketted burn-in and find ways to perform simultaneous multiple wafer level burn-in using scan/logic and memory BIST (MBIST). Strip/array burn-in is becoming more important as more packages are receiving massively parallel test in either strip or array format.

*Table TST11 - Burn-In Test Requirements*

## 7.9.2 WAFER LEVEL BURN-IN

There is no standard definition of what constitutes wafer level burn-in (WLBI). Some vendors use the term "burn-in" to refer to the application of a simple DC stress that applies opposite voltage potential to the internal nodes of a DRAM. Some say that WLBI requires full wafer contact and the application of high enough temperature over enough time to activate thermal defects, while also applying voltage stress with the device operating in "normal" mode. Some vendors enable the use of WLBI for low-end micro-controllers or SoC through DFT functions such as scan or BIST.

Key challenges are to quantify how to measure the effectiveness of these options and to develop standards that define WLBI and the methods that are used to confirm the effectiveness of this wafer level treatment. The challenge for DRAM in particular, as a device well suited for WLBI, is to provide a burn-in environment for wafers that provides the same functionality, is as effective as package-level burn-in, and yet is no more costly. The concept is to leverage the time spent in burn-in by using the burn-in environment as a massively parallel testing opportunity.

The need for WLBI is increasing. The infant mortality rate is getting worse due to transistor scaling effects and new processing technology / materials for devices. Decreasing operating voltages and margins for devices are reducing the ability to use voltage acceleration / voltage stress testing to guarantee reliability. KGD is becoming a more significant need by the customers due to requirements for chip scale packaging and multi-chip modules. Decreased cycle time and the need for faster feedback of yield / defect information to the wafer fab can be assisted by moving burn-in earlier in the overall semiconductor process. Finally, detection and removal of defective devices prior to the packaging process eliminates packaging scrap costs based on intrinsic device defects.

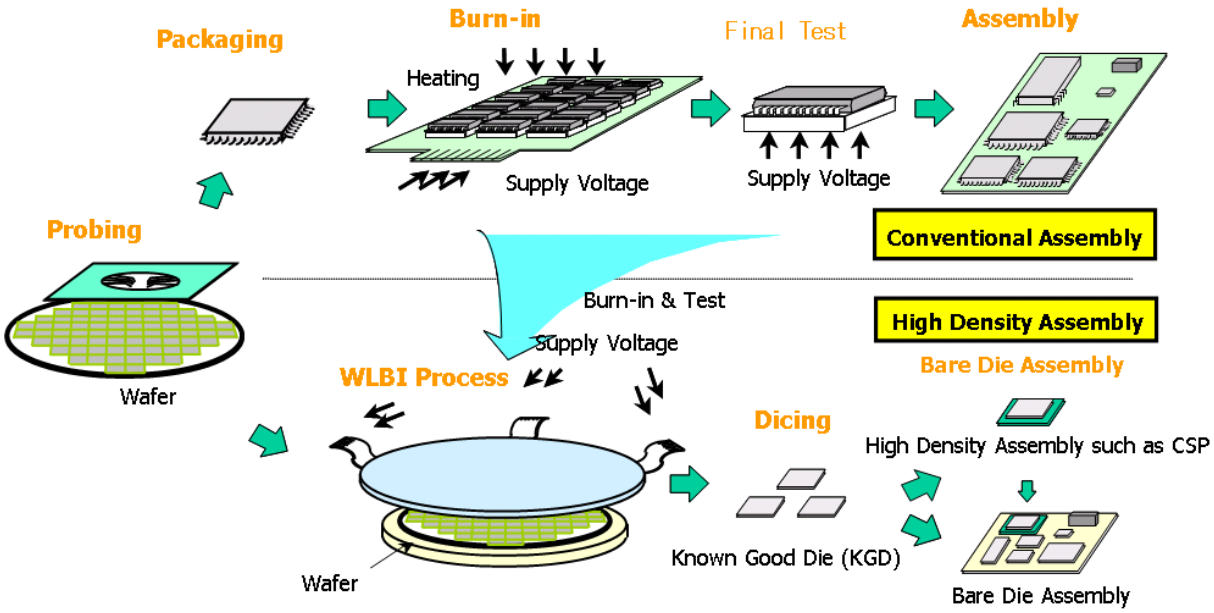



*Figure TST11 - The Production Process with WLBI Compared with Package Burn-in*

## 7.9.3 PROBING TECHNOLOGY FOR WAFER LEVEL BURN-IN

Contactors for whole wafer contact include TPS probe and micro pogo-pin contactor. TPS probe consists of a substrate board, membrane with bumps, and PCR sheet, where the PCR sheet between two components absorbs the uneven height of bumps to achieve uniform and stable contact. Significant features of this system include the ability to concentrate pressure efficiently at each bump top and the ability to achieve over 20,000 bumps contact with Al pads by control of bump material and surface condition. Materials with coefficient of thermal expansion (CTE) similar to Si (such as glass and ceramics) are used for the substrate board to prevent CTE mismatch.

A micro pogo-pin contactor consists of a CTE matched probe housing and pogo-pins with moving plungers at both sides. The pogo-pins stand vertically and have enough compliance and independent travel to accommodate height variations between adjacent contacts. The probe pitch is technology dependent.

Other contactors such as spring-functioned material on wafer pads as a wafer level package technology are expected for whole wafer contactor usage. For contactor roadmaps, DRAM is selected as the target application due to its large predominance in general memory burn-in. DFT is considered for system LSI.

## 7.9.4 OTHER WLBI TECHNOLOGY CONSIDERATIONS

The current consumption of a wafer increases by sub-threshold leakage from shorter transistor channel lengths and an increased number of transistors per unit area. The high temperature of burn-in also increases sub-threshold leakage. Therefore, the burn-in equipment must be capable of supplying over 1000 A of current per wafer. Also, to manage current

appropriately, wafer temperature control/uniformity becomes necessary. Finally, the burn-in equipment must be able to accommodate different quality distributions across each wafer.

BIST is capable of decreasing the number of pins under test per device, but die shrinks and tighter pad pitches more than offset this advantage by increasing the total number of die and pads per wafer. The increased number of pins being tested also increases the force required to contact the wafer. In order to enable the use of WLBI through DFT functions such as scan, BIST, and JTAG[3], the number of tested pins per device and total cost per device must be decreased and performance of the WLBI technology must be improved.

The probing technology for WLBI is described above. However, probing technology faces several challenges in order to meet the technology trend for the future. When the probe pitch is studied with TPS contactor technology for the devices with LOC and peripheral, a pitch around 70 µm seems feasible. For a pitch less than 70 µm, MEMS technology by use of photolithography is an option. This technology, however, does not yet have a solution for 300 mm wafers. While probing technology for tighter pitches is required, the intelligent use of DFT during pad layout may provide some relief by bypassing every other pad in order to double the probe pitch effectively, as compared to pad pitch. Application to high pin count and low force probing due to low-κ materials will also be required. This will help drive new probing technology.

## 7.10 TEST MECHANICAL HANDLING REQUIREMENTS

Wafer probe and component test handling equipment face significant technical challenges in each market segment. Common issues on both platforms include higher parallelism and increasing capital equipment and interface cost.

Increased parallelism at wafer probe drives a greater span of probes across the wafer surface which ultimately results in full wafer test across a 300 mm wafer. The increased probe count is driving interface complexity to route signals. Prober and probe card architecture will need to evolve to simplify the interface.

Reducing the cost of wafer-level and package-level test in the face of more challenging technology and performance requirements is a constant goal. The demand for higher throughput must be met by either increased parallelism (even with short test times), faster handler speed or process improvements such as asynchronous test or continuous lot processing.

Packages continue to shrink, substrates are getting thinner, and the package areas available for handling are getting smaller at the same time that the lead/ball/pad count is increasing. In the future, handlers will need the capability to very accurately pick and place, small, fragile parts, yet apply similar or increasing insertion force without inducing damage.

Temperature ranges are expanding to meet more stringent end use conditions and there is a need for better control of the junction temperature at the start of test. Power dissipation overall appears to be increasing, but multi-core technology is offering relief in some areas.

It is unlikely that there will be one handler that is all things to all users. Integration of all of the technology to meet thermal, throughput, placement accuracy, parallelism, and special handling needs while being cost effective in a competitive environment is a significant challenge. The 2007 roadmap defined three handler groupings based upon the power requirements of the DUT. High power DUTs consume greater than 10 watts, medium power devices are between 0.5 and 10 watts and low power device have less than 0.5 watts per DUT. For 2009, the high power handler category has been divided into 10-50W and >50W to reflect typical and extreme devices.

[3] *an IEEE standard 1149 boundary scan*

| | |
|---|---|
| Med-High and High Power Handler | *Temperature control and temperature rise control due to high power densities during test* |
| | *Continuous lot processing (lot cascading), auto-retest, asynchronous device socketing with low-conversion times* |
| | *Better ESD control as products are more sensitive to ESD and on-die protection circuitry increases cost.* |
| | *Lower stress socketing, low-cost change kits, higher I/O count for new package technologies* |
| | *Package heat lids change thermal characteristics of device and hander* |
| | *Multi-site handling capability for short test time devices (1–7 seconds)* |
| Medium Power Handler | *Support for stacked die packaging and thin die packaging* |
| | *Wide range tri-temperature soak requirements (-45ºC to 150ºC) increases system complexity* |
| | *Continuous lot processing (lot cascading), auto-retest, low conversion times, asynchronous operation* |
| | *Shielding issues associated with high frequency testing (>10 GHz)* |
| Low Power Handler | *A wide variety of package sizes, thicknesses, and ball pitches requires kitless handlers with thin-die handling capability* |
| | *Package ball-to-package edge gap decreases from 0.6 mm to 0 mm require new handling and socketing methods* |
| | *Handling package smaller then 1x2 mm for specialist device MEMS or photo sensor with low damage concern* |
| | *Parallelism at greater than x128 drives thermal control and alignment challenges* |
| Prober | *Consistent and low thermal resistance across chuck is required to improve temperature control of device under test* |
| | *Heat dissipation of >100 Watts at > 85ºC is a configuration gap in the prober industry* |
| | *3DI and MEMS application require very thin wafer and special shape wafer testing (handling) technology, but no industry standard resist its widely spread.* |
| | *Advances in probe card technology require a new optical alignment methodology.* |
| | *Dicing flame wafer probing require high-temp. dicing flame sheet to be developed.* |
| | *New wafer materials require heavier wafer handling.* |
| | *Power Device application require very thin and warp wafer, thus 'Taiko Wafer' and ' Ring attached wafer ' handling technology became trend.*<br>*Also no thin wafer / warp wafer standard cassette / FOUP nether exist.* |
| | *Beyond 300mm wafer size, Industry have to consider how to implement tester / prober to the test floor. Those size, weight and height (more than 2x for each) may change test floor layout, operation and its building design.* |

*Table TST12- Test Handler and Prober Difficult Challenges*

*Table TST13 - Prober Requirements*

*Table TST14 - Handler Requirements*

## 7.11 DEVICE INTERFACE TECHNOLOGY REQUIREMENTS

As device I/O bandwidth and power demands increase there is a corresponding requirement for high performance power and signal delivery during electrical test. These requirements drive challenges for the assemblies used to interface the test equipment to the device under test. The highest performance interfaces require complete power and signal path modeling from the source instrument to the die, requiring accurate simulation models of the test instrument, path, probe or socket, and die. Shrinking die and package geometries as well as manufacturing productivity further complicate these interfaces with decreasing pitch, increasing pin count and multi-DUT requirements.

'More than Moore' improvements are being tackled by innovative heterogeneous architectures such as SiP and Die Stacking. Innovations such as TSV and Proximity communications (non-contact signaling) are being explored to address the density issue for SiP and stacked dies. SiP and staked die (KGD) present major challenges from a yield and test cost point of view. Three dimensional interconnects, whether wired or wireless, create alignment and physical challenges for testing and probing.

Test Insertion into packaging lines is a potential solution for 'More than Moore' problems. Traditional Semiconductor fabs have multiple off- line and in-line in process testing. Packaging has traditionally been carried out with a small number of active components. Stacked die and SiP packaging combine multiple active dies and passive components. Testing for packaged parts has traditionally been carried out using post assembly testing.

To achieve 'More than Moore' improvements packaging will likely need to adopt semiconductor like practices with respect to test; including in-line test for yield learning and production.

For 'More than Moore' the need for higher density of interconnects and KGD both present a challenge to probing. Migration of tester resources onto probe cards or DUTs is continuing thru BIST or enhanced probe cards. A potential solution to parallelism and testing is to use non-contact interconnects as a method of high density wireless probing. Coordination or test coverage is an ongoing challenge with heterogeneous designs. Solutions may involve the addition of test structures to assemblies or instantiated dies.

## 7.11.1 PROBE CARDS

Wafer probe technologies face complex electrical and mechanical challenges driven by product specifications, test implementation requirements, test productivity goals, and reduced test cost demands. Across the device spectrum, these challenges include: higher average power demands, higher frequency response (bandwidth), rising pin counts across tighter pitches and smaller pads/bumps, increasing switching currents (di/dt), alternative pad/bump metallurgies and increasing test parallelism. Research and development of new or improved probe technologies is required to meet these challenges to ensure that the basic probing requirement of ensuring reliable, sound, and cost-effective electrical contact to the device(s) under test is achieved. Recent developments in contactless probing technology may help address future density, speed and 3D requirements for probe.

Improvements in passive probe cards are at appear to be approaching mechanical and electrical limits for increased functionality. Intelligent Probe Cards are potentially capable of solving problems of both parallelism and speed. Advances in materials MEMS and heterogeneous integration suggest that intelligent probe cards can be made economically with parallelism and performance to match DUT technologies. There are some extant commercial offerings of intelligent probe cards for use in 3D packaging test.

The tables contained in this section derive trends based on product families similar to the layout of the Test Technology Requirements section above.

## 7.11.2 TRENDS AFFECTING PROBE CARD TECHNOLOGIES

Along with addressing the key challenges listed below, research and development is urgently required to bring to the market cost-effective probe technologies directed at trends in product offerings and the testing environment.

The continuing volume growth of bumped devices, often with I/O in area arrays, points to the escalating demand for "vertical" style probe card technologies, with a rising need in multi-DUT configurations as well. Multi-row wirebond also supports this vertical style need and is particularly challenging due to tighter pitches.

Some microprocessor products and high end ASIC devices are driving power levels to 500 Watts and 1000 Watts with associated current/probe and thermal issues. Current/needle is also an issue for cantilever and MEMS technology as wire bond devices move into higher technology silicon.

Manufacturing test of devices has moved to parallel test. For some product groups (e.g., memory), wafer probe technologies are available that handle parallel testing of 512 and more devices. Probe technologies capable of full wafer contacting are in use already for 200 mm and300 mm wafers. Increasing the contacts/DUT for these massively parallel probes is the next challenge.

Innovation in test is required for effective use of new interconnect technologies such as TSV or proximity communications.

*Table TST15 - Probing Difficult Challenges*

| | |
|---|---|
| Geometry | *Probe technologies to support peripheral fine pitch probe of 23 µm peripheral staggered pad probes at effective pitches of 20/40, and fine pitch (45 µm) for dual row, non-staggered probing on all four die sides.*<br><br>*Fine pitch vertical probe technologies to support 130 µm pitch area array solder bump and 50 µm pitch staggered pad devices.*<br><br>*Multi-site pad probing technologies with corner pitch capability below 125 µm.*<br><br>*Reduction of pad damage at probe commensurate with pad size reductions (or better).*<br><br>*Alternative probe technology for 75 µm on 150 µm pitch dense array (vertical probe; bumped device).*<br><br>*Increasing probe array planarity requirements in combination with increasing array size.* |
| Parallel test | *Need a probe technology to handle the complexity of SoC devices while probing more than one device.*<br><br>*Current probe technologies have I/O limitations for bumped device probes.* |
| Probing at temperature | *Reduce effects on probes for non-ambient testing -50°C to 150°C; especially for fine-pitch devices.*<br><br>*For effects on Handlers and Probers, see that section.* |
| Product | *Probe technologies to direct probe on copper bond pads including various oxidation considerations.*<br><br>*Probe technologies for probing over active circuitry (including flip-chip).* |
| Probe force | *Reduce per pin force required for good contact resistance to lower total load for high pin count and multi DUT probe applications. Evaluation and reduction of probe force requirements to eliminate die damage, including interlayer dielectric damage with lo*<br><br>*A chuck motion model is required to minimize probe damage* |
| Probe cleaning | *Development of high temperature (85°C–150°C)* in situ *cleaning mediums/methods, particularly for fine pitch, multi-DUT, and non-traditional probes.*<br><br>*Reduction of cleaning requirements while maintaining electrical performance to increase lifetime.*<br><br>*A self cleaning probe card is required for fine pitch bumped pad devices* |
| Cost and delivery | *Fine pitch or high pin count probe cards are too expensive and take too long to build.*<br><br>*Time and cost to repair fine pitch or high pin count probe cards is very high.*<br><br>*The time between chip design completion ("tape-out") and the availability of wafers to be probed is less than the time required to design and build a probe card in almost every probe technology except traditional cantilever.*<br><br>*Space transformer lead times are too long, thus causing some vertical probe technologies to have lengthy lead-times.* |
| Probe metrology | *Tools are required that support fine pitch probe characterization and pad damage measurements.*<br><br>*Metrology correlation is needed for post repair test versus on-floor usage.* |
| High power devices | *Probe technologies will need to incorporate thermal management features capable of handling device power dissipations approaching 1000 Watts and the higher currents (≥ 1.5 amp) flowing through individual probe points.* |
| Contact resistance | *Probe technologies that achieve contact resistance <.5 Ohms initially and throughout use are needed.*<br><br>*A method to measure contact resistance is needed. The traditional continuity test is insufficient to monitor contact resistance.* |
| High frequency probing | *Traditional probe technologies do not have the necessary electrical bandwidth for higher frequency devices. At the top end are RF devices, requiring up to 40 GHz.* |

### 7.11.2.1 PROBE CARD TECHNOLOGY REQUIREMENTS

Many probe card technology types are available in the marketplace, each with suitability (technical and/or test operations driven) for probing certain device types and limitations that prevent more widespread use. There is no single probe technology capable of addressing the requirements across the entire device spectrum.

This section explores the challenges of probe technologies including those that are independent of the devices being probed. These include the resulting behavior of the probe when/after contacting the wafer, the design of the probe card to realize the productivity benefits of probing multiple die at the same time and the environment that the probe card is expected to operate within.

### 7.11.2.2 PITCH AND INTERCONNECT DEFORMATION

I/O density requirements are driving pad/bump sizes to ever-smaller sizes. It is well known that on the leading edge, wirebond pad pitches are under 30 µm (with resulting pad sizes naturally less than that). It is a formidable challenge for traditional probe technologies to scale down continually since with this scaling comes a parallel scaling-down of the permissible probe mark.

The use of cantilever probe cards for probing wirebond technologies, though still today's leading solution, is seen to be reaching practical limits in pitch and scrub within the nearer term horizon. Thus the newer emerging technologies, many using "semiconductor-like" processes (e.g., MEM and membrane structures) offer solutions for reduced pitch scrub requirements.

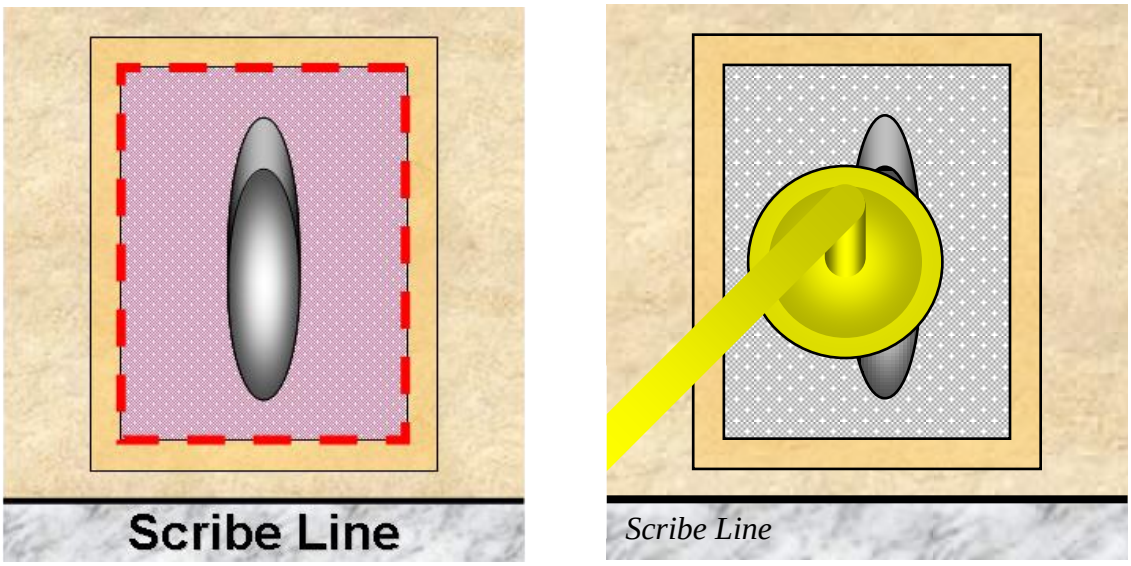



*Figure TST12 - Probing and Wirebond Contacting a Bond Pad*

Area array solder bumps are seeing growing application and driving the commensurate need/demand for vertical probing technologies. As the pitch/bump dimensions get smaller, current vertical technologies, typically an array of guided wires may also see their practical limit, thus requiring development of newer technologies.

### 7.11.2.3 MULTI-DUT

Productivity gains are often realized when testing (probing) more than one device in parallel. Memory testing has been a leader in this area, with leading edge approaching 500 devices in parallel. As *Table TST2* indicates virtually all memory testing is done in multi-DUT fashion. The move to multiple DUT testing within other product categories is already underway and is accelerating: with the use of DFT and "smart test" techniques, 16, 32, and even 64 DUTs is realizable and up to 4 DUT for high end microprocessors.

Multi-DUT probing requirements drive the need for more and more probe contacts across an ever-growing area. Today some new contact/probe technologies claim full wafer contact capability for 300 mm wafers. Ultimately increasing the contacts/DUT to hundreds will be required.

### 7.11.2.4 ELECTRICAL PERFORMANCE

Wafer probe technology, the probe card, provides electrical contact between the device(s) under test on a wafer and the test system electronics. The probe card must faithfully transmit/deliver device under test power and signals from/to the test system.

Within this ITRS document information can be found concerning device operating voltages and AC Characteristics. Additionally, within this Test and Test Equipment chapter tester performance information is provided on a wide range of electrical characteristics that may be helpful in understanding requirements for wafer probing.

There appears to be growth in the current carrying capability of individual probes contacts. At the same time the aggregate total current across the DUT is expected to rise with growing circuit densities and pin counts. Of note is that there are some selected applications that are seeing the need for higher and growing current carrying capability, approaching 1.5 amps and more. Of note is that peak values for transient currents are growing as well.

Contact resistance is always a closely watched probe technology parameter. It is influenced by many factors such as pad/bump metallurgy, contamination from pads/bumps, multi-DUT "off-stepping," contact force, scrub, cleaning, etc. The values shown in *Table TST16* requirements reflect contact resistance under 'normal' usage conditions over the practical lifetime of the probe. Initial and after cleaning requirements are often considerably lower, typically in the 200 milli-Ohm range or lower. There is a growing requirement for lower contact resistance values for longer periods (numbers of touchdowns) before cleaning.

High frequency testing in probe remains a challenge due to the lack of constant impedance structures to the die contact. The roadmap shows digital I/O performance will increase to greater than 20 GHz over the roadmap. Analog pin performance may reach 100 GHz driven by high precision proximity radar and automatic landing and automotive parking systems. High frequency test under High parallelism conditions remains a challenge and significant development is needed to support wafer level KGD (Known Good Die) test.

#### 7.11.2.5 THERMAL PERFORMANCE

Though stable through the roadmap horizon, the thermal environment for the probe is demanding. With low end chuck set-point requirements well below the freezing point and the upper end past the boiling point, the total range is wide - placing difficult demands on selecting materials that handle the extremes, but possibly more notably to deal with temperature co-efficient of expansion issues and high current demands.

Additionally, handling the heat produced by very high transient current heating effects and/or by high power products may drive the need for active thermal management within probers as well as an improved wafer to chuck thermal interface.

#### 7.11.2.6 UNIT COST AND COST OF OWNERSHIP

Probe card unit cost and cost of ownership (CoO) trends are not currently covered in this roadmap document. Though individual member companies may have their own approaches to unit cost and cost of ownership measurements and goals, there is a need to develop consistent models that can be used industry wide and cover the wide range of probe card technologies that are in the marketplace.

#### 7.11.2.7 CLEANING

Generally, online cleaning frequency for cantilever type probes rises slightly through the roadmap horizon, however increasing probe usage (touchdowns) before being taken offline for cleaning is being seen for many of the product families. The goal is better utilization of the test systems and the probe card.

For vertical probes, the rapidly growing number of touchdowns before online cleaning reflects a desire to reduce the vertical technologies' online cleaning frequency to match/better cantilever technologies more closely. Similar to cantilever technologies, the touchdowns before offline cleaning is increasing but across all product categories.

Notably, in some instances there is a move to eliminate online cleaning for memory products in the outer years of this roadmap's horizon. This is likely reflective of the design and/or complexity of probes with pin counts approaching full wafer contact.

*Table TST16 - Wafer Probe Technology Requirements*

### 7.11.3 TEST SOCKETS

The test socket is an electrical and mechanical interface responsible for good electrical connection and transference of high integrity signals between the DUT and the PCB/tester through a mechanical contact mechanism in order to determine the electrical characteristics of DUT. As the semi-conductor designing and manufacturing capabilities have progressed in recent years, the testing process keeps raising the electrical and mechanical requirements of test sockets. Therefore, the socket technologies have been rapidly driven by significantly enhanced electrical and mechanical requirements, both of which are instigated by higher power/voltage/current, reduced package size, tighter pitches, higher pin counts, smaller solder resist opening, and so on. It has been indicated that electrical properties are determined by not only the electrical but also by the mechanical requirements. The multi-physics problems have made socket designs progressively challenging along the higher requirements. Current models show difficulty in making sockets for high ball count devices and achieving I/O bandwidths of > 20GHz.

*Table TST17* contains the test socket technology requirements. The requirements have been divided into contacting NAND, DRAM, and SoC devices that are contained in TSOP, BGA, and BGA SoC packages respectively. The TSOP package is assumed to be contacted using a blade; the DRAM BGA is contacted with a spring probe and the SoC BGA is contacted with a 50 Ohm spring probe. The test socket performance capability is driven by the pitch between balls or leads, so the lead spacing of the assembly and packaging roadmap was used to determine the pitch.

Contact blades are generally used for testing TSOP NAND Flash and contain a spring function in its structure, which is loaded by compressing the DUT into the socket. The structure is very simple and suitable for HVM, however, the contactor blade must be long to maintain the specified contact force, stroke, and achieve a long mechanical lifetime. A weak point is that the blade contactor is not suitable for fine pitch devices due to the need to have isolation walls between adjacent pins. The thickness of the isolation wall must be thinner for finer pitches, which makes fabrication of the isolation wall more difficult. At the same time, the contactor blade thickness needs to be thinner for finer pitch, which complicates achieving the specified contact force, stroke requirement, and mechanical lifetime.

Spring probes are mainly used for testing BGA-DRAM device are formed by use of small-diameter cylindrical parts (probe and socket) and coil springs. Compression of the spring probe creates the contact load. In order to guarantee sufficient mechanical life, the probe diameter should to be large enough to guarantee strength and durability and the length should be long enough to maintain sufficient travel under compression. Negative characteristics of the spring probe are that it can only contact the pad/lead at one or two points and the contact resistance tends to be higher than other types of contactors. However, the spring probe structure is relatively simple and easy to maintain and it is also easy to design a DUT loadboard.

According to the BGA-DRAM roadmap, the spring probe diameter will need to be smaller over time driven by the finer pitch of the package ball roadmap. In addition, the spring probe will need to be shorter to meet the lower inductance values required to support the high frequencies of the roadmap I/O data rate.

Spring 50 Ohm probes are required for BGA-SoC high frequency devices have coaxial structures that can reduce probe length transmission issues through impedance matching. However, advances in the package ball pitch through the roadmap will create restrictions to the coaxial pin arrangement structure (0.5 mm pitch in year 2016). The data rate will increase to 20GT/s in 2016, but the spring 50 Ohm probe will not have good electrical performance due to its multiple parts structure having higher contact resistance than other contactors. To support 50milli-Ohms of contact resistance starting in 2010, advances will be required in materials, plating, and structure.

Conductive Rubber type contactors are used for BGA high frequency SoC devices. Conductive metal particles are aligned vertically in insulating silicone rubber which enables vertical contact and adjacent conductor isolation. Compared to other contacts, it is superior for used with high frequency device test since due to its low inductance and low contact height, but compression travel is limited. Conductive rubber will meet the fine pitch requirement in the road map, but it difficult to reduce contact force without decreasing the compression travel.

Contact blade + Rubber, generally used for testing QFP/QFN high frequency Soc, is a combined structure of a short length metal contact and compression rubber that makes contact thru force and travel. The required compression force can be varied by changing the rubber material, but the cycle life is normally shorter than a Contact Blade type contact.

It is capable to meet 15GT/s I/O data requirement in 2013 by shortening the metal contact, however, it is a challenge to ensure enough compression travel which is required in mass production environment. 40GT/s requirement in 2019 is also a challenge to make the metal contact 0.1nH inductance or less. It will need improvement of the structure, contact material, and plating.

Socket lifetime has not been pursued in this roadmap, but the lifetime problem will become more important in the near future as lead, ball and pad pitch becomes finer and pin counts get higher which drives lower contact force to avoid lead / ball damage. Pb-Free devices require higher contact forces than are required for non Pb-Free packages.

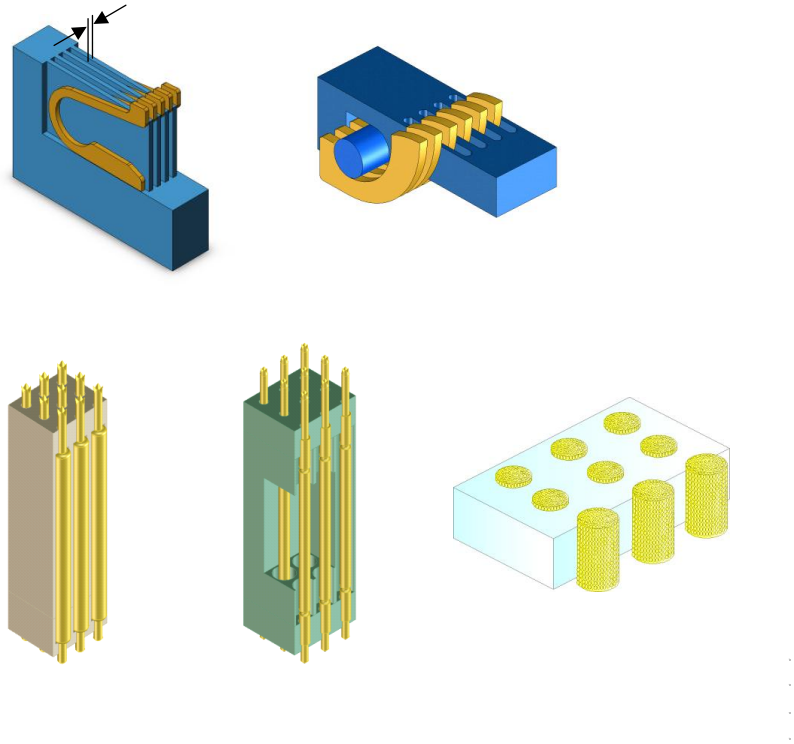


*Figure TST13 - Contactor Types*

### 7.11.3.1 ELECTRICAL REQUIREMENTS

Socket electrical requirements include current carrying capacity (CCC) per pin, contact resistance, inductance, impedance, and signal integrity parameters such as insertion loss, return loss, and cross-talk. The higher the power and bandwidth the packages are designed for, the higher the CCC, the lower the resistance, and the better matched the impedance of the pins and/or sockets need to be. Data rate requirements on over the roadmap are expected to exceed 20 GHz, which will greatly challenge impedance matching and the potential signal loss. As package size, solder resist opening, and pitches become smaller and pin counts higher, the smaller pins required to fit within tighter mechanical constraints will greatly increase contact resistance and signal integrity issues. One of the critical parameters to stabilize the electrical contact and insure low contact resistance is the contact force per pin, which generally ranges from 20 ~ 30 grams. As pitches get finer, smaller and more slender pins will be required, which may not be able to sustain a high enough contact force to have reasonable contact resistance. Due to the negative impact of mechanical requirements on electrical properties, it will be necessary to have improved electrical contact technologies or socketing innovations, in which the electrical properties and signal integrity will not be significantly impacted by or will be independent from raised mechanical requirements. To handle these high frequency signals, user has to carefully consider the signal integrity of overall test system including board design/components/socket.

### 7.11.3.2 MECHANICAL REQUIREMENTS

The mechanical requirements include mechanical alignment, compliance, and pin reliability. Mechanical alignment has been greatly challenged by higher pin counts and smaller solder resist openings, particularly in land grid array (LGA) applications. Currently, the majority of test sockets use passive alignment control in which the contact accuracy between pin and solder resist opening is determined by the tolerance stack-up of mechanical guiding mechanisms. The limit of passive alignment capability is quickly being reached because the manufacturing tolerance control is approximately a few microns. The employment of active alignment or an optical handling system is one of the options to enable continuous size reduction of package and solder resist opening, smaller pitches, and higher pin counts.

Compliance is considered as the mechanical contact accuracy in the third dimension (Z-dir.), in which the total contact stroke should take into account both the co-planarity of operating pin height and the non-flatness of the DUT pins, in addition to a minimum required pin compression. In general the total stroke of the contact is between 0.3 mm and

0.5 mm. However, as required pin sizes get smaller, it may not be feasible to maintain the same stroke and thus the compression issue may become the bottleneck of electrical contact performance.

Contactor pin reliability and pin tip wear-out have also experienced challenges because tight geometric constraints prevent adding redundant strength to the pins. The testing environment becomes more difficult with higher temperatures, higher currents, smaller pin tip contacts, etc.

*Table TST17 - Test Socket Technology Requirements*

## 7.12 SPECIALTY DEVICES

The test roadmap is not all inclusive so it does not contain test requirements for all devices. Many of the test requirements for some omitted devices fall within the bounds specified for devices within this roadmap. Other devices stretch the bounds specified in this chapter and need to be mentioned for completeness. Devices included in the specialty device section are high volume devices that are generally contained in and driven by the requirements of mobile communication and computing. The intent of this section is to document the challenges of specialty devices. For 2011, LCD display drivers, imaging devices and MEMS multimode sensor devices are featured in this section.

LCD display drivers are unique because of die form factor, which can have larger then 10:1 aspect ratio, and around thousand of very narrow gold bump pads requiring contact for test. In2011, LCD display drivers probing pad already was down to 10um width and 20um pitch in line structure which will continually narrow down in the future. Right now only cantilever probe card could achieve mass production probing such narrow and fine pitch pad with gold bump. The new feature of LCD display driver is high speed I/O protocol interface and the data speed is 2.5 Gbps in 2011 and will up to 4.8Gbps in 2019. The probing challenges become more severe and will motivate new technology to solve the probing problems with economic solution.

Imaging devices are required for every digital camera and have become a standard feature in cell phones and mobile computing devices. Digital cameras with 14 mega pixels of resolution are typical consumer devices in 2011 and will increase in resolution in the future. Automotive industry applications and surveillance need fast frame rates over 60 fps in2011. Each pixel of the image sensor has a micro lens to increase the light intensity and improve the signal to noise ratio. Some of personal assistants already implemented camera function which requested thin camera module down to 0,6mm high in 2011. Due to application of optical system, the micro-lens need to offset from optical axis of each sensor pixel a Current production solutions are generally proprietary and composed of a pupil optical system to optimum each sensor under test over a range of various angles. Back-side illumination BSI is another approach to increase incident light intensity and fill factor of sensor cell of each pixel. When image pixel size smaller then 1.4um, BSI process have advantage then FSI (front side illumination) process. Image sensors coupled with image processing digital logic in a single chip through 3DS( three dimension stack) IC and WLCCM (wafer level camera module) which integrated image sensor with lens system was achievable and applied to cell phone, But WLCCM was still limited on low resolution VGA application due to yield problems and optical quality. The production yield improvement and increasing through put of automation handling of image sensor 3DS IC and WLCCM still have some test challenges need to overcome in 2011. MEMS sensor devices are increasingly being incorporated into personal electronic devices and automotive application. Some smart phones and personal digital assistants contain multi-mode MEMS sensor devices which integrated functionalities of accelerometers, gyro, e-compass and pressure in one package which demand was grow up annually over 10% from 2011. The complexities of testing input requirements increase the challenges of test cost and data integrity. There are different test methodology and criteria between consumer, industry and automotive application. The consumer application of multi-mode MEMS sensor devices will be driven toward high parallelism package test for cost reduction or BIST on wafer level test to increase the coverage rate and cost reduction.

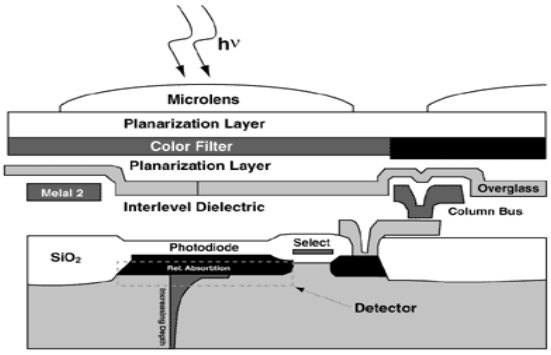



**Image sensor structure cross**

*Figure TST14 - Image Sensor Cell*