HETEROGENEOUS INTEGRATION ROADMAP

2019 Edition

Chapter 17: Test Technology

http://eps.ieee.org/hir

The HIR is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

We acknowledge with gratitude the use of material and figures in this Roadmap that are excerpted from original sources.
Figures & tables should be re-used only with the permission of the original source.

Table of Contents

To download additional chapters, please visit
http://eps.ieee.org/hir

Chapter 17: Test Technology

Executive Summary and Scope

This bi-annual update to the heterogeneous integration testing roadmap contains our best estimates of key trends influencing this industry over the next 15 years. This roadmap includes trends in semiconductor device technologies and their impact on test, as well as roadmaps for key test enablers (Device Handlers, Test Interfaces, and Test Methods). The resulting Cost of Test is also analyzed and discussed.

Key Implications per Device Technology

RF Devices: The ramp of 5G devices is going to challenge the test industry in terms of frequency, port-count, and lower noise margins. Additionally, production test methods for confronting beam-steering are of concern.

High-Speed Digital Devices: Signal delivery through clean traces and a test interface (probe or socket) is becoming a major concern as speeds move past 16 Gbps (today) to more than 60 Gbps during this roadmap timeline. Deployment of multi-level technologies (PAM4) challenge noise margins and instrument designs. The large quantity of high-speed digital interfaces per part is another challenge lacking an obvious solution.

Photonic Devices: The large number of photonic standards together with an ever-increasing number of ports, while geometries continue to shrink, is going to be a significant challenge for the industry to address in a cost-effective manner.

Logic Devices: Growing device complexity is having a big impact on the industry today. This trend is expected to continue with pattern depths and structural test times doubling every three years over the roadmap period. Higher complexity also drives up the device cooling requirements during test.

Specialty Devices: The trend for tighter-pitch displays and cameras is resulting in a need for probing contacts of finer pitch and smaller size. At the same time, higher volumes are driving the need for more test parallelism. The large volume of inexpensive IoT sensor devices of various types will continue to challenge the industry.

Memory Devices: NAND Flash device test times will continue to scale with higher densities. The introduction of high-speed serial memory interfaces may change the test approach. Achieving Known-Good-Die (KGD) and Known-Good-Stack (KGS) for memory devices destined for 2.5D and 3D integrations will continue to be of growing importance.

Analog/Mixed Signal Devices: Technology advances are expected to challenge mixed-signal devices in about five years as speeds and resolutions increase while voltage swings continue to drop.

Key Implications per Device Handling and Contacting

Device Handlers: Wafer probe and component test handling equipment face significant technical challenges in each market segment. Common issues on both platforms include higher parallelism and more capable and flexible thermal solutions, which result in increasing capital equipment and interface cost.

Probes: The probe complexity roadmap continues to accelerate at challenging rates. Pad size and pitch are expected to keep compressing with site counts increasing. With the manufacturing of super-high-speed serial and 5G devices, probe technology is being forced to provide cost-effective 40-80 GHz bandwidth solutions.

Key Implications for Test Methods

System Level Testing (SLT): A resurgence in functional/SLT testing, especially at the module level, is a major industry trend. The system-level boot-up test remains the critical acid test for most electronic systems. Efforts continue to pinpoint the faults responsible for structural test failures. In the meantime, SLT deployments are expected to rise.

Adaptive Testing: This technology continues to ramp in many ways throughout the industry. Expansion will be challenged by the explosion of data, the need for effective analysis, and response tools, which critically need a universal data model/format to enable “Big Data” analysis independent of the source Further challenging this methodology is multi-vendor integrations and data security concerns.

Concurrent Testing: Expanded usage of concurrent test techniques is constrained by limitations in device access (pins) as well as a lack in standardized tools and methods. A breakthrough for this technology will come when efficient and standard test interfaces are implemented for analog and RF testing. Analog scan is one technology starting to show progress.

2.5D/3D Testing: Known-Good-Die (KGD) are very important today and will become even more important in the future (larger parts with more gates, more devices per assembly). Achieving KGD will push test content earlier in the manufacturing flow, challenging wafer- and/or die-level testing. Testing the multi-chip assembly is expected to be done by traditional approaches including boundary-scan and system-level testing.

Reliability Testing: Increased quality and reliability needs of the automotive and mobility integrated circuit markets are driving additional test and burn-in requirements, including system-level test. Packaging innovation, including the need for KGD, drives additional test fixturing and testing challenges, as does multi-die packaging.

Design-For-Testability (DFT): Ongoing use of pattern compression and the use of hierarchical design techniques is continuing to have a positive impact on pattern depths and device test times. The quest for higher compression ratios is proving to be a difficult one.

Yield Learning: Improvements are needed in resolving cell-internal defects. Achieving this goal will generate significant amounts of data requiring careful analysis. Automotive reliability requirements are driving a change of mindset away from structural test coverage being adequate to functional, system level, and in-situ testing techniques becoming the new focus.

Key Implications for Cost of Test

The cost of consumable products in device testing is starting to have a dominant impact on the industry moving forward. The result is many anecdotal comments that moving to a new transistor geometry node has cost more in terms of consumable products (probes, sockets, interface boards) than the cost of the test instrumentation to do the new level of testing. Compounding this problem is the requirement for ongoing maintenance of these components as well as their limited lifetime.

Another major trend having a significant impact on the cost of semiconductor testing is longer test times. Increased device complexity drive up the pattern depths and test times at a rate significantly faster than the rate of complexity increases (exponentially). We no longer talk about test times of a few seconds and often end up with test times of a few minutes. Without a breakthrough in scan testing speeds, this trend is expected to continue for monolithic devices during this roadmap period.

Heterogeneous integration will allow the large devices to be split up into smaller chips with the test time impact now being exponentially reduced. Additionally, multiple smaller devices can usually be tested concurrently for additional savings of test costs.

Many efforts are expected to try and reduce test costs moving forward including higher parallelism, reduced test time overhead, reduced equipment and consumable costs, etc. Despite these efforts, increased complexity and demands for higher performance testing will likely cause the cost of leading-edge test solutions to increase significantly.

Detailed Discussion of the Trends

Each of the topics highlighted above as key implications for test moving forward are explored in precise details in the next sections.

Acknowledgments

This update to the test roadmap is output from a team of more than 120 test experts hailing from more than 50 companies world-wide. The leadership team for this effort is shown below:

Dave Armstrong:	Chairman, Test Technology Working Group, Heterogeneous Integration Roadmap

2.5D/3D device test	Bill Eklow	Photonic Test	Dick Otte
Adaptive test	Phil Nigh	Probe Cards	Dave Armstrong (Acting)
Analog and Mixed-signal	Don Blair	Prober	Yoichi Shimizu
Concurrent Test	Kenichi Anzou	Radio Frequency	Don Blair
Cost of Test	Ken Lanier	Reliability Testing	Carl Buck
DFT & SOC Device Testing	Kenichi Anzou	Specialty Device Testing	Wendy Chen
Handlers	Masahito Kondo	System Level Test	Harry Chen
High Speed IO Testing	Yi Cai	Test and Yield Learning	Anne Gattiker
Logic device test	Marc Hutner	Test Sockets	Rick Marshal
Memory Test	Dave Armstrong (Acting)		

Test Technology Working Group Contributors:

Stacy Ajouri, Lionel Kimerling, Mike Alfano, Craig Kirkpatrick, Kenichi Anzou, Leo Kluger, Takenori Aoki, Masahito Kondo, Jeff Arasmith, Hiroshi Kurosaki, Dave Armstrong, Greg Labonte, Keith Arnold, Ken Lanier, Roger Barth, Lenny Leon, Brady Benware, Ira Leventhal, Keren Bergman, Clark Liu, Paul Berndt, John MacWilliams, Mark Berry, TM Mak, Anil Bhalla, Yiorgos Makris, Don Blair, Adriano Mancosu, Shawn Blanton, Peter Maxwell, Daniel Bock, Fabio Morgana, Bill Bottoms, Franco Motika, Carl Bowan, Nilanjan Mukerjee, Carl Buck, Chris Nigh, Ken Butler, Phil Nigh, Yi Cai, Peter O'Neill, John Caldwell, Richard Otte, John Carulli, Sule Ozev, Krishnendu Chakrabarty, Gilles, Wendy Chen, Jayson Park, Harry Chen, Mike Peng Li, Wu-Tung Cheng, Jan Peters Weem, Vivek Chickermane, Robert Pfahl, Tim Cleary, Robert Polster, Chris Coleman, Chris Portelli, Zoe Conroy, Dieter Rate, Adam Cron, Jeff Rearick, Al Crouch, Graham Reed, Robert Daasch, Herb Reiter, Dirk de Vries, Joseph Reynick, Rao Desineni, Jason Rivers, Makoto Eiki, Mike Rodgers, Bill Eklow, Jeff Roehr, Dan Evans, Pete Rogan, Ira Feldman, Mark Roos, Shawn Fetterolf, Rajiv Roy, Derek Floyd, Sam Salloum, Bert Gab, Ben Sarpong, Michael Garner, Philip Schofield, Anne Gattiker, Michael Schuldenfrei, Xinli Gu, Rene Segers, Richard Hanvey, John Shelley, Ian Harrison, Yoichi Shimizu, Don Hartman, Mustafa Slamani, Kazumi Hatayama, Wesley Smith, Klaus-Dieter Hilliges, Stephen Sunter, Andy Hsu, Aoki Takenori, Tom Hsu, Steve Tilden, Marc Hutner, Srikanth Venkataraman, Shuichi Ito, Li-C Wang, Erik Jan Marinissen, Thomas Williams, Rene Jonker, Scottie Wyatt, Matthias Kamm, Tian Xia, Rohit Kapur, John Yi, Brion Keller, ChuHui Yu

Section 1: RF Test

Key Test Trends

Table 1: RF Test Requirements

Year of Production	2018	2019	2020	2025	2030
Leading Edge (Note 1)					
Mobile Devices (General Radio) (Note 13)					
Carrier Frequency (GHz)	8	8	8	8	8
Number of simultaneously active RF Ports per Die (Note 11)	12	16	16	16	16
Total number of RF Ports per Die (Note 11)	48	64	64	64	64
Modulation RF SSB BW (MHz) (Note 2)	80	80	80	80	80
Amplitude Accuracy (dB) (Note 3)	<0.25	<0.25	<0.25	<0.25	<0.25
ACLR (dB) (Note 4)	77	80	80	80	80
Phase Noise (dBc/Hz @ 100kHz offset @ Fc=1GHz) (Note 12)	-143	-146	-146	-146	-146
Error Vector Magnitude 3G/4G (Note 5)	0.5%	0.5%	0.5%	0.5%	0.5%
IIP3 (dBm) (Note 6)	36	36	36	36	36
OIP3 (dBm) (Note 6)	60	60	60	60	60
Infrastructure/Focused applications (Radar, WiGig, Backhaul)					
Backhaul Carrier Frequency (GHz) (Note 7, 9) 115 is W band.	115	115	115	115	115
Collision Avoidance Radar Carrier Frequency (GHz) (Note 8)	81	81	81	81	81
Modulation RF SSB BW (MHz) (Note 10)	1760	1760	1760	1760	1760
Amplitude Accuracy (dB) (Note 3)	<0.5	<0.5	<0.5	<0.5	<0.5

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

Notes for Table 1:

1. Leading Edge versus Volume: This distinction is to serve two purposes. One purpose is to approximate an adoption cycle from when new technologies emerge and require characterization vs. volume production, while a second is to highlight that certain test methodologies may be required in characterization (typically more complete), while production test may rely on simplified conditions.
2. The leading-edge modulation bandwidth increased significantly from year 2010 to year 2011 to cover proposed wireless standards such as WirelessHD. 1760 MHz driven by WiGig @ 57-66GHz.
3. Amplitude accuracy below 0.5 dB requires de-embedding, which adds significant cost.
4. For WCDMA @ 5MHz offset channel.
5. Error Vector Magnitude (EVM) testing is very prevalent in high volume manufacturing today. The challenge in a table such as this is that each standard (WLAN, WiMAX, W-CDMA, etc) has a different requirement. The values represented here are a blended average of EDGE, and W-CDMA variants (HSDPA, HSUPA, etc).
6. OIP3 / IIP3: These are (best for the chain) figures of merit for distortion performance (O = Output, I = Input). Generally speaking, systems with larger peak to average power excursions require higher linearity. Examples are systems employing code division multiple access (CDMA) technologies and Orthogonal Frequency Division Multiplexing (OFDM) systems.
7. 40/100 GHz Point to point connectivity (WW freq. allocations differ by country)
8. 75-81 GHz for collision avoidance RADAR & Other sensors, IEEE 802.11ad (WW freq. allocations differ by country)
9. 66-67 and 76-86 GHz is for backhaul point to point
10. 1760 MHz is for WiGig
11. Device is defined as a single IC (not chip set or other). Testing can be done with less simultaneously ATE active ports.
12. DUT Requirements of the source
13. Cell Phone 5G base station (standard not finalized) 28 & 40 GHz carrier. Wafer level OK (BW is a challenge and number of ports is high). Package level with antenna is still a challenge
14. 40 GBPS @10 GBPS 4 twisted pairs -- 4 twisted pair (10 GBPS)

Short Term Trends (<5 years)

Mobility

The mobility area of the table is a dynamically evolving topic. LTE and LTE-advance have emerged as the world-wide 4G standard for all mobility devices and UWB was never commercially adopted. WiGig has taken the place of UWB. The market was driven by evolving existing standards to larger bandwidths like WCDMA to LTE-Advance and 802.11a/b/g/n to 802.11ac WiFi to drive faster data rates. The primary drivers of consumer mobility devices are smartphones and tablets that use LTE/LTE-advance and WiFi as predominant standards.

The Mobility portion of table has been reworked to reflect those changes. RF SSB BW remains constant at 80MHz (driven by 802.11ac). The WiGig is in place in the near future, requiring a 1760 MHz modulation as reflected in the tables.

The system sensitivity, driven by better phase noise and less spectral growth from non-linear distortion, drives adoption of higher density modulation standards to carry forward faster data rates. As the number of QAM constellations increase, the EVM accuracy that plots the amplitude and phase error also needs to improve.

The RF port count per device increases to 64 over time because frequency allocations are non-standard across the globe and the need for backwards compatibility to the many existing digital communication standards need to be fulfilled.

The Internet of Things (IoT) is a market that is high-volume cost-driven, more than a test technology driver. It primarily consists of huge numbers of RF sensors that transfer data into a global data distribution center which will then communicate to the outside world and internet. Data processing will be used to identify trends and make the data meaningful to the world. There are two schools of thought for the data transfer for IoT: 5G speed could make this attractive for the massive amounts of data. On the other hand, if cellular phones quickly adopt the 5G standard, then there will be 4G infrastructure capacity available at a cheaper cost for IoT data transfer.

Infrastructure/Automotive RADAR/Industrial

The popularity and acceptance of collision-avoidance radar detection systems and point-to-point backhaul is driving the second area of the table – Infrastructure/Automotive RADAR/Industrial. In the Automotive Radar space, the trend is towards more ports and multiple transmit and receive channels. Similarly with backhaul transceivers (802.11 AY, MU MIMO) the trend is towards more antennae and arrays (4 antennas per array and possibly as many as 64 arrays) as well as frequency increases upwards of 115 GHz.

In the 5G space, it is already out and becoming more prevalent and widely used. It is also driving the need for more RF ports with its antenna numbers ranging from a 2X2 antenna array up to 64. Another possible driver for 5G is that it is less able to penetrate walls and glass, so that could drive the repeater trend to mitigate this.

Difficult Challenges in the Short Term

- 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 at the 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.
- More ports, multiple transmit and receive channels, more antennas and arrays will require more RF ports from ATE.
- In general, for some of the leading-edge frequencies and number of ports required, characterization and debug solutions are known, but will be a challenge in production. For example, the 802.11 AY band requires a 60 GHz carrier1 and W Band carriers are at 115 GHz. The newer standards for MU-MIMO are using 8X8 arrays2.
- There are some challenges specific to a production ATE environment including: Integrating waveguides into ATE for applications >65 GHz (mechanical and cost challenge), Integrating 3rd party components into the test-head or probe card for applications >60 GHz, and blind mate connector BW performance at 40-80 GHz.
- In the 5G space, there are more carrier bands/aggregation and more BW(IF - 2 GHz). Power management may become an issue both in terms of number of domains and cooling dissipation during test.
- 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 are still trending from RF-to-BB and vice-versa. RF-to-Digital or line-to-line is limited but widely discussed.
- 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.
- OEM is conceptualizing RF BIST/loop-back test methodology, but functional and parametric tests still dominate the market.

Medium-Term Trends (6-10 years out)

In the backhaul transceiver space, phase array antennas will have a larger number of RF ports. For example, satellite apps in the 12-40 GHz range with MCMs combining front end integration: ASICs, ADC and DACs.

Other examples include backhaul: transceivers with 3 ports (RX, LO, TX) per band combined with multi-band for a total of as many as 9-11 ports; and 5G with a minimum 4 ports (2X2 antenna) and maximum of 64 ports for an 8X8 antenna.

With this increase in ports and IP blocks, there will be an increase of power needs in terms of the number of domains and wattage (heat).

Difficult Challenges in the Medium Term

- Test for phase-array transceivers will require cooling mechanism(s) during test.
- There will be a need to integrate temperature sensors into the handler, device (on die) and wafer probe. The higher the power, the faster the rate of change of temperature.
- There will be a requirement to integrate cooling into the handler and wafer chuck.
- COT challenges will be exacerbated for these solutions for cooling and number of RF ports.

LONG Term Trends (10+ years out)

- 6G future possible improvements include:
  - Integrate terrestrial wireless with satellite systems.
  - Ultra-dense cell networks, millimeter waves for user access, enhanced optical-wireless interface.
- Difficult Challenges in the Long Term
  - The definition of 6G needs to firm up for the challenges to be known.

Trends impacting the roadmap 2017-19

The model presented in this roadmap was generated in 2017 and there are areas that need to be looked into further for the next update.

- 5G both in the base station and appliances have been gaining significant momentum and will need to be integrated into the tables with more details about frequency bands, different modes, channel bandwidth, EVM requirements and number of channels.
- An additional 5G topic, but one that deserves its own space in the tables as well is the 5G antennae which are required for beam forming. Here the entries would include the frequency bands, array sizes, looking at phase shifts between the antenna elements in the array, and the number of RF ports required to test the array along with the other elements, such as a transceiver(s)

Summary

The number of RF ports increases significantly in categories such as backhaul transceivers and multi-band 5G. With the increase in ports also comes higher power consumption, leading to higher heat generation. The Mobility market continues to have higher cost pressures than the Infrastructure/Automotive RADAR/Industrial. ATE needs to follow these trends with more RF ports, more power options and ways to deal with heat generation, while maintaining or even decreasing the ATE costs.

References

https://www.keysight.com/upload/cmc_upload/All/20170608-A3-BrianSu.pdf
https://www.eetindia.co.in/news/article/new-wi-fi-standard-extends-mu-mimo-to-8x8-arrays
https://www.linkedin.com/pulse/beyond-5g-roadmap-6g-stephen-patrick/

Section 2: Test of Photonic Devices

Executive Summary

This section addresses overall lifetime test issues resulting from the inclusion of photonic capabilities into devices and products. The emphasis is on silicon wafers and die with photonic functionality, and assemblies and products that include these devices. Systems in Package (SiP) assemblies and systems are addressed to the extent viable given the diversity of test needs that are specific to applications. The test issues for wafers, die, SiPs and systems will be addressed at the Design, Qualification, Validation, Production and In-Use stages of life cycles. Current and anticipated optical parameters to be tested and their value or level are considered along with the test access issue at each stage of the product life cycle.

Telecommunications test equipment, components and methods were and are being adopted for optical testing of products used for non-long-haul applications. The traditional methods are being extended and new methods developed to address test needs for photonic wafers, photonic integrated circuits, System in Package (SiP) that utilize optics, and complete systems. Utilizing these extended methods requires optical probing of both wafers and die combined with electrical probing, resulting in a series of mechanical issues. The inclusion of optical probing, especially single mode probing, requires gratings or other access points on wafers. For individual die, dual mode (electrical and optical) probing is especially difficult due to the small size of die and difficulty of holding and locating probes accurately. At the SiP level, the problems are easier because the device is larger, not as fragile, and is often designed to facilitate dual-media probing. The wafer, die and SiP probe fixtures tend to be expensive due to the complexity and accuracy required. System-level test access is usually easier because at that level, electrical interfaces and optical connectors are included as part of the DUT.

In addition to probe access, optical test methods to simultaneous characterize and compare multiple optical lanes and/or ports at the same time are needed. One need is comparative simultaneous testing of multiple signals from arrays of ribbon fibers, waveguides or chip sources or detectors for optical skew, jitter, etc. A related need is simultaneously evaluating optical signals multiplexed on one fiber or waveguide. Applications with arrays of up to 256 ports (fibers, waveguides chips) or ~256 multiplexed wavelengths are forecast in the next 10 years.

In addition to the standard telecom optical parameters (power, wavelength, attenuation, jitter, SNR, etc.), emerging applications utilize virtually every parameter that light can have, potentially requiring the extension of test capability in multiple dimensions such as polarization, phase noise, spatial modes, multiple fiber cores, etc. While these emerging needs are potentially very broad, the near-term emerging needs seem most likely to be extensions of data communications needs.

Optical communication applications are likely to utilize 650 nm to 1700 nm wavelengths, multiplexed wavelength spacing down of 25 GHz, detector efficiency of ~1Amp/Watt, receiver sensitivity as great as -45 dBm, power levels of 1 watt or less, symbol rates of 100 Gbaud per lane, modulation schemes utilizing up to 10 bits per symbol, polarization multiplexing, BERs of 10-12, etc. Over time these parameters will improve, so test capabilities will need to stay ahead of them. Data rates as high as 500 Tbps per fiber are likely to emerge in the next 10 to 15 years.

Design for test by including optical test access points, Built In Self Test (BIST), redundancy for self repair and prognostics to report changes and deterioration during operation over the life cycle of optical products are desirable and of value in an increasing number of applications. These should be considered for inclusion not only in designs but in software design tools as well.

This section on photonics test contains our best estimates of key trends influencing this industry over the next 15 years. This roadmap includes trends in semiconductor device technologies and their impact on test, as well as roadmaps for key test enablers (Device Handlers, Test Interfaces, and Test Methods). The resulting Cost of Test is also analyzed and discussed.

INTRODUCTION

This section focuses on unique attributes of testing optical devices. No attempt has been made to duplicate required and typical electrical or mechanical testing.

The chapter is open ended on optical applications testing with much of the material broadly applicable. It does, however, concentrate primarily on testing data communications products.

The section addresses photonic testing for:

a. photonic integrated circuits (PICs) on wafer
b. individual PIC die
c. photonic System in Package (SiP) devices
d. system level optical functionality, such as a complete transceiver or Active Optical Cable (AOC).

Another dimension addressed is testing needs over a product life cycle:

a. during development to prove functionality and de-bug devices
b. qualification testing
c. pre-production validation
d. in-process production testing to assure product quality and improve yield.

Links vs Channels

An optical link utilizes one wavelength traveling from one point to another in one fiber or waveguide. Information may be imposed on the beam utilizing any methods such as On-Off-Keying, PAM XX, dual polarization or any method that affects only that wavelength.

A Channel may consist of a single lane but often, even usually, has multiple lanes. The lanes may be on multiple parallel fibers or waveguides, or on the same fiber or waveguide utilizing wavelength division multiplexing. Many datacom standards utilize multiple lanes to achieve their data rates.

Evaluating technical capability is most easily done utilizing lanes because channels that combine many lanes make it difficult to understand the underlying technology. System designers, however, find the channel view more useful as the technology details are not important at their level.

SITUATION (INFRASTRUCTURE) ANALYSIS

Generic Photonic Device Testing

Figure 1 illustrates the general test requirement; the need for both electrical and optical test inputs, and then analysis of the electrical and optical outputs from the device under test (DUT). In addition, environmental parameters, such as temperature, humidity, vibration, etc. may be test inputs. Finally, in addition to the optical and electrical test responses, physical factors such as temperature rise may be outputs that are monitored during testing.

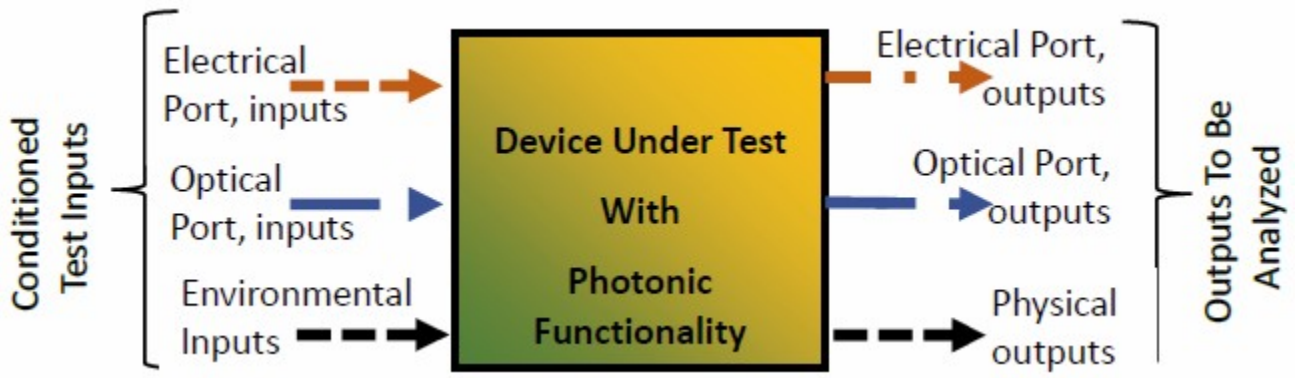


Figure 1. Generic Photonic Product Test Environment

The electrical ports are electrical contacts, or arrays of contacts, for power, control, monitoring of functionality and of course data inputs and/or outputs.

Optical input ports may provide optical beam/s to be modulated or data streams, often in arrays, to be analyzed. The optical "connections" may be a butt coupling, an air gap with a beam bridging into the device or an evanescent coupling resulting from proximity of waveguides. Making these test connections, especially the optical connections, is frequently a major project.

The photonic input and output signals may have multiple parameters. Specifically, optical signals may have:

- Intensity
- Polarization

- Direction
- Mode profile
- Wavelength
- Variation over time:
  - modulation (fast)
  - drift (slow)
- Skew between beams
- Signal-to-Noise Ratio (SNR, RIN, Crosstalk)
- Bit error rate (BER)

One or more of these parameters must be imposed on a suitable optical beam or beams and injected into devices on the optical ports to provide a drive signal. The device under test will perform a function determined not only by the input optical signals, but by the electrical inputs as well. The resulting electrical and optical signals are on the corresponding output ports.

Environmental inputs may include all of the usual variables: temperature, humidity, temperature cycling, Highly Accelerated Stress Test (HAST), vibration, shock, etc.

Physical outputs include temperature rise, mechanical changes such as delamination, cracking, swelling, wire breaks or optical chain interruptions, etc.

Photonic test requirements vary by the test level (wafer, die, photonic SiP, system) and test need (access, sources, detectors, functions). Table 1, Photonic Test Requirements, gives a generic view of the testing needs for items containing photonic elements at the various levels.

A similar table can be developed for specific applications to provide some insight into the related requirements for each application.

Several types of product testing of devices, including those with photonic capability, are usually required:

- Test during development to ensure the design “works”.
- Qualification testing, typically done before a product is committed to wide use.
- In-process testing to monitor manufacturing process quality.
- Final testing before each individual product is shipped to a customer.

Every application, including photonic products, has a specific set of these tests that is applicable to that product. Data communication is an application of immediate interest, one of the most important at the moment, and an application about which much is known; hence we will concentrate on it.

Table 1. Photonic Test Requirements

	Test Need			
Test Level	**Optical Access**	**Sources**	**Detectors**	**Functions**
Wafer	45o mirrors, vertical grating couplers, cleaved fiber, tapered fiber, lensed fiber, focused free-space beam, evanescent coupling	External sources injected via fiber or free-space access Integrated sources	External photo-detectors, potentially in arrays; Imaging sensors (eg. CCD/CMOS FPAs); Optics to collect and/or image light to be detected. Integrated photodetectors	Wide variety of device characterization and functional tests; media loss/cm, insertion loss, modulation depth/ bandwidth, polarization control, wafer uniformity, detector sensitivity/responsivity, temperature sensitivity, die-to-die variation, skew between outputs
Chip	Wafer options plus edge coupling to embedded or surface waveguides.	Wafer options	Wafer options	Wafer options plus edge coupling impacts on loss, spectral bandwidth and polarization
Photonic SiP	Butt coupling or expanded beam connector, evanescent coupling, fiber splice, ribbon fiber splice	External or on-chip laser source to simulate application related requirements.	External detector or detectors, potentially in an array, gathering light from an edge emitting waveguide or vertical emitting 45^{o} mirror, or vertical emitting grating	Wafer and chip options, plus characterize package connections, and application specific tests such as eye diagrams, BER, environmental sensitivity
System	Conventional optical connector, fiber splice	External or internal laser source or sources to simulate inputs.	As needed to measure and evaluate system outputs.	Intensity, skew between lanes, polarization, eye diagram, SNR
In Use, Over Lifetime	Limited, if any. Primarily wireless or electronic	Both self and remotely initiated data reporting	Primarily wireless or electronic	Monitor & report performance changes. Initiate self-repair.

Data Communications Device Testing

Some specific points and issues important in testing data communications photonic products, especially transceivers, are these:

i) Test time is being increased by IEEE Standards that "stretch" the required reach. That reduces SNR and rapidly raises the BER. Thus, "wisely" managing the required reach reduces test cost.

ii) For some data communication applications, a simple, "worst case" Eye Diagram Test is sufficient. This point illustrates that choosing the right criteria that properly balances the test need and potential faults can reduce test cost.

iii) Some Optical Components, especially lasers, are nonlinear, so testing is harder to do and more demanding, but also more important.

The need for specific, high-tolerance physical location of optical components during testing makes changing optical test configurations difficult and time consuming. Gradually reducing input signal strength and measuring the decrease in performance can sometimes provide a way to determine margin and robustness of Optical Systems. This approach, of course, is widely used as one criteria/methodology for electronic products.

Developing new test software, fixtures, sources, detectors, etc. often takes a long time and depends on the materials and devices to be tested. Optical engineers are innovative and continually developing new design and test methods. Fortunately, low volume test capability is improving to support researchers. (These capabilities may eventually impact high volume needs.)

Test Equipment

Optical device test equipment is available from multiple suppliers. Historically, the telecommunications industry was the major consumer, but in recent years the use of optical communications for short distances, such as LANS, FTTX, and AOC and in Data Centers, as well as a variety of sensors, has broadened the demand. Much of the demand emerging for these new optical applications is filled by utilizing equipment developed for and derived from that used by the Telecom sector. As these applications grow in importance, specialized equipment is emerging and becoming available.

Test Processes

Bit Error Rate (BER) Testing is the most time-consuming and therefore more expensive than other testing.

Eye diagram evaluation is of great value and one of the most common test methods for data communications devices.

A related test methodology is "constellation measurement" used for characterizing complex modulation schemes such as QAMXX. Several versions of this method are in use depending on the modulation scheme to be evaluated.

Test Access, Fixtures and Methods

PIC Layout for Test:

- Layout photonic chips and substrates to facilitate testing and packaging by enabling optical access for simultaneous multi-channel optical probes as well as DC, RF and microwave electrical probes.
- Route all optical I/O to the same device edge, organized in a standard linear array; e. g., 127 um or 250 um pitch. All optical ports can be accessed simultaneously by alignment of a single fiber array probe.
- DC and RF pads should similarly be routed to a single, although different, device edge to allow simultaneous probing with multi-conductor probe cards. Ideally only four independent probe heads need to be aligned, to the North, South, East and West edges of the device. Anticipate interference of probe structure mounts when designing compact devices.
- Include on-wafer/chip test structures for calibrating probes.
- Include on-wafer/chip test devices for characterizing individual components, in isolation from a complex, integrated system.
- Incorporate Built In Self Test (BIST) whenever possible.
- Enable debug of faulty circuits by designing test points (grating couplers or photodiodes) to tap off signals. Might make use of the CLIPP (Contactless Integrated Photonic Probe).

Wafer DUT Interface:

- Optical probes must interface by surface grating couplers which impose limits of narrower spectral bandwidth and polarization dependency. Alternative vertical coupling technologies could include etched turn mirrors. Possible for grating couplers to be diced off during singulation, so the final chip is edge-coupled.
- Optical probes can be: single-channel, multi-channel, SMF, PMF, flat facet (cleave, UPC), APC.
- facet, lensed, tapered.
- Optical probe alignment should ideally have 6 degrees of freedom, with 0.1 micron translational precision, and 0.3 arcminute rotational precision.[1]
- Test on an opto-electronic probe station (manual or automated) with a temperature-controlled wafer chuck. A commercially available system from PI is shown in figure 2. A video of it in action is at the following link: https://www.youtube.com/embed/_TG3lUu-k0k?rel=0

[1] The 0.3 arc minute requirement is pointing accuracy required for free space beam such as encountered in Lidar and other free space applications. The requirement in most datacom applications is significantly less, such as 1° which is required for simultaneous alignment of multiple fiber probes in a linear array.

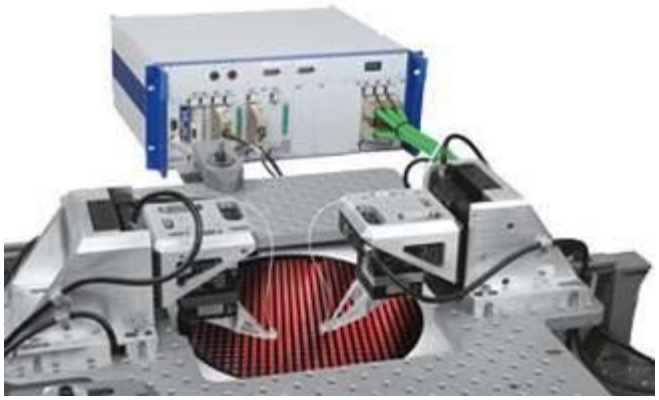


Figure 2. A PI Wafer Probe Station for photonic integrated circuits.

Chip DUT Interface:

Figure 3. PI Probe Station Concept for Individual PICs.

Probing individual chips as suggested in Figure 3 and earlier in Table 1 is viable under some conditions but is generally to be avoided, with wafer-level testing preferred. Individual PIC testing has not only the usual optical and electrical probe interface issues but the added issues of handling and aligning an individual die. Handling is particularly difficult for thin die, meaning 100 microns, that are commonly used.

Whenever necessary, however, test capability for chips is available commercially, even for PICs.

Packaged device DUT interface:

Testing during development on a lab bench: Temperature control is possible through the use of a thermal-stream forced air system. Connections are made through standard connectors (optical and electrical). No probes are required.

One version of a test fixture suitable for products with both electrical and photonic inputs and outputs is illustrated below in Figure 4. This is a highly custom fixture designed specifically for the device to be tested. The configuration is viable because the device configuration, meaning locations of optical and electrical access points, is chosen with the fixture needed to make connection in mind. The electrical connections through pogo pins is well known but does put stress on the DUT, causing it to deform to varying degrees depending on dimensions, temperature, etc. These physical changes make interfacing the optical ports more complex as they must move not only laterally, but somehow be keyed to align in the other 2 dimensions as well as 2 angular dimensions. This might be accomplished with pins, for example, commonly used with the US Conec optical connectors. That, of course, will make the fixture more complex than suggested by Figure 4.

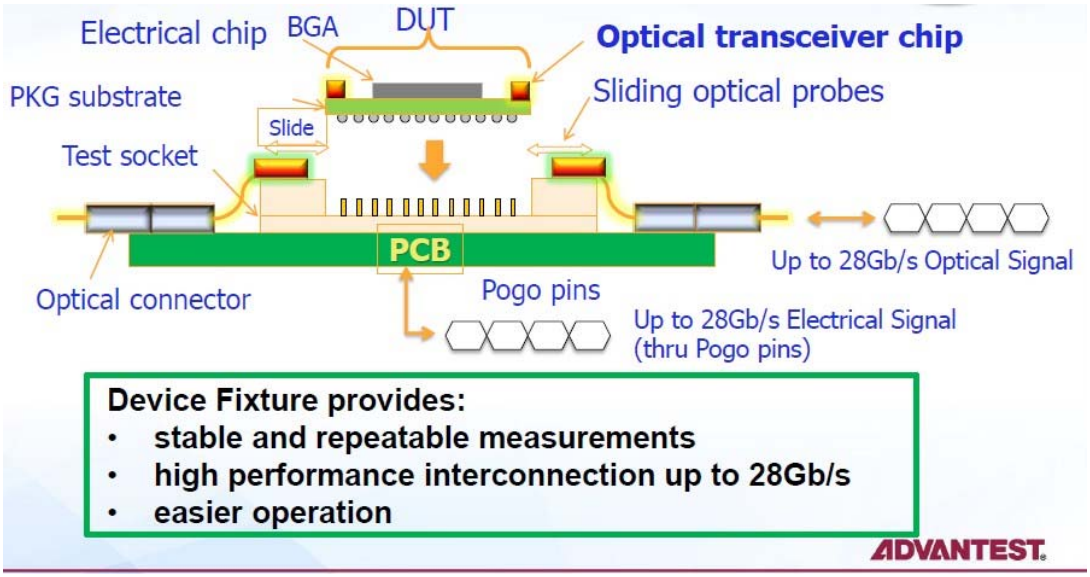


Figure 4. A Test Fixture with both Optical and Electrical probes.

Design Tools and Design for Test

Design tools for PICs are starting to emerge. Including consideration of test will contribute to minimizing the cost and difficulty of optical element test and raise the reliability of the end devices. PICs with many optical elements and/or optical functions may benefit from on/in-chip optical ports to inject optical test signals or ports to probe optical signals. Optical IO test ports can be included as gratings or splitters that allow injection or tapping of signals. These could be in addition to vertical and horizontal ports that communicate on-to and off-of chip for regular IO functions.

Test Standards

The most complete set of optical data communications standards are published by Telcordia. The standards for Optical Fiber and Equipment can be found at: http://telecom-info.telcordia.com/site-cgi/ido/docs2.pl ?ID=187033671&page=docs_doc_center. These standards related to telecommunications, of course, and are applicable to system-level test requirements and to high reliability systems. Less demanding standards are needed for primarily commercial and industrial applications.

Standard Physical Platforms, meaning photonic wafer fabrication methods and a related process design kit (PDK) that Applications can be built on, have not emerged. Fortunately developing and making available such a platform is one of the objectives of AIM-IP. The photonic industry will be able to support a minimal number of innovations, so the sooner standardized packaging emerges, the better.

The PDK that AIM-IP is offering will limit innovation, so it is important that AIM-IP choose its PDK wisely and that the PDK evolves in response to user needs and demands. The AIM MPW capability will be important to Universities and Innovation.

The kinds of testing required vary over the life cycle of a product. Figure 5 below lists typical optical device test activities and requirements over the life of a device from conception through the in-use and end of life phases.

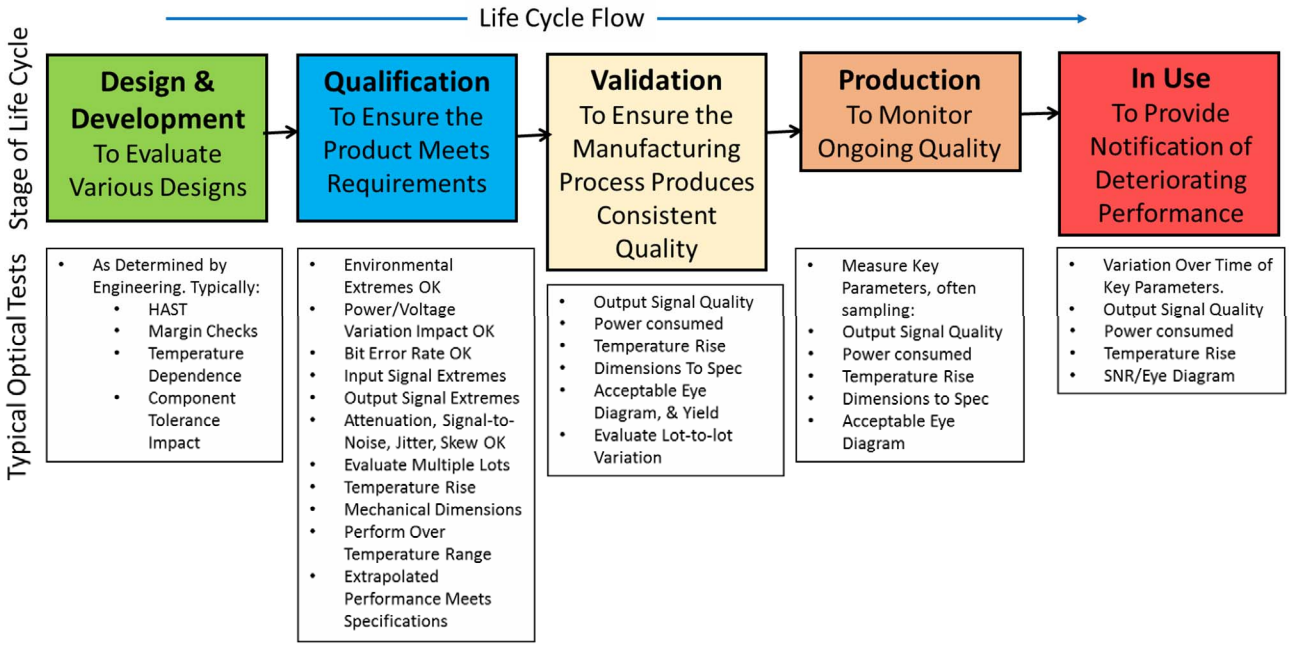


Figure 5. Test Needs over An Optical Product Life Cycle

Environmental Technology

Optical Test has minimal environmental issues other than a few safety requirements associated with laser intensity or UV hazards.

Separately, the environment in which optical devices must perform, and the related proof of performance test evaluations they must be subjected to, have a wide range. A very demanding requirement is the classic Telecordia telecommunications standards to ensure 40-year life. These are still required for some products, such as sub-sea amplifiers. At the other extreme are consumer disposable that have minimal requirements. Between these extremes are a variety of environments in which development and qualification tests must ensure performance.

Generic Test, Inspection and Measurement (TIM)

Types of measurements, O/O, O/E, E/O, E/E: Wafer characterization:

- Wafer mapping: inking for good die identification, yield calculations, process improvement.
- Wafer metrology: layer thicknesses, surface roughness, haze, feature definition, step heights, sheet resistance, doping profiles, mature industry.

Device characterization:

Waveguides: Modal structure, group velocity, dispersion, loss, power handling, polarization dependence, Rayleigh scatter, nonlinear limits (i.e., characterize second and third order susceptibility coefficients, SHG, SFG, DFG, Raman, Brillouin, SPM, XPM, FWM)

Photodiode: Responsivity vs wavelength, polarization dependent loss, dark IV Curves (extract dark current, diode ideality factor, series resistance), bandwidth vs bias, linearity, power handling/compression, over-temperature sensitivity, input capacitance and resistance

Laser diode: LIV curves (extract threshold, slope efficiency, power saturation), RIN, SMSR, center wavelength, linewidth, frequency noise, ASE, tuning coefficients, tuning range/rate, direct modulation bandwidth, over-temperature sensitivity, tolerance to optical feedback

SOAs: spectral gain, efficiency, bandwidth, saturation power, noise figure, amplified spontaneous emission, input power dynamic range

Modulators: extinction ratio vs input signal swing, input capacitance and resistance, electrical to optical bandwidth, spectral bandwidth, insertion loss, polarization dependent loss, resonance frequency and free spectral range and reflected power if applicable, dependence of extinction and optical bandwidth on biasing point

Splitters/Combiners: split ratio, insertion loss, wavelength dependence, polarization dependence, reflected power

Filters: band pass/band reject, attenuation/loss, center wavelength, bandwidth, 'Q' and free spectral range and reflected power if applicable, dependence of extinction and optical bandwidth on biasing point, polarization dependence, tunability: speed, efficiency, performance impact

Attenuators: center wavelength, bandwidth, attenuation, polarization dependence, reflected power vs wavelength and polarization

Polarizers: degree of rejection, insertion loss, reflected power vs wavelength and polarization

Fiber Couplers: optical bandwidth, insertion loss, polarization dependent loss, reflected power

Switches: crosstalk, extinction ratio vs input signal swing, input capacitance and resistance, electrical to optical bandwidth, optical bandwidth, insertion loss, polarization dependent loss, resonance frequency and free spectral range and reflected power if applicable

Optical Connector Characterization:

The incorporation of optical connectors to build systems is increasingly important as a means to eliminate fiber pigtails to reduce handling and the size of systems. Connectors, however, introduce another variable that must be controlled and measured/tested.

Specific component and system characterizations to be performed related to connectors include:

- Connector loss.
- Wavelength dependence of connector loss.
- Connector return loss.
- Connector polarization-dependent loss.
- Connector re-mating loss variation.
- Dust contamination induced connector loss (test TBD).
- Telcordia GR-1435 Uncontrolled Environment Thermal Aging, Humidity Aging, Thermal Cycling, and Humidity/Condensation Cycling testing.
- Signal Bit Error Rate vs. connector number and loss (25 Gbps/channel).
- Estimated system implementation cost.

Functional tests

The most demanding test requirements are found in single mode applications so those are addressed below. Multimode signal testing is usually less demanding.

The Telecom Industry utilizes single mode technology over hundreds of kilometers and has led the development of optical test equipment and capability. Much of that equipment can be adopted for use with products being developed for the emerging needs for shorter distances.

Data communications test needs differ from Telecom in that they tend to utilize more parallel signal transmission through parallel media, either ribbon fiber or waveguide arrays, transmit light shorter distances are impaired by modal dispersion and sometimes utilize more complex modulation schemes.

The general optical signal technical properties and test parameters follow in Table 2 below.

Table 2. Optical Test Parameters, Values, Media and Ranges		
Parameter	**Range**	**Comment**
Optical Signal Characteristics		
Wavelength	750 to 1,650 nm	These are the primary wavelengths used for optical communications. Longer, and sometimes shorter, wavelengths are used in sensors and analytic applications.
Optical power	<1 watt (30 dBm). usually < 0.1 watt (20 dBm)	This value applies to most communications, sensor and analytic applications. Much higher power levels are used for industrial processes. Laser safety must be considered.
Wavelength spacing	Down to 25 GHz or ~0.2 nm at 1.5 microns	Applies in dense wavelength division communications multiplexing (DWDM) applications. More demanding in some sensors.
Optical Modulation Rate	~50 GHz near term, 100 GHz long term	This is the single lane modulation rate.
Laser Sources	40 Gbps/link and higher	Reliable laser sources for 40 Gbps/link and higher rates utilizing higher order modulation are needed.
Optical Amplitude Modulation	Up to 64 levels (6 bit) per single phase near term, 1024 levels (10 bit) long term	For QAM modulation, other formats will also be supported, such as OOK, PAM and PSK
Polarizations	2	X and Y for SMF. More complex SDM being explored for FMF (few-moded fiber).
Detectors	Responsivity	~1Amp/Watt (These values assume conventional photodiodes. Avalanche photodiodes provide higher sensitivity but introduce added noise.)
Detector bandwidth	~50GHz near term, 100 GHz long term	Waveguided photodetectors, higher BWs may require UTC structures
Probing		
# of simultaneous optical test signals needed	1 to 12 near term, up to 200 long term	Some number of optical test signals may need to be injected/received simultaneously using ribbon fiber or parallel optical waveguides with a combination of the following characteristics; one or more wavelengths modulated with controlled polarization, phase and/or amplitude with known and controlled skew between fibers.
Physical connections; Input of test signals and output of device signals	1. Conventional optical fiber connectors 2. Specialized for- test-only gratings built into substrates and products 3. Focused beams 4. Spliced fibers	A variety of probes (methods to get light into and out of optical ports, such as fibers, waveguides or elements such as lenses, mirrors, etc.) are likely to be required. For SM applications, alignment of the probes to the DUT (device under test) of < 0.1 microns will be required. MM applications require <5 micron alignment. Cleaning and inspection are required for each connector end contact face before mating with another connector to perform a test.
Test Receivers	Up to -80 dBm sensitivity, 650 nm to 1,700 nm, up to 50 GHz BW	Need to measure power level, wavelength, polarization, latency and eye diagrams with up to 1024 signal levels (32 x 32 constellation). Also phase and skew between parallel signals.
Bit Error Rate (BER)	< 10-9 to < 10-12	BER is highly dependent on signal-to-noise ratio, signal conditioning, the application and the degree of error correction coding used, if any.
Optical Communication Signal Media Properties		
Single Mode Fiber	Typically 6 micron diameter high index glass core, step-index 125 micron diameter lower index outer glass cladding, overall diameter of 250 microns with 125 micron polymer buffer.	
MultiMode Fiber	50 to 62.5 micron diameter high index glass core, graded-index 125 micron diameter lower index outer glass cladding, overall diameter of 250 microns with 125 micron polymer buffer.	
MultiCore Fiber	Recently developed for SM applications. Initially 7 SM cores in a 125 micron diameter with other combinations under consideration.	
Ribbon Fiber	Either SM or MM fibers built as a linear array, usually on 250 micron centers.	
Waveguides	Single mode from 0.2 microns to 6 microns, with strip or rib geometry. Both SM and MM waveguides are built in silicon, InP, glass and polymers. Waveguides typically have higher loss than fiber.	

Education and Training

Test is essentially a manufacturing activity and thus requires education and training in a series of disciplines and skills. Tables 3 and 4 below provide some guidance on these needs.

Table 3. Academic Education Requirements

Knowledge Required	Content
AC/and DC electricity & electronics	Voltage, current, frequency, power, electronics, transformers, capacitors, inductors, transistors, ions, conductors, semiconductors, non-conductors electrical to optical and optical to electronic conversion.
Basics of Optics	Ray tracing, lenses, mirrors, prisms, wavelength, phase, polarization, intensity, beam divergence, beam focus, optical modes, E and H fields as related to the Poynting Vector, light in fibers, both single and multimode, etc.
Characteristics of Signals	Power, transmitting information, signal to noise ratio, modulation methods including OOK, orthogonal signals, multiplexing, demultiplexing, Shannon Limit, etc.
Basics of Statistics	Gathering data, maintaining integrity, managing data bases, standard deviation, mean, median, Parato charts, statistical process control, control limits, Cp, Cpk, etc.
Measurements	Basics of mechanical, electrical, optical metrology. Repeatability, gage studies, etc.
Financial basics	Basic business financial concepts; revenue, costs, elements of cost, product cost elements, overhead, cash, AR, AP, depreciation, equity, etc. "The $ in must be greater than the $ out". "We make investments in order to make more money back utilizing the result of the investment," etc.

Table 4. Training Requirements

Skill Required	Areas of Training*
Personal Behavior	Show up on time. Be prepared to perform your job. (Be present mentally and not preoccupied with a non-job related issue, rested, healthy, properly dressed, etc.)
Safety	Rules, behavior, precautions, etc., related to safety for machinery, chemicals, slips and falls, people related, spills, MSDSs.
Quality	Follow the rules. Ensure procedures are followed. Go beyond the formal requirements and propose improvements. Follow the Japanese "5S" rules. Follow "Deming's 14 Rules For Management". Use statistics to improve yield and minimize variation.
Cost	Why cost is important, sources of cost, minimizing cost, proposing cost reductions, minimizing waste, maximizing reuse and recycling.
Equipment operation	Safe operation, instrument setup, calibration, standard operations, maintaining records, impact of each process on cost, use of the operating manual, machine maintenance.
Metrology	Use of calipers, electronic and optical measurement methods, storage of data and analysis, ensuring accuracy.
Interpreting Instructions	Read what it says, ask question, make sure you understand, do not "assume", eliminate and resolve ambiguities,
Completing Jobs On Time	Ensure you understand what is required; ensure all of the instructions, materials equipment and other resources are available. Start as soon as possible. Look for potential barriers ahead and ensue they are eliminated. Be prepared to revise your approach. Ask for help. When you error, admit you made a mistake, learn from it, ensure you do not make it again. Do not hide your errors.
*While training is often highly specific to each job, basics apply to all jobs.	

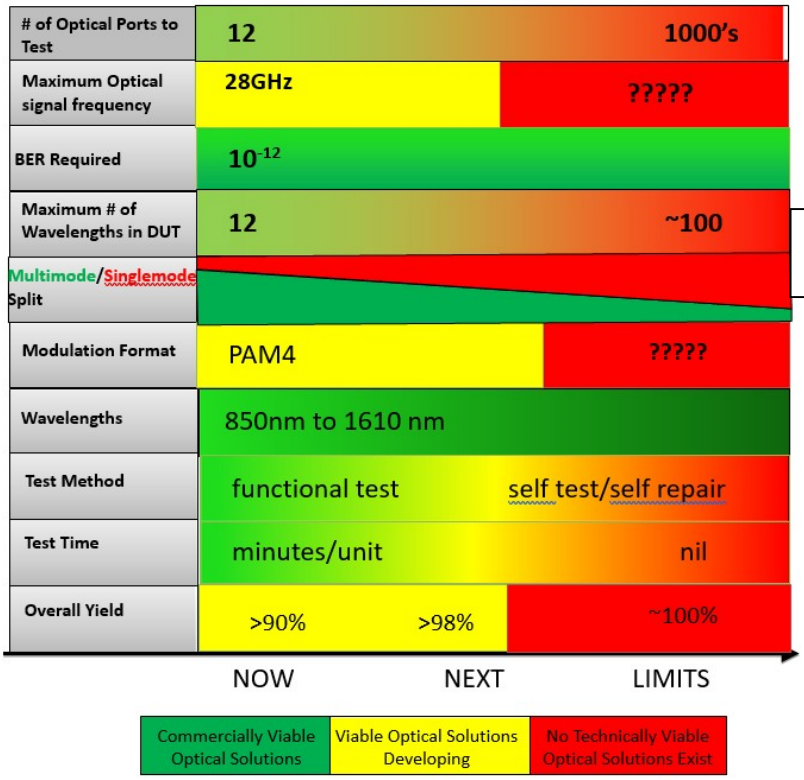


Figure 6: Test, Key Attributes

TECHNOLOGY NEEDS

Prioritized Research Needs (< 5 year results)

- Processing ever faster (100Gbps+) data streams.
- Test time is often determined and limited by memory IO data rates, so increasing these will remove a barrier to lower cost.
- Developing test equipment with more capability than the devices to be tested is a continually moving target!
- Flexible Test Platform, compatible with the test needs of different applications.
- Ability to test photonic properties of wafers during fab to ensure wafers are good.

Prioritized Development & Implementation Needs (> 5 year result)

Eventually, the ability to support 500 Tbps/fiber data transfer rates is going to be needed. An important issue is the nature of the data stream; how much parallelism, what modulation format, etc.

Gaps & Showstoppers

Table 5. Gaps and Showstoppers
The 50 GHz barrier resulting from conventional CMOS capability forcing parallel solutions rather than higher baud rates.
Low speed of suitable assembly, test and other process equipment resulting in high costs.
Inability to overcome the cost driving, rate limiting step/bottle neck of manufacturing/testing such as the number of assembly steps or length of time to perform test, especially BER testing. "Time is money"
Limits resulting from adapting existing equipment, materials and methods to optical test because more specific equipment is not available because the demand is not sufficient to incentivize equipment manufacturers to make it available.
Designing for Manufacturing and test: • Maximizing output to reduce cost • Studying designs to trade off accuracy and speed
Inability to utilize materials or processes due to environmental related constraints (RoHS, REACH, WEEE, etc.)

RECOMMENDATIONS ON POTENTIAL ALTERNATIVE TECHNOLOGIES

Table 6. Recommendations for Potential Alternative Technologies
Utilize laser processing to make optical waveguides in-situ to effective optical connections and optical structures.
Utilization of plasmons to minimize size and maximize functionality

Table 7. Types of Instruments Used for Optical Device Testing	
Optical Vector Network Analyzer (OVNA)	Optical Wavemeter
Lightwave Communication Analyzer (LCA)	RIN measurement system
Electrical Vector Network Analyzer (VNA)	Frequency noise test system
Optical Backscatter Reflectometer (OBR)	Polarization controllers/analyzers
Swept tunable laser source (TLS)	Optical attenuators and amplifiers
Fixed laser sources	Real-time Oscilloscope
Power meters – optical and RF/microwave	Digital Communication Analyzer (DCA)
Fast photo-receivers	Bit-Error Rate Tester (BERT)
Optical Modulation Analyzer (OMA)	IR camera to look at mode profiles, and scattered light
Optical Modulation Generator	Ellipsometer
Optical Spectrum Analyzer (OSA)	White light surface profilometer
Electrical Vector Spectrum/Signal Analyzer	Optical microscope; SEM
Electrical Vector Signal Generator	AFM

Section 3: Logic Device Testing

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 the "consumer SoC" devices are chosen as the primary reference because the most trend data are available for them. Specific test requirements for embedded memory (such as cache), I/O, mixed-signal, or RF are addressed in their respective chapters and must also be comprehended when considering complex logic devices that contain these technologies.

Key Logic Device Testing Trends

The trends in Table 1 are extracted from other parts of the roadmap and are reproduced here to form the foundation of key assumptions used to forecast future logic testing requirements. The first two line-items in the table show the trends of functions per chip (number of transistors) and chip size at production. Chip size in terms of area 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 – the diminishing returns of clock frequency increases. There is a trend to greatly increase the number of cores within each process generation and these will include multiple cores of a particular instruction set, but also include other types of cores, such as graphics units (GPUs), Specialized I/O units (e.g. USB) and various other cores not necessarily specific to a microprocessor.

The internal scan data rate of a device is the rate at which ATPG data can be shifted across the scan latches between test sequences. These are commonly known as load/unload sequences. In the ITRS roadmap this was assumed to be increasing by a percentage increase per year (15%). There are several reasons why this trend has a fundamental limit. The toggle rate for scan, which translates to power utilization of the device under test, is much higher than for functional operation of the device. The device would consume significantly higher power during test, which translates to power degradation of supply levels, higher temperature of operation and a shift in operating point of the logic elements under test. As devices move to finer-pitch geometries, the voltage levels required are also reducing, which further impacts the effects mentioned. In addition to power demands, the timing closure of scan paths is also more difficult at higher frequencies, resulting in a significant amount of work to ensure high speed operation. As a result, the HIR roadmap projects that as devices reach frequencies of 125-150MHz for internal scan rate, the rates will no longer increase compared to historical rates.

The effective scan compression rate will continue to increase, but the increase is accounted for in a different way. The Logic Assumption sheet reflects the effectiveness of compression of a single IP block or subsystem. Effectiveness of compression at this level of the design is slowing. Several years ago, the concept of hierarchal scan was introduced by the EDA community. This is accounted for in the Logic Test Data Volume requirement, Table 2, and realized with the number of identical IP blocks, where the same patterns could be applied at the same time.

In recent years, the fault models used for logic test have also improved. The shift was from a fault model based on a Boolean equation to one based on transistor configuration. This has been accounted for in the models by starting the projections from a transistor count rather than equivalent gates/functions. Presently the increased pattern count has been counterbalanced by improvements in compression and scan-rate increase. As new fault models are introduced to reach enhanced quality goals, the data volume may increase beyond what is projected in our model. This will be addressed in a future revision of the HIR roadmap.

Table 1: Logic Assumptions

Year of Production	2018	2019	2020	2021	2026	2031
Design Related Backgound Data						
Chip size at production (mm^2)						
MPU-HP - High Performance MPU (Server)	261	261	261	261	261	261
MPU-CP - Consumer MPU (Laptop/Desktop)	140	140	140	140	140	140
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	140	140	140	140	140	140
# of Transistors (M) [T]						
MPU-HP - High Performance MPU (Server)	15306	19284	22495	26241	56684	122447
MPU-CP - Consumer MPU (Laptop/Desktop)	5768	7268	8478	9890	21363	46147
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	2756	3438	4291	5357	16372	50448
% Logic Transistors						
MPU-HP - High Performance MPU (Server)	35%	35%	35%	35%	35%	35%
MPU-CP - Consumer MPU (Laptop/Desktop)	63%	63%	63%	63%	63%	63%
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	81%	82%	83%	83%	87%	88%
% Memory Transistors						
MPU-HP - High Performance MPU (Server)	65%	65%	65%	65%	65%	65%
MPU-CP - Consumer MPU (Laptop/Desktop)	37%	37%	37%	37%	37%	37%
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	19%	18%	18%	17%	13%	12%
% Random Transistors						
MPU-HP - High Performance MPU (Server)	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%
MPU-CP - Consumer MPU (Laptop/Desktop)	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%
VDD(V)						
MPU-HP - High Performance MPU (Server)	0.78	0.77	0.75	0.74	0.66	0.64
MPU-CP - Consumer MPU (Laptop/Desktop)	0.78	0.77	0.75	0.74	0.66	0.64
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	0.57	0.55	0.53	0.51	0.43	0.41
Test Related Assumed Data						
Maximum Power Consumption at Test (W)						
MPU-HP - High Performance MPU (Server)	400	400	400	400	400	400
MPU-CP - Consumer MPU (Laptop/Desktop)	250	250	250	250	250	250
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	40	40	40	40	40	40
Effective Number of External Scan Pins (TDI+TDO) [1]						
MPU-HP - High Performance MPU (Server)	256	256	256	256	256	256
MPU-CP - Consumer MPU (Laptop/Desktop)	256	256	256	256	256	256
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	64	64	64	64	64	64
Effective External Scan data rate (Mbps) [1]						
MPU-HP - High Performance MPU (Server) [2]	131	137	144	150	150	150
MPU-CP - Consumer MPU (Laptop/Desktop)	131	137	144	150	150	150
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	86	92	98	105	147	150
% Scan Chains Shared Btwn Homgeneous Logic (all device types) Homogeneous logic refers to logic contained in IP cores.						
MPU-HP - High Performance MPU (Server)	66%	67%	68%	70%	75%	80%
MPU-CP - Consumer MPU (Laptop/Desktop)	59%	61%	62%	64%	71%	76%
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	37%	40%	42%	44%	55%	63%
Target Compression Ratio (all device types) [3]	134	143	154	165	251	410

[1] The external scan pin count and scan data rate are intended to suggest a scan bandwidth (BW=Pins*Data Rate). Since some devices use higher external data rates (via a SerDes) these terms are indicated as "effective" in order to normalize the results.

[2] slow growth rate maxing out at 150MHz [3] Was 10% per year growth, changed to base growth + acceleration

Table 2: Logic Test Data Volume

Year of Production	**2018**	**2019**	**2020**	**2021**	**2026**	**2031**
Worst Case (Flat) Data Volume (Gb)						
MPU-HP - High Performance MPU (Server)	6802	9256	11366	13957	38973	108829
MPU-CP - Consumer MPU (Laptop/Desktop)	4130	5620	6901	8475	23664	66079
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	2133	2907	3964	5406	25444	116008
Best-Case Test Data Volume (Hierarchal & Compression) (Gb)						
MPU-HP - High Performance MPU (Server)	10.3	12.0	12.6	13.1	15.3	17.0
MPU-CP - Consumer MPU (Laptop/Desktop)	8.5	9.8	10.2	10.7	12.3	13.0
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	7.0	8.4	9.6	11.5	24.2	45.5
Best-Case Compression Factor (Hierarchal & Compression) (Compression rate)						
MPU-HP - High Performance MPU (Server)	658	770	905	1069	2547	6388
MPU-CP - Consumer MPU (Laptop/Desktop)	488	571	675	795	1924	5084
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	303	347	414	472	1050	2551

Table 3: Logic ATE Requirements

Year of Production	2018	2019	2020	2021	2026	2031
Minimum Required ATE Scan Pattern Depth (Relative to 2016 Pattern Depth)						
MPU-HP - High Performance MPU (Server)	X 1.34	X 1.54	X 1.60	X 1.64	X 1.84	X 1.97
MPU-CP - Consumer MPU (Laptop/Desktop)	X 1.34	X 1.55	X 1.58	X 1.63	X 1.80	X 1.81
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	X 1.41	X 1.68	X 1.87	X 2.19	X 4.32	X 7.76
Maximum Required ATE Scan Pattern Depth (Relative to 2016 Pattern Depth)						
MPU-HP - High Performance MPU (Server)	X 1.85	X 2.52	X 3.09	X 3.80	X 10.61	X 29.63
MPU-CP - Consumer MPU (Laptop/Desktop)	X 1.85	X 2.52	X 3.09	X 3.80	X 10.61	X 29.63
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	X 1.85	X 2.53	X 3.45	X 4.70	X 22.12	X 100.84
Typical Test Time (Relative to 2016 Pattern Depth) [1]						
MPU-HP - High Performance MPU (Server)	X 1.68	X 2.17	X 2.54	X 3.00	X 8.36	X 23.34
MPU-CP - Consumer MPU (Laptop/Desktop)	X 1.68	X 2.17	X 2.54	X 3.00	X 8.35	X 23.30
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	X 1.62	X 2.06	X 2.62	X 3.33	X 11.15	X 49.87

[1] Test times assume that all the ATE Patterns are executed once without any looping or repetition

Trends impacting roadmap 2017-19

The model presented in this roadmap was generated in 2017, it is fair to ask about the impact of Artificial Intelligence (AI) and Automotive devices. Both of these markets are emerging as important drivers to our industry. What is known is that present AI devices best fit into the SOC category in 2019. While the number of compute elements are highly replicated, the way the designs are partitioned in the DFT tools do not take advantage of the hierarchy as do GPU device types. Over the near term we expect this to continue as there is a push for new architectures and product validation in the marketplace. As the AI segment matures, the effective compression rate and ATE Scan Pattern Depth will more closely track the MPU-HP segment for the large AI training and inference device types. Automotive also provides additional demands on this roadmap with the need for additional self-test coverage as well as increased quality which may drive further test intensity. Both topics will require further study for the next update.

Summary

The tables reflect the continuation of the trends that are known and that are accepted by industry. Historically, innovation in digital test has been driven by the economic need to keep the cost of test relatively flat. If Moore's Law scaling continues, it can be seen in the 2025 projections that an economic need for innovation is emerging. The cost problem may be further compounded by emerging automotive desire for part-per-billion (PPB) failure rates on leading semiconductor process nodes. These projected trajectories may be affected by alternative scan loading methodologies using high speed IO, redundant multi-core systems and the adoption of inline diagnostics. In addition to the test data increases which we are projecting, the team has been observing a category of "system level" faults in

digital systems that are not caught by BIST or traditional coverage techniques. Since these faults could lead to a negative user experience, some product companies are augmenting traditional test coverage with a System Level Test insertion, or crafting System Level Tests at ATE wafer/package test. This group will continue to monitor future areas of growth and adjust the roadmap as they become generally accepted methodologies.

Section 4: Specialty Device Testing

A classification of specialty devices was defined in industry roadmaps beginning in 2006, driven by strong high-volume market demand, but having odd test requirements. Examples are CMOS image sensors, LCD drivers, MEMS devices (including multimode sensors), actuators, bio-MEMS, and similar non-standard devices.

Trends Impacting this Technology Area

The applications of Mobile personal devices, IOT, Healthcare, Automotive/ADAS and Robotics are key drivers of specialty devices to motivate innovative technologies and the high growth rate of volume.

The trends of technologies: (Near Term < 5 years)

The mobile and wearable devices for IOT and healthcare applications are major drivers, impacting technology trends in the near term.

- The trends for multi-mode MEMS sensors is fusing multiple functionalities together in one device and also reducing the size of the package to be smaller and thinner for adding value in a compact unit. See Figure 1.

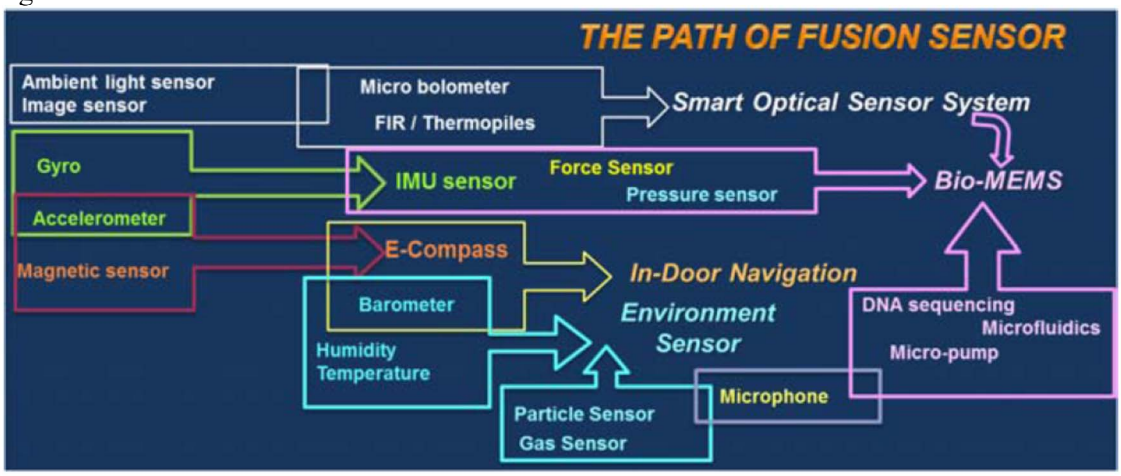


Figure 1: The path of MEMS sensor fusion

- The technology trends for image sensors lead to highly integrated multiple wafers using a 3DS (three-dimensional stacking) process and novel packaging technologies for enhancing image performance with cost-effective mass production solutions. The first successful step involving 3DS wafer processing of image sensors was the BSI (Back Side Illumination) process which bonded a photo-sensor wafer together with a back-side mixed-signal data processing wafer, connecting their signals through TSVs. When image-sensor pixel numbers increased and sensor size shrank below 1.4um, the BSI process had the advantage of a lower signal to noise ratio than the traditional FSI (Front Side Illumination) process. The next step in the image-sensor wafer-integration process adds a memory-cell wafer between the photo sensor wafer and the mixed-signal data processing wafer, which could enhance image performance and the speed of data processing in a variety of imaging applications such as 3D imaging, face recognition, and image capture, with frame rates over 1000 frames/second. See Figure 2.
- Image-sensor packaging innovation began using WLCSP (Wafer Level Chip Scale Package) to develop the glass-based stacking wafer-level package with TSV or Shallcase type processing, as shown in Figure 3.
- The trends for new WLP for image sensors are WLO (Wafer Level Optics) and WLCM (Wafer Level Camera Module) which stack optical systems on the image-sensor wafer using a wafer-level packaging process to reduce the size of optical systems and increase the efficiency of mass production. See Figure 4.

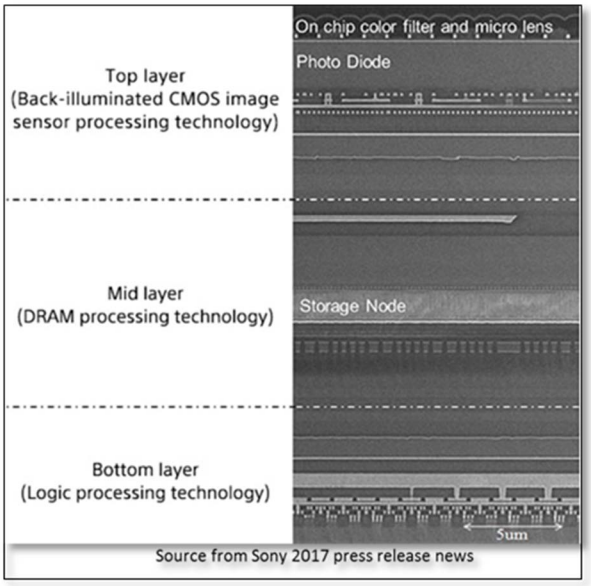


Figure 2: Image sensor 3D stacking wafers
(Backside illumination photo sensor wafer + memory cell wafer+ data processing wafer)

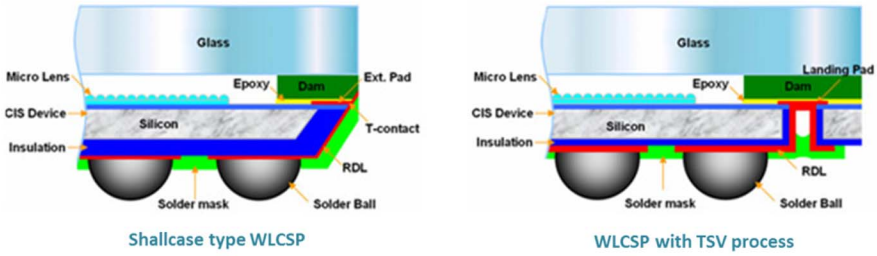


Figure 3: The structure of Image sensor WLCSP(wafer level chip size package)

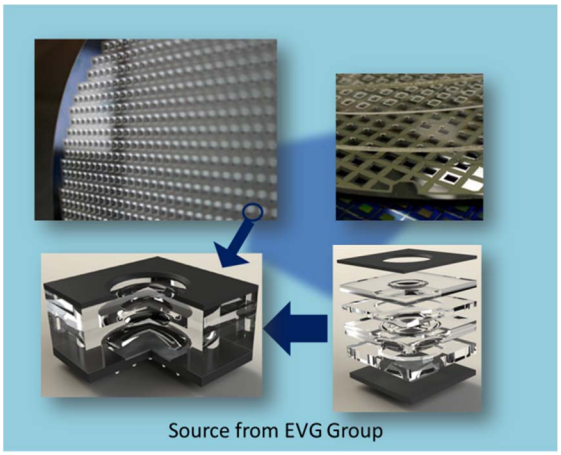


Figure 4: Image sensor WLO (Wafer Level Optics) packaging

The trends in technologies: (Medium to Long Term < 15 years)

The automotive, robotic, medical and intelligent artificial organ fields are the next wave of drivers for specialty devices which impact technologies in the medium and longer term.

- Reliability will become a very important subject for specialty devices. Burn-in and tri-temperature testing will become necessary test procedures during mass production.
- Built In Self-Diagnostic, Self-Calibration & Compensation and Self-Repair technologies will become important design skills to apply on specialty devices for enhancing reliability performance.

Concerns: Test Challenges

LCD display drivers:

LCD display drivers are unique because of their die form factor, which can have larger than a 10:1 aspect ratio and thousands of very narrow gold-bump pads requiring contact for test. In 2017, in-line probing pad pitch for LCD display drivers already was down to 18μm, and stager pad pitch was 13μm. Right now, only the cantilever probe card is a major cost-effective solution for achieving probing of LCD drivers with such narrow and fine pitch pad with gold bump in mass production.

An upcoming test challenge is the data transfer speed of I/O, which will increase to 2.5 Gbps in 2019 and is predicted to be up to 6.5 Gbps within 10 years. We need to overcome the challenges of probing fine-pitch bumping pads with high-speed signals with economical probing solutions.

Image sensor devices:

The testing of image sensor devices needs to consider special test requirements for optical systems and huge image data processing. The innovative technical trends of highly integrated 3DS CMOS image sensors increase the difficulties of special test requirements and challenges.

		Year of Production	2017	2018	2019	2020	2021	2026	2031
Process Integration	**Test Method**	**Challenges**							
Wafer level probe (BSI process + Memory+ ASIC , 3 layer W to W)	CP	Wafer probe with multi-sites	Multi- insertion						
			Single insertion (ATE)						
		Wafer probe by full wafer contact	Optical system (Visible light)						
			Technology of optical asserroy & probe card						
			(ATE) / challenges test system resource						
WLCSP (BSI process + Memory +ASIC , 3 layer W to W)	FT	Test after singular (Pkg form)	Multi- insertion						
			Single insertion (ATE)						
		Test after dicing (wafer form)	Optical system (Visible light)						
			Probing methods and accessories, multi-sites						
		Test after dicing with full wafer contact (wafer form)	Optical system (Visible light)						
			Probing methods and accessories						
			(ATE) / challenges test system resource						
WLO-P (Multi-layer W to W)	FT	Test after singular (Pkg form)	Single insertion (SLT)						
		Test after dicing (wafer form)	Optical system (Visible light)						
			Probing methods and accessories						
		Test after dicing with full wafer contact (Full functional test under wafer form)	Optical system (Visible light)						
			Probing method and accessory						
			Challenges ATE or SLT test system resource						

Research Required
Development Underway
Qualification / Pre-Production
Continuous Improvement

Table 1: Specialty device odd test potential solutions table – Image sensor device

Automotive ADAS applications and intelligent machine vision especially need the functionalities of image sensors with a wide spectrum (from UV to FIR), high dynamic range, good S/N (Signal to Noise) ratio, fast data frame rate, better quality and reliability, which challenges test system design. Burn In solutions also need to include optical stress for sorting out defects in the coating process on the photo sensor surface.

MEMS devices (Sensor, Actuator and Biological):

MEMS were successfully applied for various sensors for sensing motion, magnetic field, optics, sound, air pressure and vibration, flow, chemical composition of air, DNA sequencing, and other characteristics, and the market volume is increasing rapidly due to IoT, healthcare and automotive applications. Testing MEMS sensor devices with suitable physical stimulus and cost-effective solutions for the various types of sensors is difficult and tricky. Especially testing the expanding kinds of fusion sensors will bring many test challenges.

			Year of Production	2017	2018	2019	2020	2021	2026	2031
Process integration		**Test Method**	**Challenges**							
IMU sensor (Accelerometer + Gyro)	CP / wafer probe	Probing MEMS wafer (DC only)	Probe card technology/multi-sites vs. cost							
		Probing each wafer with full functions (Multi-insertion)	Motion Prober system and probe card technology/multi-sites vs. cost							
		Probing 3DS wafer with full functions (Single-insertion)	DFT design and implement							
	WLP	Test before dicing (Wafer form)	DFT design and implement							
		Test after singular (Pkg form)	Handling small size package							
	FT	Final test with full functions (Multi- insertion)	Test cost is high							
		Final test with full functions (Single- insertion)	Reduce test coverage rate							
			DFT design and implment							
		Burn In Test	SLT							
			BISX (Build In Selft X) X= (Test, Diag-nostic, Correlation,Compensation/Repair)							
Navigation (G-sensor+ Gyro+ Magnetic sensor + Barometer)	CP / wafer probe	Probing MEMS wafer (DC only)	Probe card technology/multi-sites vs. cost							
		Probing each wafer with full functions (Multi-insertion)	Probing system with situmlus & probe card technology							
	FT	Final test with full functions (Multi- insertion)	Test cost is high							
		Final test with full functions (Single- insertion)	Reduce test coverage rate							
			DFT design and implment							
		Burn In Test	SLT							
			BISX (Build In Selft X) X= (Test, Diag-nostic, Correlation,Compensation/Repair)							
Enviromential Sensor (Pressure + Humidity + Gas) Sensor	CP / wafer probe	Probing MEMS wafer (DC only)	Probe card technology/multi-sites vs. cost							
		Probing each wafer with full functions (Multi-insertion)	Probing system with situmlus & probe card technology							
	FT	Final test with full functions (Multi- insertion)	Test cost is high							
		Final test with full functions (Single- insertion)	Reduce test coverage rate							
			DFT design and implment							
		Burn In Test	SLT							
			BISX (Build In Selft X) X= (Test, Diag-nostic, Correlation,Compensation/Repair)							

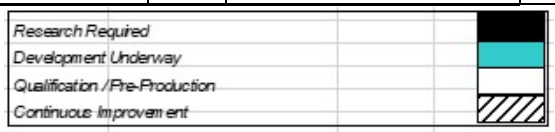


Table 2: Specialty device odd test potential solutions table – MEMS Fusion Sensor

DFT for MEMS sensor devices is a new technology and needs research and innovative development for different kinds of sensor structures. MEMS sensors DFT needs to develop the stimulus source and the sensor together in the MEMS structure as a BIST (Build-In-Self-Test) cell. When testing, the cloned control signal of physical stimulus is generated from the MEMS ASIC to enable the MEMS BIST cell to imitate physical stimulus for testing the sensor cell to achieve the DFT concept. This concept could also implement the technologies for BISD (Build-In-Self-Diagnostic), BISC (Build-In-Self-Correlation/Compensation) and BISR (Build-In-Self-Repair) to enhance reliability of MEMS sensors for automotive and medical applications. The key during testing is to make sure the BIST cell works well.

Beyond MEMS sensors, there are also actuator and biological applications such as micro-mirrors, MEMS speakers, RF switches, energy harvesting, microfluidics, micro-dispensers and artificial organs, plus others. The challenges of testing MEMS actuators and biological devices are that test methods are hard to standardize and depend on the structure for each different kind of MEMS device. Especially for the testing of biological devices, the test environment can be severe and needs to pass safety certification based on the laws of different grades and countries.

Summary

Specialty devices as defined need odd test requirements and are driven by strong high-volume market demand. Under these two conditions, the trends of specialty devices will be drive toward highly integrated multi-functions in one smaller unit to overcome ASP (Average Sale Price) erosion, and testing procedures will move toward high parallelism to reduce test cost. Test challenges will follow the same trends to overcome testing evolutionary HI (Heterogeneous Integration) specialty new product through cost effective solutions.

References

"Sony Develops the Industry's Fist 3-Layer Stacked CMOS Image Sensor with DRAM for Smartphone", Sony press release, February 7 ,2017

"Xintec 12-inch WLCSP Line Ready for Mass Production in 2H 2015" , DIGITIMES, March 2015

"EVG's wafer-level optics (WLO) manufacturing solutions", EVG press release, September 11 ,2017

Section 5: Memory Test

There have been significant changes in the Memory environment in recent years. Up until 2003, DRAM bits comprised approximately 90% of bits shipped per month. By 2009, NAND had become the dominant form of memory and comprised 85% of monthly bits shipped. During the same period, NOR Flash has largely migrated from parallel interface devices to serial interface devices with extremely small form factors in order to reduce PCB size, complexity and power. With the introduction of mobile smart phones in 2007 and tablets in 2010, the traditional dominance of PC DRAM has been eroded by lower-power LPDRAM which is required for longer battery life. New generation memory types such as PRAM, RRAM, STT RAM, and CBRAM along with 3D multilayer instantiations of NAND will further change the memory environment over the next decade. The continuing shift to a battery-powered wireless environment will continue to drive changes in memory usage. Existing memory types will likely not be replaced, but will be used in joint solutions with newer memory types.

Memory density has kept pace with Moore's Law since the first DRAM was manufactured in 1969, but lithography cycles as well as the performance roadmap are expected to push out for litho nodes less than 20nm, so the typical 2-year technology pace is forecasted to stretch to initially 3 years, and then 5 years later in the roadmap. Multi-layer 3D NAND technologies allowed the use of longer gate lengths, so the 2-year technology pace should continue for the near term.

From a test perspective, most memory will be structurally tested at wafer test utilizing low-pin-count interfaces of 4 to 10 pins per device. Structural test, when used, will likely include a self-test performance validation.

Table 1: Memory Test Requirements

	2018	2019	2020	2021	2026	2031
DRAM Characteristics						
Capacity (bits) [1]	16G	16G	32G	32G	64G	128G
I/O data rate (Gb/s)						
PC DDRx	3.2	4.4	4.4	6.4	8.4	8.4
GDDRx	10.0	12.0	16.0	16.0	16.0	16.0
LPDDRx	4.0	4.8	4.8	4.8	8.4	8.4
Hybrid Memory Cube (HMC [SerDes])	15.0	15.0	15.0	15.0	25.0	30.0
Wide IO	0.3	1.1	1.1	1.1	2.2	2.2
High Bandwidth Memory (HBM)	2.0	2.6	2.6	2.6	3.0	3.6
NAND Characteristics						
Capacity (bits) [2][3][4]	1T	1T	2T	2T	16T	64T
Maximum I/O data rate (Gb/s)	0.53	0.53	0.53	0.67	1.07	1.60

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

Notes:

1. DRAM bit capacity per die
2. 3 bits per cell introduced in 2009
3. 4 bits per cell introduced in 2012
4. 3D multi-layer introduced in 2014

DRAM

Historical trends show that PC DRAM has doubled its performance every 5 years and will reach a data rate of 8.4 Gb/s per I/O in 2022 with DDR6. However, DDR4 appears to be the end of the DDRx era as there are no further enhancements of the DDRx architecture beyond DDR4 in definition or development. DDR4 itself has severe PCB design restrictions in order to meet the I/O performance requirements, so an enhancement of the current DDRx single-ended interface will present additional challenges and constraints. The Wide IO memory architecture is targeted to mobile applications such as phones and tablets and is an evolution of the DDRx that decreases I/O bit rates while expanding the number of I/Os up to 512. Hybrid Memory Cube (HMC) is targeted to servers, where it provides very high performance over a SERDES interface, though at increased cost. High Bandwidth Memory (HBM) is similarly

targeted to graphics video cards and applications. Based on existing roadmaps, DDR3/4 will continue to serve the PC market for the foreseeable future. LPDDRx and GDDRx DRAM families will continue to be a driver for the near future.

DRAM will become increasingly difficult to scale in sub-20nm nodes, and transistor wear-out will increase the frequency of errors. On-chip error correction and memory management will likely become a requirement before 2020. Dynamic failure detection, analysis, and repair will become necessary over the product life. To enhance test productivity, new test-oriented architectures will be required. On-chip correction may also change the DRAM test paradigm.

Maintaining high ATE test parallelism is required over the roadmap period to manage test cost. However, probe card performance test at high parallelism may be a challenge at high GT/s due to the interface routing complexity required. These challenges will ultimately drive the need for die self-test.

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 increase in the number of bits stored in a single memory cell from one and two to four on some cell types. Use of error correction allows greater uncorrected error rates and enables increased bits per memory cell. NAND bus width has continued to be dominantly 8-bit with a decreasing number of products at 16-bit I/O. As large amounts of NAND are being consumed in Solid State Drives (SSD), a new NAND interface may emerge that is more optimized for SSD use.

The need for internal voltages that are 3-8 times the external supply requirements is expected to continue in the test process, 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 operational 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 are required.

NOR memory density is expected to increase slowly over the roadmap period and remain flat toward the end of the roadmap. NOR has been transitioning from a parallel to a serial interface since 2007 to reduce package size and power. Further increases in NOR performance along with increasing test requirements are not expected.

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.

Embedded Memory

Embedded memory consumes greater than 80% of transistors in many MPU and SoC designs and will scale with the increase of transistors in these devices. Embedded Flash and DRAM bits will not match the density of standard DRAM and NAND. Newer memory types such as RRAM or STT RAM may become embedded over the course of the roadmap.

To enhance test productivity, new test-oriented architectures and/or interfaces will be required. Built-in self-test (BIST) and built-in self-repair (BISR) will be essential to test embedded DRAM and Flash memories cost effectively. 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 are tested and repaired on a memory tester, while the logic blocks are tested on a logic tester.

Section 6: Analog and Mixed Signal Test

Forward

The economic benefit of monolithic integration (SoC) and system in package (SiP) is well established and continues on. This integration has combined digital logic and processing, analog, power management, and mixed signal 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.

An 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 and Parallel Test Efficiency (PTE), also known as Multi-Site Efficiency (MSE), high; this is of increasing importance to avoid severely impacting Units Per Hour (UPH).

A similar concept but in a dimension relating to the single device itself is testing multiple IP cores within the device in parallel (concurrent test). This has many of the requirements and challenges of parallel test, but also includes some unique ones. A key one is having the ability in the design of the IC to test IP cores independently, in parallel. Test Access Mechanisms (TAMs) are the ability of IP cores to be accessed and controlled independently from other IP cores. The most powerful economic advantage results when being able to test multiple IP cores in parallel, while at the same time testing multiple devices in parallel.

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 typically non-deterministic and must be post 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 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. The requirements for the TAM are covered in the DFT SOC Device Testing section. The requirements for DC trim accuracy are included in the Mixed Signal tables (see Table 1).

Table 1: Mixed-signal Test Requirements

	2018	2019	2020	2021	2026	2031
Low Frequency Waveform [Note 1]						
SFDR	145	145	145	145	145	145
SNR	120	120	120	120	120	120
THD	140	140	140	140	140	140
BW-Minimum (kHz)	50	50	50	50	50	50
BW-Maximum (kHz) [Note 2]	500	500	500	500	500	500
High Frequency Waveform Source / Measure [Note 3]						
Level V (pk–pk)	<4	<4	<4	<4	<2.5	<2.5
BW (MHz)	250	250	250	250	500	500
Sample rate (MS/s) [Note 5]	500	500	500	500	1000	1000
Resolution (bits) AWG/Sine	16	16	16	16	18	18
Noise floor (dB/RT Hz)	-140	-140	-140	-140	-150	-150
Very High Frequency Waveform Source / Measure [Note 4]						
Level V (pk–pk)	<4	<4	<4	<4	<4	<4
Accuracy (±)	0.50%	0.50%	0.50%	0.50%	0.50%	0.50%
Measure BW (GHz) (under sampled)	6.4	6.4	9.6	9.6	15	15
Capture Depth Mwords	4	4	4	4	4	4
Min resolution (bits)	8-10	8-10	8-10	8-10	8-10	8-10
DC Accuracy (Note 6)						
DC force (uV)	100	100	50	50	50	50
DC measure (uV)	100	100	50	50	50	50
DC force (nA) (Note 7)	10	10	5	5	1	1
DC measure (nA) (Note 7)	10	10	5	5	1	1
Ethernet						
Speeds (GBPS)	40	40	40	40	100	400

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

NOTES:

1) Audio / Precision; Source & Measure specifications (22 KHz BW)
2) Major testing condition
3) Target Devices are Wireless Baseband, xDSL, ODD, Digital TV (Track Mobile Baseband)
4) Target Devices are HDD, Radar, WiGig
5) For Measure Sample Rate: Dependent on method, tracking or Front End filter.
6) The purpose of DC accuracy for this table is for high reolution force/measure and trim
7) Devices may also need high current with the less accuracy

Key Test Trends

Short-Term Trends (< 5 Years)

There are three important trends. The first is to deliver adequate quality of test. Most analog/mixed-signal testing is done through performance-based testing. This includes functional testing of the device and then analyzing the quality of the output(s). 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. See the Mixed Signal test tables (Table 1) for updates and future needs.

The second key trend is the need for higher DC accuracy. Many of the converters and precision references are made more accurate by doing a measure and trim step. The trim can be accomplished through several means; one of the more recent and cost-effective ways is through register programming of the device. The trim takes a relatively lower performance device and adds high accuracy to it through a DC test and register programming. In the past, this was done for medium performance devices, but now the test methodology has matured and it is being applied to high accuracy/resolution devices. The change is that in this class of devices, much higher DC accuracy is required to make a valid test.

The third key trend is to enable the economics of test through instrumentation density and Parallel Test Efficiency (PTE). The level of parallelism requires 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 continue to lag. No proven alternative to performance-based analog testing has been widely adopted 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, partial BIST, or elimination of the need for external instrumentation altogether.

The Ethernet trends are continuing into higher speeds – 28, 40 Gbps per channel and even beyond. [1] There continues to be the need for backwards compatibility to the many existing digital communication standards.

Difficult Challenges in the Short Term

- As reflected in the tables, manufacturing solutions exist for the immediate future testing needs. However, high DC accuracy for sourcing, measuring and for trim/fuse blowing/register-setting in a manufacturing environment could be at issue depending on how high a resolution/accuracy the DUT is. Also 40 Gbps Ethernet has known manufacturing solutions, but none are optimized.
- Time-to-market and time-to-revenue issues are driving test to be fully ready at first silicon. The analog/mixed-signal test environment can seriously complicate the test fixtures and test methodologies. Noise, crosstalk on signal traces, added circuitry, load board design complexity, and debug 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, require 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 the items above.

Medium-term Trends (6 to 10 years out)

- For Wireless Baseband, xDSL, ODD, and Digital TV (Track Mobile Baseband) devices, the source and measure bandwidths, sampling rates and resolutions increase, while the noise floors are decreasing.
- Additionally, DC force and measure accuracies get more challenging.
- Ethernet speeds trending to 100 Gbps [2] have only interim solutions identified.
- Higher speeds and modulation will necessitate PAM to handle the increased data bandwidth – for example, PAM 4, 8 or 16 at speeds of 32 GBPS. [3], [4]

Difficult Challenges in the Medium Term

- As the capability requirements increase, there are solutions available, but they do not lend themselves easily to high volume manufacturing.
- Basic physical and electrical properties come more into play. For example, a -150 dB noise floor is possible, but special fixturing is required that is difficult to deploy into a manufacturing environment.

- Ethernet speeds of 100 Gbps [2] have only interim solutions identified.

Long-term Trends (10 years+ out)

- Ethernet speeds trending to 400 Gbps [5], [6]

Difficult Challenges in the Long Term

- Ethernet speeds of 400 Gbps do not have known manufacturing solutions identified.

SUMMARY

Cost continues to be the most critical pressure and concern for analog mixed signal because much of the volume for this is consumer oriented. However, in the medium and long term, performance starts becoming an issue for high-volume manufacturing in terms of bandwidth, sample rate, resolution and noise floor to keep up with the newer devices on the horizon. Ethernet in the medium and long term has manufacturing challenges both in optimization and known solutions.

References

1. https://www.cisco.com/c/dam/en/us/products/collateral/switches/catalyst-6500-series-switches/white-paper-c11-737238.pdf
2. https://www.networkcomputing.com/data-centers/25-50-and-100-gigabit-ethernet-data-center/1422885308
3. http://www.ieee802.org/3/100GNGOPTX/public/mar12/plenary/tremblay_01_0312_NG100GOPTX.pdf
4. http://www.ieee802.org/3/bm/public/sep12/ghiasi_01a_0912_optx.pdf
5. http://searchdatacenter.techtarget.com/feature/Understanding-tomorrows-Ethernet-speeds
6. https://www.keysight.com/main/application.jspx?ckey=2716254&id=2716254&nid=-32176.0.00&lc=eng&cc=US

Section 7: Wafer Probe and Device Handling

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.

Device Handling Trends

Increased parallelism at wafer probe drives a greater span of probes across the wafer surface and significantly increased probe card complexity. Prober and probe card architecture should evolve to simplify the interface, however just the opposite is happening: ATE tester complexity is decreasing and more technology and complexity is built into the probe card interface. A better thermal solution is a very important parameter along with performance for better yield management. Memory applications are increasing the total power across a 300mm wafer, and wafer probe needs to dissipate this total power to sustain the set-temperature during test. Power density per DUT is increasing and it's very challenging to manage a stable wafer-level test temperature. 3D integration technology requires very precise probing technology in X, Y and Z, as micro-bumps may be easily damaged during the probing process. MEMS applications require a variety of testing environments such as pressure, magnetic, and vacuum environments; also, wafer shape and package style are becoming very unique depending on the application type.

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 reduced test times), faster handler speed, or process improvements such as asynchronous test or continuous-lot processing. 3D integration technology requires new contact technology for the intermediate test insertion which will be added between conventional front-end process and back-end process. New contact technology to probe on the singulated and possibly thinned die's micro-bumps or C4 bumps after the die is mounted on an interposer is needed. For the die-level handler, the main tasks are the alignment accuracy to enable fine pitch contact, die level handling without damaging the die, and the tray design that supplies/receives the die.

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, die-level handlers as well as package 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, immediate heat control technology, and temperature control to enable stable DUT 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 wide temperature range, high temperature accuracy, high throughput, placement accuracy, parallelism, and special handling needs while still being cost effective in a competitive environment is a significant challenge.

Gravity feed, turret, and strip handlers have been added to the table while retaining the pick and place type handler. The gravity feed handler is used on SOP, QFN, and DIP packages. Turret handlers are widely used on discrete-type QFN devices. Strip handlers are used on the frame before singulation. Strip test enables high parallelism with fewer interface resources, which enables cheaper test cost. These additional three types of handlers are widely used on relatively low-end or low-cost devices. Evolution of these handlers is quite different but important for various type of LSI.

Table 1: Test Handler and Prober Difficult Challenges

Pick and Place Handlers (High Performance)	Temperature control and temperature rise control due to high power densities
	Continuous lot processing (lot cascading), auto-retest, asynchronous device socketing with low-conversion times
	Better ESD controls as products are more sensitive to ESD. 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)
	Force balancing control for System in Package and Multi-Chip Module
Pick and Place Handlers (Consumer SoC/ Automotive)	Support for stacked die packaging and thin die packaging
	Wide range tri-temperature soak requirements (-55ºC to 175ºC) increases system complexity for automotive devices
	Device junction temperature control and temperature accuracy +/-1.0°C
	Fine Pitch top and bottom side one shot contact for Package on Package
	Continuous lot processing (lot cascading), auto-retest, low conversion times, asynchronous operation
Pick and Place Handlers (Memory)	Thin die capable kit-less handlers for a wide variety of package sizes, thicknesses, and ball pitches < 0.3mm
	Package ball-to-package edge gap decreases from 0.6 mm to 0 mm require new handling and socketing methods
	Parallelism at greater than x128 drives thermal control +/-1.0°C accuracy and alignment challenges <0.30mm pin pitch
Prober	Consistent and low thermal resistance across the chuck is required to improve temperature control of the device under test. There is a new requirement of active/dynamic thermal control, which can control junction temperature(ΔT) during test
	Both Logic and Memory wafer generates more wattage/heat, demand of Heat dissipation performance improvement is expected. Especially Heat Dissipation at Hot temperature is challenging technology for wafer prober.
	There are wafer handling requirements of non-SEMI standard such as 3DI, MEMS, WLCSP and PsP applications. Those are thin, thick, unique shape so customized wafer handling technique/technology is needed. Wafer cassette is needed to be customized to meet the request as well.
	Probing on micro-bump is technically proven but there are many challenges "parallelism/multi-site", "Thermal conduction" and "bump damages/reliability"
	Advances in probe card technology require a new optical alignment methodology.
	Dicing frame probers can cover a wide temperature range, but a dicing sheet cannot cover the full range.
	Greater parallelism/multi-site, and higher pin counts require higher chuck rigidity and a robust Probe Card changer.
	Power Device application requires very thin wafer which drive need for 'Taiko Wafer' and 'Ring attached wafer' handling and more high voltage chuck technologies.
	Enhanced Probe Z control is needed to prevent damage to pads, there are solution in the market but those must be optimized to integrate onto wafer prober to meet needs of test cost requirement.
Gravity Feed Handlers	Thinner packages and wafer will require a reduction in the impact load to prevent device damage
	Test head size increase due to higher test parallelism may alter handler roadmap
	Reduction of static electricity friction and surface tension moisture friction on very small packages (<1 x 1 mm)
Turret Handlers	Test contactor support for > 100A current forcing on power devices
	Kelvin contact support (2 probes) to very small area (0.2 x 0.2mm) contacts on small signal devices
Strip L/F Handlers	Testing process infrastructure configuration
	Accuracy of the contact position for high temperature testing environment

Table 2: Wafer Probe Technology Requirements (part 1)

Year of Production	2018		2019		2020		2021		2026		2031	
MPU, ASIC, SOC and Mixed Signal Products												
Wirebond - inline pad pitch [1]	40		40		35		35		35		35	
Wirebond - stagger pad pitch	45		45		30		30		30		30	
Bump - array bump pitch	30		30		30		30		25		25	
Sacrifical pad pitch in a field of bumps	100		100		100		100		100		100	
I/O Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Wirebond	35	35	30		30		30		25		25	
Bump	30		30		30		25		25		25	
Sacrifical pad in a field of bumps	45		45		45		40		35		30	
Wafer Test Frequency (Hz)	1.6G		2.4G		2.4G		3.2G		4G		5G	
Wafer Test Frequency (Hz) for HSIO	16Gbps/8GHz		25Gbps/12.5GHz		25Gbps/12.5GHz		33Gbps/16.5GHz		50Gbps/25GHz		50Gbps/25GHz	
Probe Tip Diameter Wirebond	8		7.5		6.5		6.5		6		6	
Probe Tip Diameter Bump	25		25		25		25		25		25	
Probe Force Bump(gf) - at recommended overdrive	1.5		1.5		1.5		1.5		1.5		1.5	
Size of Probed Area (mm2)	20000		20000		20000		20000		20000		20000	
Number of Probe Points / Touchdown	150000		180000		200000		200000		250000		300000	
Maximum current per probe >130um pitch[2]	1.5A		2A		2A		2A		2A		2A	
Maximum current per probe <130um pitch[2]	1A		1A		1A		1A		1.5A		1.5A	
Maximum contact resistance	<0.5		<0.5		<0.5		<0.5		<0.5		<0.5	
Probe test temperature range	-55	185	-55	200	-55	200	-55	200	-55	200	-55	200
Automotive Radar												
Wafer Test Frequency (GHz)	80GHz		80GHz		80GHz		80GHz		80GHz		80GHz	
RF Pad Geometry [3]	Solder Ball		Solder Ball		Solder Ball		Solder Ball		Solder Ball		Solder Ball	
Pad Size (um)	250 µm SB		100 µm Cu Pillar SB		100 µm Cu Pillar		100 µm Cu Pillar		100 µm Cu Pillar		100 µm Cu Pillar	
Pad Pitch (um)	500 µm		300 µm		300 µm		300 µm		300 µm		300 µm	
RF Ports per Site	13		13		13		13		13		13	
Sites being probed together	2		2		4		4		4		4	
Total # of RF Ports	26		26		52		52		52		52	
High Speed Digital (TIA, CDR, VCSEL, etc)												
Wafer Test Frequency (GHz)	67 GHz		67 GHz		67 GHz		67 GHz		67 GHz		67 GHz	
RF Pad Geometry	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Pad Size (um)	55	55	50	50	50	50	50	50	40	40	40	40
Pad Pitch (um)	90 µm		80 µm		80 µm		80 µm		60 µm		60 µm	
RF Ports per Site	24		24		24		24		24		24	
Sites being probed together [4]	2		2		4		4		4		4	
Total # of RF Ports	48		48		96		96		96		96	
802.11ad												
Wafer Test Frequency (GHz)	64 GHz		64 GHz		64 GHz		64 GHz		64 GHz		64 GHz	
RF Pad Geometry	Solder Balls		Solder Balls		Solder Balls		Solder Balls		Solder Balls		Solder Balls	
Pad Size (um)	80 µm		80 µm		70 µm		70 µm		60 µm		60 µm	
Pad Pitch (um)	150 µm		150 µm		125 µm		125 µm		100 µm		100 µm	
RF Ports per Site	32		32		32		32		32		32	
Sites being probed together [5]	8		8		8		8		8		8	
Total # of RF Ports	256		256		256		256		256		256	
5G												
Wafer Test Frequency (GHz)	45 GHZ		45 GHZ		73 GHZ		73 GHZ		73 GHZ		73 GHZ	
RF Pad Geometry	Solder Balls		Solder Balls		Solder Balls		Solder Balls		Solder Balls		Solder Balls	
Pad Size (um)	100 µm		100 µm		70 µm		70 µm		60 µm		60 µm	
Pad Pitch (um)	150 µm		150 µm		125 µm		125 µm		100 µm		100 µm	
RF Ports per Site	32		32		32		32		32		32	
Sites being probed together	2		8		8		8		8		8	
Total # of RF Ports	64		256		256		256		256		256	

[1] In lieu of tighter pitch more staggared rows being used.
[2] CCC per IMSI
[3] WLCSP or eWLB packaging
[4] Parallelism is pushing capability on largest probe card
[5] The x8 is not possible today, but has been requested

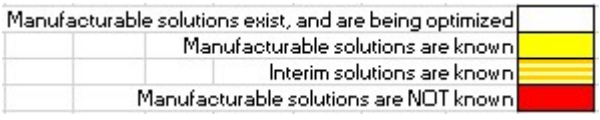


Table 2: Wafer Probe Technology Requirements (part 2)

Year of Production	2018	2019	2020	2021	2026	2031
Optical Probe						
Optical ports per site	4	12	48	96	192	192
Minimum pitch between fibers (um)	127	127				
Fiber optical alignment accuracy (Multi-Mode)	< 5um	< 5um	< 5um	< 5um	< 5um	< 5um
Fiber optical alignment accuracy (Single-Mode)	< 0.1um	< 0.1um	< 0.1um	< 0.1um	< 0.1um	< 0.1um

DRAM												
Wirebond - inline pad pitch	55		50		50		45		40		40	
I/O Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Wirebond	40	55	40	50	35	40	35	35	35	35	35	35
Sacrificial Pads	45	50	45	50	40	40	40	40	40	40	40	40
Wafer Test Frequency for Sort(Hz)												
Frequency(Hz)	250M		250M		400M		400M		400M		400M	
Shared Signal Line Test Frequency(Hz)	125M		125M		200M		200M		200M		200M	
Minimum pulse width	2.0nS		2.0nS		2.0nS		2.0nS		2.0nS		2.0nS	
At Speed Wafer Test												
Test Frequency(Hz)	2.2G		3.2G		3.2G		3.2G		4G		4G	
Probe Tip Diameter	8.5		8.5		8.5		8.5		8.5		8.5	
Probe Force(gf) - at recommended overdrive	2.5		2.5		2.5		2.5		2.5		2.5	
Size of Probed Area (mm2)	100% of wafer		100% of wafer		100% of wafer		100% of wafer		100% of wafer		100% of wafer	
Number of Probe Points / Touchdown - Memory	120000		130000		150000		150000		200000		200000	
Maximum Current (mA)/pin	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage
	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001
Maximum Resistance (Ohm)	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series
	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3
Probe test temperature range	-45	150	-45	150	-45	175	-45	175	-45	175	-45	175

NAND												
Wirebond - inline pad pitch	80		80		80		80		80		80	
I/O Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Wirebond	50	65	50	60	50	60	50	60	45	55	45	55
Wafer Test Frequency for Sort(Hz)												
Wafer Test Frequency(Hz)	100M		100M		133M		133M		266M		266M	
At Speed Wafer Test												
Test Frequency(Hz)	400M		600M		600M		600M		800M		1G	
Probe Tip Diameter	10		10		10		10		10		10	
Probe Force(gf) - at recommended overdrive	3		3		3		3		3		3	
Size of Probed Area (mm2)	100% of wafer		100% of wafer		100% of wafer		100% of wafer		100% of wafer		100% of wafer	
Number of Probe Points / Touchdown - Memory	80000		80000		80000		80000		80000		80000	
Maximum Current (mA)/pin	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage
	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001
Maximum Resistance (Ohm)	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series
	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3

Table 2: Wafer Probe Technology Requirements (part 3)

Year of Production	2018		2019		2020		2021		2026		2031	
LCD driver Products												
Bump - inline pad pitch	18		18		16		16		14		12	
Bump - stagger pad pitch	13		10		8		8		6		4	
I/O Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Inline	11	50	11	50	10	50	10	50	10	40		
Stagger	15	30	15	30	12	40	12	40	10	25	10	25
High speed I/O pin freq (Hz)	2.0G		2.5G		3.0G		3.5G		5G		6.5G	
Probe needle structure	Cantilever		Cantilever/Vertical		Cantilever/Vertical		Cantilever/Vertical		MEMS		MEMS	
Probe Tip Diameter (mil)	0.4		0.4		0.3		0.3		0.2		0.2	
Probe Force(gf)	2		2		2		2		2		2	
Size of Probed Area (mm2)	4900		5600		6800		6800		12000		16000	
	4000		4000		4000		4000		8000		12000	
Maximum Current (mA)/pin	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage
	300	<.001	300	<.001	300	<.001	300	<.001	300	<.001	300	<.001
Maximum Resistance (Ohm)	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series	Contact	Series
	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3	<0.5	<3
CMOS Image Sensor												
Wirebond - inline pad pitch	95		90		80		80		75		70	
I/O Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
Wirebond	70	70	60	70	60	65	60	65	55	60	45	50
WLCSP	46	100	46	100								
WLP/ WLO / WLC (TSV process)	55	55	40	40	40	40	40	40	40	40	30	30
	200M		200M		400M		400M		400M		400M	
	2.0G		2.5G		3G		3G		5G		6.5G	
Probe needle structure	Vertical/MEMS		Vertical/MEMS		Vertical/MEMS		Vertical/MEMS		MEMS		MEMS	
Probe Tip Diameter Wirebond (mil)	0.5		0.5		0.4		0.4		0.4		0.3	
Probe Force(gf)	3		2		2		2		2		2	
Size of Probed Area Diameter Φ (mm) [6]- Visible light	300x300		300x300		300x300		300x300		300x300		300x300	
Number of Probe Points/Touchdown	3200		5000		10000		10000		20000		35000	
Maximum Current (mA)/pin	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage
Visible light sensor	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001
IR sensor	1000	<.001	1000	<.001	1200	<.001	1200	<.001	1200	<.001	1200	<.001
Visible light sensor [7]												
Maximum Current (mA)/pin	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage	Probe Tip	DC Leakage
Visible light sensor	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001	250	<.001
Parametric (Process monitor)												
Inline pad pitch	40		40		40		40		40		40	
Inter-row pad pitch	35		35		35		35		35		35	
Pad Size (µm)	X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
In line pads	20	20	20	20	20	20	20	20	20	20	20	20
Probe Tip Diameter	6		6		6		6		6		6	
Number of pad rows	2		2		2		2		2		2	
Probe Force(gf) - at recommended	2		2		2		2		2		2	
Number of Structures /Touchdown	8		8		8		8		8		8	
Maximum Leakage (pA)/pin (10V / 1 Sec	0.2		0.2		0.2		0.2		0.1		0.1	
Maximum Contact resistance	0.3		0.3		0.3		0.3		0.3		0.3	
Maximum Path resistance (Ohms)/pin	3		3		3		3		3		3	
Maximum Probe temperature Range	-50	200	-50	200	-50	200	-50	200	-50	200	-50	200
Maximum test Frequency (GHz)	3		3		3		3		3		3	

Notes:
[6] Probe area is limited by optical system
[7] The trend of NIR image sensor circuit design will be similar to visible image sensor. So, the wafer probe table doesn't need to seperate addional items for NIR image sensor

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 the DUT. As semiconductor design 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 for these higher requirements. Current models show difficulty in making sockets for high ball count devices and achieving I/O bandwidths of > 20GHz.

Socket Trends

Table 1 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 their 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 and stroke, and to 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, mainly used for testing BGA-DRAM devices, 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 be large enough to guarantee strength and durability and the length should be long enough to maintain sufficient travel under compression. 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 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 2016, 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 uses with high frequency device test due to its low inductance and low contact height, but compression travel is limited. Conductive rubber will meet the fine-pitch requirement in the roadmap, but it is difficult to reduce contact force without decreasing the compression travel.

Table 1: Test Socket Technology Requirements

Year of Production	**2018**	**2019**	**2020**	**2021**	**2026**	**2030**
TSOP – Flash (NAND) – Contact blade [1]						
Commodity NAND Memory						
Lead Pitch (mm)	0.3	0.3	0.3	0.3	0.3	0.3
Data rate (MT/s)	133	133	266	266	266	266
Contact blade						
Inductance (nH)	5-10	5-10	5-10	5-10	5-10	5-10
Contact Stroke (mm)	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3
Contact force (N)	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3
Contact resistance (m ohm)	30	30	30	30	30	30
Slit width (mm)	0.17	0.17	0.17	0.17	0.17	0.17
BGA – DRAM – Spring Probe [2]						
Commodity DRAM (Mass production)						
Lead Pitch (mm)	0.25	0.25	0.2	0.2	0.1	0.1
DRAM RM GT/S	5.3	5.4	6.4	6.4	8.5	8.5
Spring Probe						
Inductance (nH)	0.2	0.2	0.15	0.15	0.15	0.15
Contact Stroke (mm)	0.2	0.2	0.2	0.2	0.2	0.2
Contact force (N)	<0.2	<0.2	<0.2	<0.2	<0.2	<0.2
Contact resistance (m ohm)	100	100	100	100	100	100
BGA – SoC – Spring Probe (50 ohm) [3]						
Lead Pitch (mm)	0.15	0.15	0.15	0.15	0.15	0.15
I/O data (GT/s)	56	56	56	56	56	56
Spring Probe (50 ohm)						
Contact force (N)	<0.2	<0.2	<0.2	<0.2	<0.2	<0.2
Contact resistance (m ohm)	50	50	50	50	50	50
BGA – SoC – Conductive Rubber [4] [5]						
Lead Pitch (mm)	0.25	0.25	0.25	0.25	0.25	0.25
I/O data (GT/s)	56	56	56	56	56	56
Conductive Rubber						
Inductance (nH)	0.15	<0.1	<0.1	<0.1	<0.1	<0.1
Contact Stroke (mm)	0.15	0.15	0.15	0.15	0.15	0.15
Contact force (N)	0.1	0.1	0.1	0.1	0.1	0.1
Contact resistance (m ohm)	50	50	50	50	50	50
Thickness (mm)	0.5	0.5	0.5	0.5	0.5	0.5
QFP/QFN –SoC – Contact blade + Rubber [6]						
QFP/QFN –SoC						
Lead Pitch (mm)	0.3	0.3	0.3	0.3	0.3	0.3
Data rate (GT/s)	20	40	40	40	40	40
Contact blade + Rubber						
Inductance (nH)	0.15	＜0.1	＜0.1	＜0.1	＜0.1	＜0.1
Contact Stroke (mm)	0.2	0.2	0.2	0.2	0.2	0.2
Contact force (N)	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3	0.2-0.3
Contact resistance (m ohm)	30	30	30	30	30	30

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

Notes for Table 1:

[1] For pitches less than 0.3mm contactor molding becomes difficult due to the thin wall thickness between pins.

[2] For higher performance, a shorter probe spring is required which shortens the contact stroke. In 2018, the contact stroke will be 0.2mm so the contact resistance will be unstable.

[3] The spring probe must be coaxial for high-speed test. 20GT/s cannot be supported with finer pitches.
[4] Ball height is expected to change over the roadmap but amount of change is not known.
[5] A contact stroke of 0.15mm was assumed with a 0.5mm rubber thickness. For high ball count devices, the contact pressure has been lowered.
[6] Contact stroke will be the biggest concern to achieve 40GHz bandwidth in 2019.
[7] Bandwidth threshold is -25dB Crosstalk of GSSG model EM simulation.
[8] Shorter spring probe is required to reduce inductance, but it is challenge due to its mechanical structure.
[9] Conductive rubber type has advantage in the pitch and low inductance, but the challenge is contact stroke.

Contact blade + Rubber, generally used for testing QFP/QFN high frequency SoCs, 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 life cycle is normally shorter than for a Contact Blade type contact.

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.

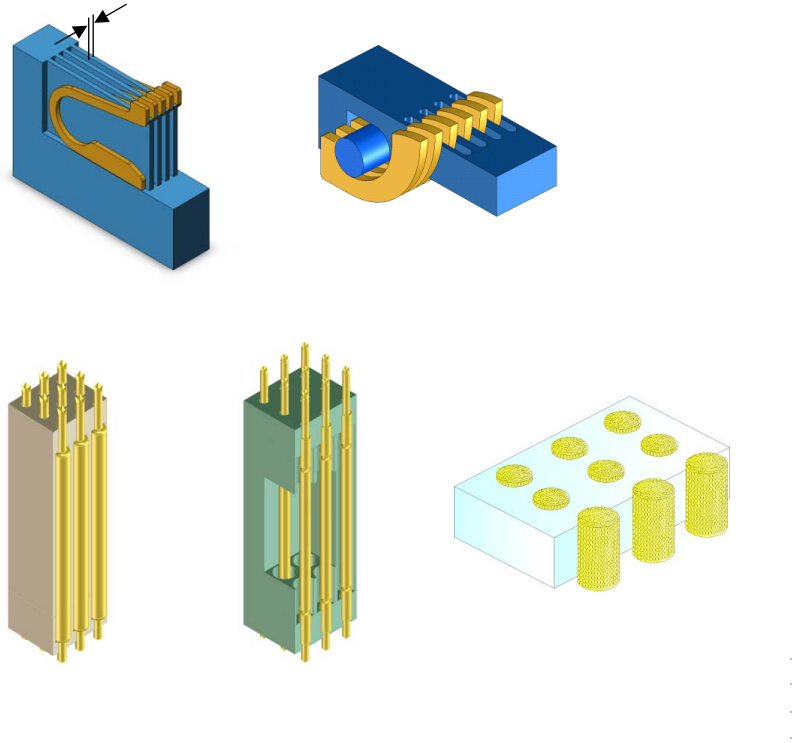


Figure 1: Contactor Types

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 over the roadmap timeframe are expected to exceed 20 GHz, which will greatly challenge impedance matching and 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 ensure 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 stringent mechanical requirements. To handle these high-frequency signals,

the user has to carefully consider the signal integrity of the overall test system including board design/components/socket.

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 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-direction), 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.

Section 8: System Level Test

System level test (SLT) refers to exercising the components of a system as an integrated whole to validate correct system operation for its intended end-use applications. We confine our notion of systems to be primarily electronics-based. As a whole, a system is comprised of both physical hardware and software programs. Electronic hardware scales from single chips, multi-die integrated packages, printed circuit boards (PCB), on up to racks of PCBs. For interacting with the physical world, some system hardware components can fall into non-electronic domains such as mechanical, optical, chemical, and biological. Software includes firmware, device drivers, operating system, and applications. The steady march of Moore's Law in semiconductors has enabled the creation of ever more complex systems. These systems are finding new applications and winding their way ever deeper into our daily lives. This trend is expected to continue at an accelerated pace as exemplified by smart cars, homes, cities, factories, healthcare, agriculture, etc. loosely aggregated under the umbrella term internet-of-things (IoT), enabled by high-data-rate 5G wireless communications.

The purpose and use of SLT differs depending on the perspective of the adopter: a component supplier versus a system integrator. Running SLT is a way for the component supplier to minimize defect returns from the system customer. Having SLT enables earlier time-to-market for the component while buying more time to improve ATE-based production test. The system integrator uses SLT to assess the quality of incoming components from multiple sources as part of an overall system validation process. Complex systems can have many configuration parameters which can be tuned by software to achieve optimal performance. Variability of components and their interactions may require the system integrator to run SLT to tune each system individually.

This section addresses a number of topics addressing SLT:

1. What's driving the trend toward more use of SLT?
2. What are the various SLT approaches and flows being used today?
3. What are the challenges and issues in using SLT?
4. What are the opportunities for improving SLT?
5. How will future system applications affect SLT development?

TREND TOWARD SLT ADOPTION

In the early days of simpler electronics, functional test predominated. As designs grew in complexity, propelled along the trajectory guided by Moore's Law, functional testing became untenable due to its high development cost and inadequate fault coverage to meet quality requirements. Scan-based structural test gained prominence since it enabled efficient automatic test pattern generation (ATPG) to achieve high fault coverage. As process dimensions shrank, fault models that underlie structural test kept pace with increasingly complex defect behaviors by evolving from stuck-at to transition-delay to cell-aware, delivering the required quality levels albeit with significantly higher test pattern size. However, the requirements are shifting yet again and the benefits of structural test may be hitting a limit. There is much anecdotal evidence that functional testing in the form of SLT is gaining more usage to augment structural test, in order to catch so-called "marginal" defects missed by ATE-based production test.

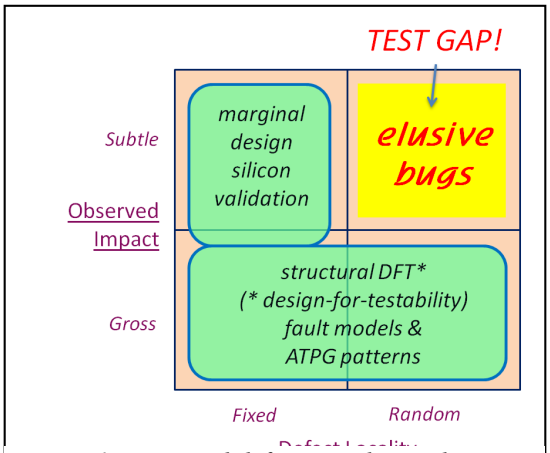


Figure 1. Marginal defects in advanced process nodes escape through test gap left open by existing test methods.

The notion of marginal defects arose in connection with inherent limits of optical lithography and process variation control in the manufacturing of devices at advanced nanometer nodes. Traditional test and debug methods are successful at catching defects with a gross observable impact that occur at random locations, or more subtle defects that are systematically localized. Left uncovered is a test gap for marginal defects that are both subtle in impact and randomly occurring (Figure 1). Exacerbating the issue is the growing divergence between test-mode versus in-system operation for today's multi-core system-on-chip (SoC) designs. For production test efficiency, complex clock and power domain interactions found during mission-mode are typically not exercised on ATE. Production test conditions can miss marginal defects which escape and cause failures under system operating conditions. These defects are difficult to resolve between field and factory, earning them the label of no-trouble-found (NTF).

Scan-based structural test is only applicable to synchronous digital logic. Specialized test schemes, usually executed via some form of BIST, exist for memory, high-speed I/O, and some analog functions. However, at the system level, defects that affect interactions among the constituent components are not well-covered by isolated

device/IP-level tests. Adding software to the mix further creates the possibility of soft failures when marginal devices are exercised under certain application scenarios that induce stress, triggering "application-enabled" defects. The trend toward product miniaturization in many end-use applications has driven packaging innovations to create multi-die system-in-package (SiP) solutions. For these advanced packages, testing to assure the structural and functional integrity of individual block-level parts may not be sufficiently comprehensive.

The test gaps identified above form the driving impetus towards SLT adoption. For the system integrator, performing SLT is par for the course, since the system is the end product. However, for the component supplier, SLT has not always been a necessity. To some extent, component functional test (in addition to structural test) performed on the ATE partially fulfills the purpose of SLT and is often good enough to ensure end-user system quality. But when both component and system complexities reach beyond a certain threshold, typical ATE-based functional test can no longer mimic the compendium of system-level interactions. For example, it is difficult to boot the entire Android operating system on a mobile phone SoC on the ATE. For the component supplier, it can be more cost-effective to build specialized SLT environments that can better mimic the end-user system and run a multitude of application scenarios to catch potential system-level defects. Thus, providers of sophisticated SoC and SiP components are seen to be early adopters of SLT.

SLT APPROACHES AND FLOWS

SLT practices vary widely, dictated by factors such as:

- System integrator vs. component supplier role
- Product market segment, unit volume, and lifetime
- Product quality and reliability requirements
- Competitive pressure in terms of time-to-market (TTM) and time-to-volume (TTV)
- Product profit margin

These factors determine how a company develops its own unique SLT approach, which can be characterized in a number of aspects:

- ATE vs. special SLT equipment
- Structural and functional SLT content
- SLT run time
- Full vs. sampled SLT
- SLT thermal, voltage, and cycling conditions
- SLT failure diagnosis

As shown in the example illustration (Figure 2) for a high-volume consumer mobile phone SoC, SLT today is inserted as the last step in the production test flow after wafer probe die test (WP), and packaged chip final test (FT). Typically, SLT is performed on special equipment distinct from WP and FT ATE. In this case, the SoC test board is comprised of the entire mobile phone system where the software stack from firmware to user applications can be run. The SLT flow is controlled by a PC-based test station with handler to load/unload multiple boards for parallel testing. If system components employ various scan-based test access standards such as IEEE 1149.x, 1500, and 1687 in a consistent and compatible manner, structural aspects of SLT to assess system assembly and connectivity can be executed more efficiently than relying purely on functional exercises. Reuse of modular BIST capabilities at the system level is another benefit.

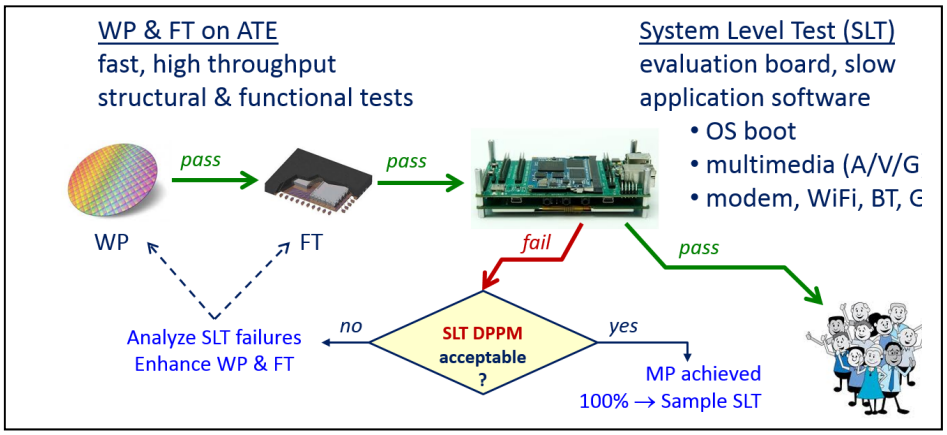


Figure 2. SLT flow for high-volume mobile phone SoC, initially 100% then reduced to sampling mode once WP/FT test quality has been enhanced to achieve acceptable SLT DPPM.

Typical SLT run times can vary from seconds to tens of minutes for high-volume products. For low-volume and long-lifetime products requiring high reliability, the system integrator may run SLT for hours

or even days. A component supplier can utilize SLT in the production test flow either as a permanent or a temporary step. Due to the relatively long run time of SLT, including it fully in production requires higher-cost SLT test equipment that can support very high concurrency in order to meet volume throughput demands. Full SLT typically applies to higher-end products selling at a premium, accompanied by higher user expectations. The cost structure of high-volume but lower-end products cannot support full SLT. Instead, to meet early TTM, full SLT is only utilized during initial production to deliver products of good-enough quality. Effort is spent on analyzing SLT failures to enhance WP & FT functional tests with the goal of reducing SLT DPPM. Once SLT DPPM reaches a level corresponding to acceptable customer quality expectations, SLT is shifted from full mode to occasional sampled monitoring. Quickly removing SLT as a throughput bottleneck enables the production TTV goal to be met.

Based on the experience of product deployment and in-field failures encountered, the system integrator may require the supplier to augment SLT with further preventive measures. For example, if power-up failure is noticeably significant, adding multiple cycles of OS boot under thermal stress to SLT may help flag problematic components. For safety-critical applications, it may require a serious and continuous effort in failure diagnosis to identify root-causes, leading to corrective actions including potential design and manufacturing changes.

There is increasing interest to employ big data analytics across the WP-FT-SLT flow to improve process efficiency and accuracy. These usually involve finding signature correlations among test measurements and failing behaviors which can be turned into predictive decisions as part of an adaptive test flow. Another approach being explored by industry is to move SLT upstream to WP, and for SLT to be ATE-based. This approach has two aspects: (1) replicate the full SLT environment and (2) add more SLT-like tests for the component. Being able to perform SLT at the wafer and/or die level will help provide Known-Good-Die (KGD) prior to subsequent multi-die integration steps. However, full replication of a highly complex SLT environment can be very expensive and is rarely justified except in special cases such as critical warfare electronics in a fighter jet. If system complexity is not high, it is possible to create the full system on the ATE load-board and run SLT using a low-cost tester (Figure 3). Besides the device-under-test (DUT), the load-board system includes flash memory to store SLT program components, a simple CPU as test controller, DRAM for system memory, and SLT self-checking circuitry to send DUT results to ATE. Such a "system functional test" approach has been used successfully to achieve exceptional quality for a line of low-cost consumer optical drive ICs.

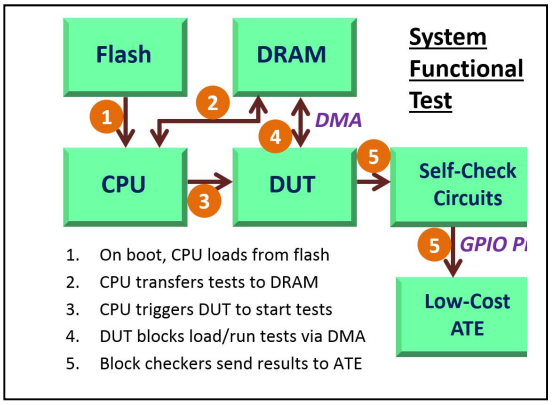


Figure 3. ATE-based system functional test.

Stand-alone component functional test executed on ATE can take advantage of on-chip processor and memory to run light-weight system programs to partially achieve the aims of SLT. Protocol-Aware (PA) testing is a recent development to ease the testing of modern device-to-device high-speed I/O interfaces that have complex non-deterministic protocols. It overcomes severe productivity challenges associated with traditional low-level fixed-timing ATE programming, and enables the test engineer to develop patterns at a higher level of system abstraction. Such patterns can cover more realistic and dynamic I/O protocol transactions to achieve better coverage.

SLT CHALLENGES AND ISSUES

Though its ability to catch defect escapes from WP/FT can be readily demonstrated, SLT's primary issue is the lack of quantifiable metrics to assess its effectiveness. Unlike structural test, SLT does not have well-developed fault models and an established statistical framework to link fault coverage to quality level. The fundamental reason underlying this widely-recognized issue has to do with the much more complex nature of SLT failures. Clearly, defects that manage to escape ATE test are harder to catch by definition. Reasons for failure in SLT are more likely to be murky and involve multiple contributing factors. For example, a compound failure may involve a piece of firmware code operating two interacting marginal devices in a manner that creates an unanticipated scenario that pushes system operation into an unsafe power-thermal zone. When running SLT, a common failure response is for the program to hang and time out without any indication of location and cause. Because SLT lacks the controllability and observability advantages of scan, there can be a long lag as in millions of clock cycles between when an error first occurs to when it propagates to a distant location where system failure is eventually observed.

Needed are methods of systematic SLT failure analysis to understand in more detail the defect mechanisms, error sequences, and associated statistical properties. That understanding then drives the development of effective test and

defect coverage strategies. Many elements of an SLT program are naturally derived from design verification where pattern development is guided by coverage of more abstract entities such as application scenarios, transactions and functions. Missing is a link from the higher-level abstraction to the underlying hardware in terms of covering where defects are more likely to occur, and de-emphasizing areas that are not problematic or already covered by WP/FT tests. Fault grading using traditional gate-level simulation is ineffectual due to severe run-time/capacity limitations and inability to model the entire system encompassing software, behavioral constructs, and non-digital hardware. Uniform treatment of low-level structural fault models without system context may waste computing resources and convey a false sense of adequate coverage. SLT program development, failure analysis, and enhancing ATE functional test remain largely a manual and inherently inefficient endeavor, contributing to its high cost of adoption.

In terms of test equipment, building an ATE-based full SLT environment in CP/FT faces ever higher barriers as system complexity increases. Even if it could be done, the cost is likely to be prohibitive in most situations. Thus, component manufacturers are more likely to choose dedicated SLT equipment used in an additional test insertion. Throughput then becomes the primary issue. A typical component SLT test station today is configured to handle four to six units in parallel. In order to meet high throughput demands under the constraint of long SLT run times, the manufacturer is forced to add more testers, more operators, and more factory floor footprint. Massively parallel SLT architectures are needed to raise throughput measured in unit-per-hour (UPH) by an order of magnitude or more. Being able to perform thermal stress testing is also gaining importance with more SoC devices implemented in FinFET nodes where self-heating is a concern as well as managing heat dissipation in tightly packed multi-die SiP components . Consequently, the SLT environment must support the evaluation of system thermal integrity.

OPPORTUNITIES TO IMPROVE SLT

Though SLT methods and practices are developed in a somewhat ad hoc fashion based on circumstances unique to each organization, many opportunities exist to introduce more formal and systematic approaches that can be shared across the industry to create commonality. Over time, an ecosystem can be built up based on standards, EDA tools, and equipment to enable more purpose-driven and efficient SLT flows.

As groundwork, research should be conducted, both theoretical and experimental, to develop fault models that embody emergent properties of complex component interactions. Complex systems are inherently hazardous since it's nearly impossible to anticipate all scenarios of failure, especially when software (rarely defect-free) and human behavior are integral to the system. Even when anticipated, failure mitigation measures constrained by cost may imply the acceptance of calculated risks during system design, calculations that may contain faulty assumptions. A system-level fault model should encompass, as a minimum, physical mechanisms of variation, noise, and aging, all the way up to application-induced stress scenarios that expose marginalities leading to heightened probability of failure. Having such a fault model will enable further development of system-level capabilities for fault grading, ATPG, and statistical correlation of coverage vs. quality. Unlike traditional device-level fault models, system-level faults may be difficult to explicate by single or simple root-causes. Thus, new approaches to fault diagnosis that consider compound causal relationships need to be investigated.

Since SLT is primarily function-oriented (as opposed to structural), much of the test patterns are procured from design verification (DV) which has also been facing growing design complexity as a serious challenge, to the point that DV now dominates the cost of SoC development. Starting from manually written tests to verify functionality, impressive progress has been made in the development of coverage metrics to assess quality, automated test generation such as constrained-random to improve coverage, and the standardized Universal Verification Methodology (UVM). Though successfully deployed widely, UVM has run into reuse scalability limitations: first in moving up the scope of integration beyond the level of verifying IP blocks to verifying complete systems with embedded software; and second in extending to non-simulation verification platforms such as in-circuit emulation, FPGA prototyping, and post-silicon bring-up.

The latest push by the DV community to overcome these limitations is to develop a common abstract model of verification intent named the Portable Test and Stimulus Standard (PSS). As shown in the PSS diagram (Figure 4), block-level abstract models can be combined directly to construct higher-level system verification tests. The same abstract stimulus model can drive test generation tools to target multiple verification platforms. PSS offers an opportunity for SLT to capitalize on major investments being made by EDA tool providers and users to realize a broad vision of verification reuse. As illustrated, the PSS abstract model can be extended to include system-level fault model and stress scenarios. Test generation tools can be enhanced to traverse the system activity graph, solving a set of constraints derived from SLT perspective to generate test patterns which can be run on production SLT

platforms. The test community should participate in the PSS Working Group (PS-WG) to promote the inclusion of SLT in its scope.

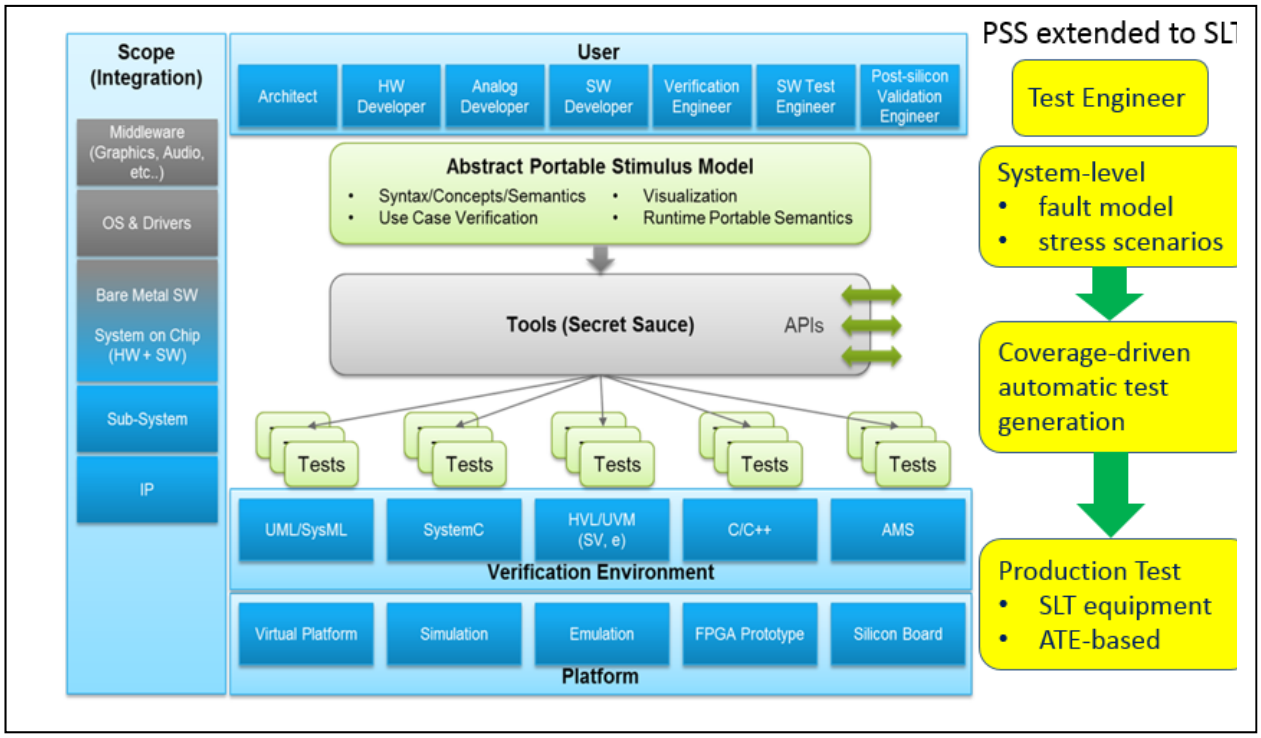


Figure 4. PSS serves as a framework to construct SLT flow that includes fault modeling and test pattern generation.

As with structural test, design modifications to increase controllability and observability should enable more effective and efficient SLT. Due to the close relationship between SLT and post-silicon validation, SLT design-for-testability (SLT-DFT) can draw on the latter's design-for-debug (DFD) practices as well as on-going industry and academic research to advance the state-of-the-art. In SoC designs, it is common to find on-chip instrumentation to support (1) profiling of performance aspects such as power consumption and speed, (2) monitoring of power grid noise and thermal excursions, and (3) debug and diagnosis of silicon failures by signal tracing. Instruments in hardware tend to favor observation so as to not disrupt normal system operation. Software also plays a crucial role in configuring initial silicon states, setting triggers for tracing, transporting logged data off-chip, and analyzing data for high-level debug. SLT can utilize data captured by embedded instruments during system operation to identify vulnerabilities in the silicon and improve failure diagnostic resolution. Of note is the recent appearance of commercial IP for embedded analytics (UltraSoC) and the growing popularity of USB as the device-under-debug to host interface. USB has the advantages of ubiquitous presence in SoC designs and high data transfer rate when compared to IEEE 1149.1 JTAG or IEEE 1687 IJTAG.

Another promising approach called Quick Error Detection (QED) aims to drastically reduce error latency via software transformations applied to existing validation tests. It has the option to insert hardware checkers (incurring a small overhead) to accelerate test execution time. Reduced error latency allows more precise localization of both functional and electrical bugs. QED transformations target bug scenarios abstracted from analysis of a diverse set of actual failures found in commercial multi-core SoCs. Thus, bug scenarios can be viewed as a kind of system-level fault model and the goal of QED is to generate tests to achieve high coverage of bug scenarios.

Embedded instruments provide a wealth of internal device data which can be exploited to optimize the entire production test flow. Adhering to the general concept of adaptive test (Fig. 5), upstream WP/FT data can drive downstream SLT decisions. For example, typical SLT failure rates (so called SLT DPPM) are quite low. Even a high SLT DPPM of 10,000 means 99% of devices will pass SLT. If one can predict based on WP/FT data even a modest portion, say 60%, of devices that are very likely to pass SLT, then limiting SLT to run only on the other 40% will realize large test cost savings. Furthermore, if WP/FT data can indicate potential weak spots in each device, SLT programs can be uniquely customized to target the weak spots and achieve more effective SLT screening, thus realizing higher quality. Customized SLT does require new flexibility for on-the-fly test program construction. To enable adaptive SLT, a learning phase is needed to find strong data correlations between WP/FT and SLT stages. The kinds of test data and correlation methods are active topics of investigation. For example, typical WP/FT test data may lack sufficient information. One approach applies delay-test scan patterns under non-destructive stress conditions on ATE to generate fine-grain localized internal health signatures which can be correlated against SLT pass/fail data to obtain predictive rules.

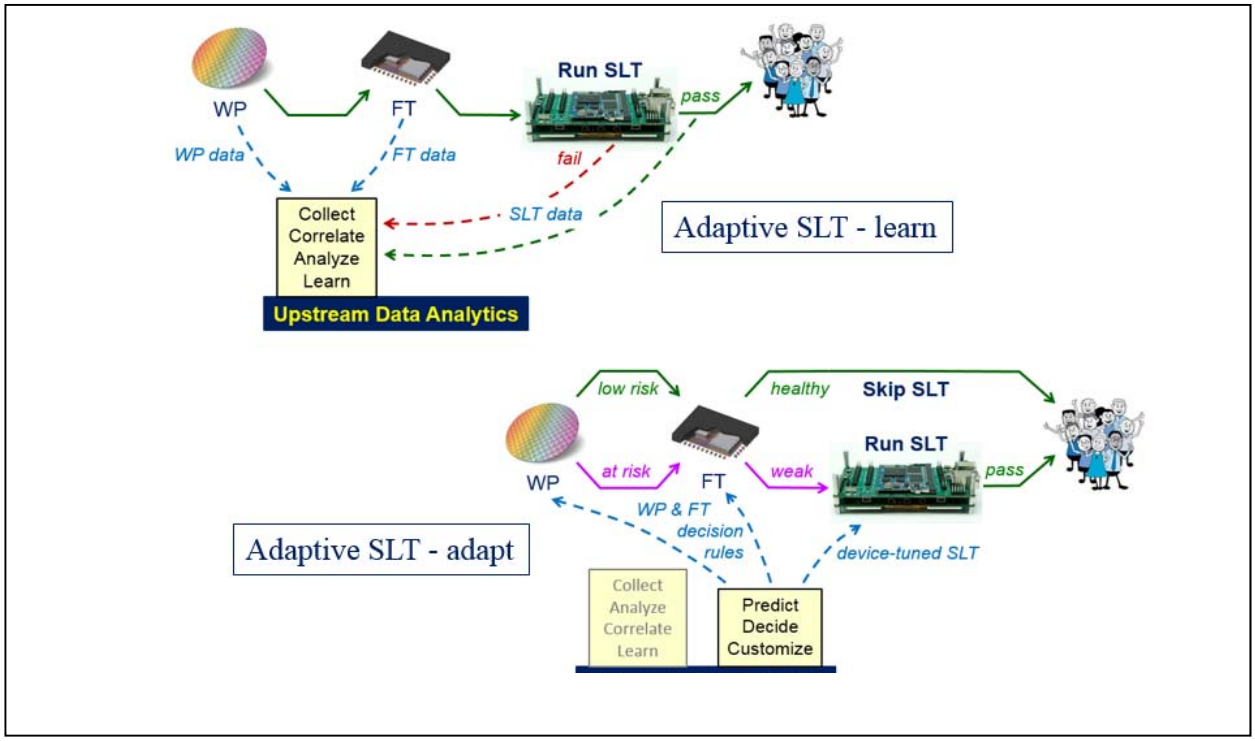


Figure 5. Adapt SLT flow using data analytics to optimize for cost efficiency and test effectiveness.

Sophisticated fault models such as cell-aware are being used today to target subtle system-level faults in WP/FT at great expense in test pattern count. Perhaps a more effective SLT can help off-load WP/FT to achieve an optimal and balanced production test flow. Creating a data loop between component suppliers' WP/FT and system integrators' SLT will require secure data exchange standards and means to protect sensitive proprietary information. Real-time data analytics also create opportunities for test equipment to add value by adopting machine learning accelerators.

IMPLICATIONS FOR SLT FROM FUTURE APPLICATIONS

The modern high-end car is a marvel of systems engineering with hundred-plus electronic control unit (ECU) sub-systems communicating over multi-level networks, all coordinated by software that can reach over 100M lines of code. Yet this level of complexity pales in comparison to the transformation currently underway that will take us into the future of self-driving electric vehicles. Withal the complexity lay the uncompromising requirements for reliability, safety, and resilience over an expected lifetime lasting ten years or more. This future robot on wheels brings together all of the key technological advances in 5G communications, artificial intelligence (AI) computing, and IoT applications. Needless to say, SLT will have to undergo a similar transformation to keep pace. The implications for SLT as applied to the robot car also extend to systems in other domains undergoing similar versions of the "smart" evolution.

Achieving the full potential of self-driving vehicles will require them to be part of an intelligent transportation system where wireless vehicle-to-vehicle/infrastructure (V2X) networking coordinates dynamic local operations. The key enabler is 5G millimeter wave (mmWave) beamforming which provides the fast network acquisition, short latency, and high data rate necessary for ad hoc network formation. Furthermore, due to space and weight constraints, the plethora of additional sensors dictate a portion of in-vehicle networks to also go wireless in support of various IoT-connected services. For 5G mmWave, the traditional cabled and controlled-environment testing approach no longer suffices. Instead, SLT has to be over-the-air (OTA) to check each individual antenna array element as well as the entire array's overall adaptive beamforming performance under realistic noisy conditions.

AI computing introduces software that is, in part, not written by humans, but instead trained by data in the form of deep-learning neural network (DNN) connection weights. The inexact nature of AI computation introduces a testing conundrum: Is a system failure due to defects in the DNN hardware implementation, or to insufficient DNN training that missed rare corner cases in the input data domain? Heavy AI computation also poses a power consumption challenge, especially if power comes mainly from stored energy, e.g., battery. Already inefficient conventional von-Neumann processors are being overtaken by highly parallel arrays of dedicated processing cores with local memory and data-flow accelerators. On the horizon are even more energy-efficient computation techniques spanning the spectrum from devices to algorithms such as analog, memristive, asynchronous, neuromorphic, and approximate. As structural testing techniques for synchronous digital designs do not easily extend to the above, it may be even more incumbent on SLT to fill the testing gap.

Future smart and energy-efficient systems may be architected from the ground up to be inherently redundant and error-tolerant much like the human brain. System construction will be based on complexity management principles of hierarchy, modularity, and abstraction. Such systems may have to operate continuously under extreme conditions with natural aging and degradation over time, thus requiring self-monitoring/testing/healing capabilities beyond T0 (defined as time of product shipment). Though component SLT may still be done as part of a less complex subsystem, SLT itself will evolve into the more encompassing built-in capability of self-aware systems.

Summary

SLT's rise in importance is inevitable as advances in semiconductor process and heterogeneous integration technology lead to ever more complex products. Its practice extends from system integrators to system component suppliers. Subtle software-hardware component interactions defy current production test screens, resulting in end-system application-dependent failures. Though SLT can help as an additional screen, there is much opportunity to improve its efficiency and effectiveness in methodology, automation, and test equipment. SLT also fits naturally within the overall adaptive test flow by utilizing cross-test data analytics. The test community should take advantage of the similarity between SLT and DV (particularly post-silicon validation) to push for PSS extensions to include SLT. Driven by upcoming 5G/IoT/AI-enabled transformations touching every aspect of our lives, SLT will most likely evolve into the self-monitoring/testing/healing capability of future smart systems.

References

S. Biswas and B. Cory, "An industrial study of system-level test," IEEE Design & Test, vol. 29, no. 1, pp. 19–27, February 2012.

P. G. Ryan et al., "Process Defect Trends and Strategic Test Gaps," International Test Conference, October 2014.

Z. Conroy et al., "A practical perspective on reducing ASIC NTFs," International Test Conference, November 2005.

A. Jutman, M. Sonza Reorda and H.-J. Wunderlich, "High Quality System Level Test and Diagnosis," Asian Test Symposium, November 2014.

H. H. Chen et al., "Predicting System-Level Test and In-Field Customer Failures using Data Mining," International Test Conference, September 2013.

Astronics Test Systems, "Massively Parallel System-Level Testing – The new Paradigm for Semiconductor Quality Control," https://www.astronics.com/test-systems/whitepaper---system-level-testing-for-semiconductors

Accellera Systems Initiative, Portable Stimulus Specification Working Group, http://www.accellera.org/activities/working-groups/portable-stimulus

P. Mishra et al., "Post-Silicon Validation in the SoC Era: A Tutorial Introduction," IEEE Design & Test, vol. 34, no. 3, pp. 68–92, May 2017.

M. Sadi et al., "An All Digital Distributed Sensor Network Based Framework for Continuous Noise Monitoring and Timing Failure Analysis in SoCs," Asian Test Symposium, November 2014.

D. Lin et al., "Effective Post-Silicon Validation of System-on-Chips Using Quick Error Detection," IEEE Trans. on Comp.-Aided Design of Integrated Circuits and Systems, vol. 33, no. 10, pp. 1573–1590, October 2014.

C. G. Shirley et al., “Board Manufacturing Test Correlation to IC Manufacturing Test,” International Test Conference, October 2014.

H. H. Chen et al., “Statistical Techniques for Predicting System-Level Failure using Stress-Test Data,” VLSI Test Symposium, April 2015.

H. H. Chen, “Data analytics to aid detection of marginal defects in system-level test,” Int. Symposium on VLSI Design, Automation and Test, April 2016.

A. Smith, “5G and IoT create a market for mmWave,” https://www.electronicsweekly.com/news/design/communications/5g-and-iot-create-a-market-for-mmwave-2017-06/

Section 9: Adaptive Test

Adaptive Test is a set of IC manufacturing test methods enabling real-time optimization of the value of production test (in a fully automated way). These methods include the practice of using production test data to reduce/optimize test cost, improve product quality and reliability, drive yield improvements and improve product performance.

To effectively use Adaptive Test, additional development is needed in data infrastructure and data analytics, production test cell design, die traceability, and the coordination between IC manufacturers, IC test and assembly. The top challenges facing industry application of Adaptive Testing are listed in the portion of this section called "Adaptive Test Challenges and Directions."

There is a diverse set of Adaptive Test applications aimed at different product markets and requirements. Companies should target specific Adaptive Test applications aimed at their product requirements and target markets. Thus, the benefits of implementing Adaptive Test will vary based on type of product, manufacturing flow, volumes and technology.

This section provides:

1. A description of Adaptive Test and terminology used by its practitioners
2. Example applications of Adaptive Test as of 2019 and future opportunities
3. A list of challenges for the development and deployment of Adaptive Test.

To ensure the industry can fully exploit the benefits of Adaptive Test, this section describes:

1. The infrastructure requirements for test cells, data systems and device and system designs
2. A description of Adaptive Test challenges and the coordination needed between IC manufacturers, OSATs (Outsource Assembly and Test providers) and fabless system integrators.

Definition

Adaptive test comprises a set of methods for automatically changing manufacturing test conditions, manufacturing flow, test content, test limits, or test outcome to reduce test cost, increase outgoing quality and reliability, reconfigure parts, or collect data to further improve test and manufacturing. Adaptive Test strives to make these changes in a manner that does not significantly increase testing time or increase human involvement in test operations. The decisions on when and how to adapt the tests are made algorithmically by the tester, other test cell equipment, or an automatic data analysis system (given automation – the analysis time is significantly reduced compared with the traditional, engineering-intensive approach).

Adaptive Test Description

Adaptive Test is generally accepted as an advanced test strategy that can be used to achieve quality, yield, and cost goals that might not be reached by normal test methods. For example, Adaptive Test may modify a production test process in a number of ways:

1. **Test Conditions** (modifying test setup conditions or limits – such as voltage or clock frequency)
2. **Manufacturing Flows** (adding or deleting test insertions such as burn-in)
3. **Test Content** (modifying specific patterns or tests such as transition fault or IDDQ, respectively)
4. **Test Limits** (changing the pass/fail limits such as DC power or Vdd-min test specifications)
5. **Test Outcomes** (changing the binning or configuration of some die based on post-test analysis of the die's test results)

Adaptive Test applications are organized by when decisions are made to modify the test flow and to which device(s) the modified test flow are applied. The four most common categories are in-situ, feed-forward, feed-back and post-test. Figure 1 is a flow diagram depicting the relationships among these four categories.

1. **In-situ:** Data collected from the part being tested is used to modify the testing of the same device during the same test insertion. These methods include not only speed-binning and trimming calibration (which are not new), but using data from any test to modify test conditions or device settings for other tests for this specific device. For example, parametric data taken from on-product process monitoring structures may be analyzed – and the results used to drive subsequent test limits or test conditions or test-driven device reconfiguration.
2. **Feed-forward:** Data collected from a previous test step stage (e.g. probe, hot probe, burn-in) is used to change how the same parts are tested at a future stage. An example of the Feed-forward category are

statistical methods which identify 'risky' dice or wafers and selects these components (only) for burn-in, or "clean" dice that may be candidates for reduced testing.

3. **Feed-back:** Data collected from a previous part (or parts) is used to modify the tests or limits of different devices yet to be tested. Skipping some test patterns on high-yield wafers, adding more tests to low-yield wafers or refining statistical models used for die classification are examples of this category.
4. **Post-Test:** Data sample statistics or other analysis is performed between test steps and is used to reclassify certain devices or to change future manufacturing flow and test conditions for these devices. Part Average Testing and outlier identification methods are examples of the Post-Test category.

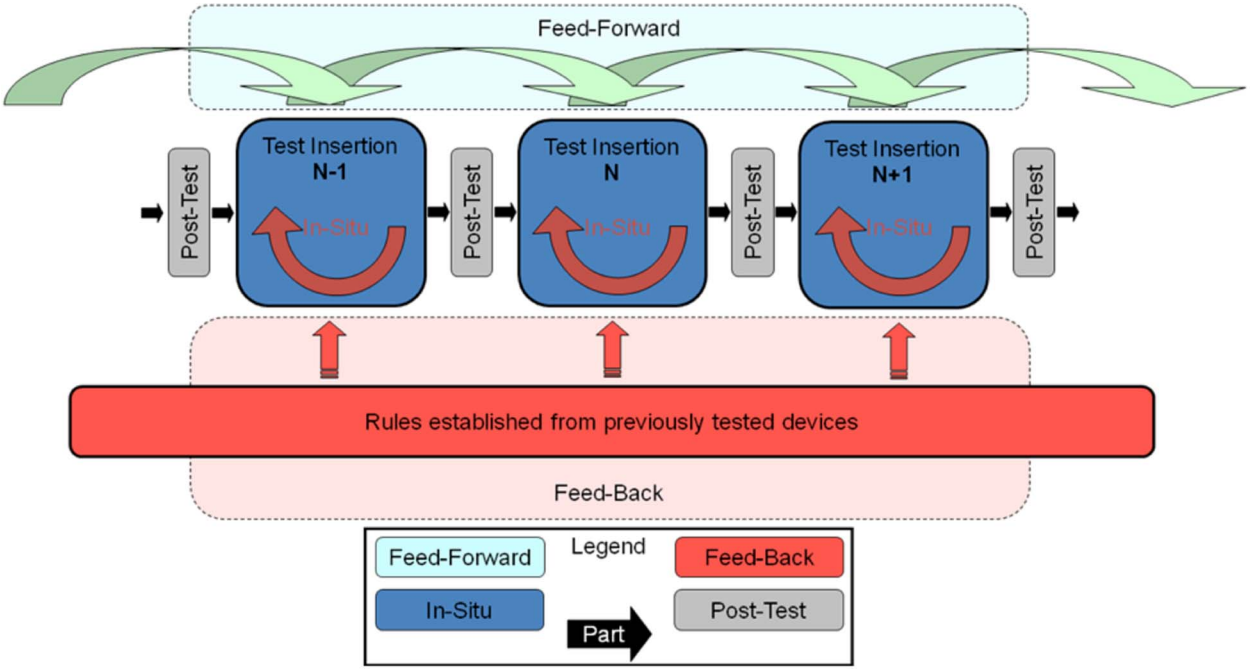


Figure 1: Adaptive Test supports feed-forward and feed-back data flows. Adaptive Test provides for data use for test decisions either in-situ (i.e. during a given test step) or post-test.

Example Applications

Below is a list of example Adaptive Test applications. Each example is labeled by one or two categories outlined earlier. In addition to clarifying the categories, the examples demonstrate the shift from manual and static methods to automatic methods with little or no human intervention during test execution. Note that the use of die ID (e.g., a die-specific identifier such as wafer/XY coordinate and lot information, or a unique identifier that is fused on each die) is a key enabler for many of these applications. There is a list of references that include many example applications at the end of this section.

Some of these applications are more widely used today than others. Common methods include:

- Data feed-forward
- Good die, bad neighborhood
- Post wafer test statistical analysis – e.g., pickmap updates
- Device trimming or reconfiguration
- PAT (Parts Average Testing) and DPAT (Dynamic Parts Average Testing) – for outlier detection

1. **Dynamic test flow changes (In-situ, Feed-forward):** Die production data is monitored within the test program to add or remove tests, to selectively perform per-die characterization for yield learning, or to collect data for later diagnosis. This application supports many common real-time Statistical Process Control (SPC) methods.
2. **Statistical screening (Post-Test, Feed-forward):** After wafer or lot data collection, identify die which are outliers or mavericks as possible sources of test-escapes spikes or reliability failures. Statistical screening is Feed-forward because results can be used to route target dies through test flows different from the main flow.

- PAT – Part Average Testing is a statistical technique in which a die is tested against static or dynamic test limits derived from other die in a common subgroup
- NNR – Nearest Neighbor Residuals is a statistical technique relating a univariate or multivariate test result to a model derived from a local region on wafer of the device under test.

3. **Single-step flow control (Feed-forward):** Data from one test step is used to optimize testing at the next test step to focus subsequent screening on issues observed in the manufactured parts.
   - For example, inline test modifies wafer test; wafer test modifies package test; burn-in modifies final test; or package test modifies card/system test.
   - One method for doing data feed-forward is to store results in on-chip memory (e.g., Flash) that will be read and used at subsequent test steps.
4. **Off-tester optimization of test flows (Feed-back):** Off-tester data analysis drives test flow changes for future devices (fully automated).
   - For example, off-line analysis could optimize test flows, test content and test measurement routines using input from many sources including historical data, test capacity, required turn-around times, DPM (defects per million) requirements, expected yields and parametric data.
5. **Production monitors and alerts (In-situ, Feed-forward, Feed-back):** Data from multiple sources is merged for statistical analysis to control production test optimization beyond what has historically been possible.
   - For example, subtle parametric shifts from marginal wafer probe contacting can be automatically identified and action taken during production testing.
6. **Die matching (Feed-forward, Post-test):** Production data from various sources is used to support the build/test process for multi-chip applications and many of today's board build process to match specific die combinations during assembly.
   - Note die-matching data transfer may require world-wide data sharing, across multiple companies and throughout the entire supply chain.
7. **On-chip test structures and sensors (In-Situ, Feed-forward, Feed-back):** Data collected from auxiliary on-chip test structures such as ring oscillators, selected critical paths, on-chip voltage and thermal sensors, or on-chip reliability monitors is used to modify the die's test content, test limits or future test flow.
   - Sensor measurements can be used at all levels of assembly and test (including system operation) to monitor and adjust functionality.
8. **On-chip configuration (In-situ, Feed-forward):** Production test data (including test structure data) is used to adjust features in the design to improve a die's performance, power, margin, yield or reliability.
   - Emerging ICs have more on-chip configuration and adaptability such as clock tuning, partial goods (redundant spare cores), and voltage and frequency adjustments (including per core).
9. **Component System Level Test (SLT) test optimization (In-Situ, Feed-forward, Feed-back, Post-test):** Test results from current or prior operations are used to customize the test flow or enable test sampling.
10. **Card/System configuration and test (Feed-forward, Post-test):** Component test results (such as parametric data, yield characteristics or partial good data) are used to customize the card/system test flow or customize card/system test conditions.
    - Data feed-forward from die testing are fed-forward and used by the board test program to make decisions on whether to add specific content to test for marginality.
    - Data feedback is used to deliver card/system test results to IC suppliers – who use this data to adjust test content, test condition or test limits.
    - In-situ card/system test measurements (such as on-chip sensors for voltage or temperature) are used to modify board testing (e.g., adjust margins and/or performance). On-chip sensors can also be read during field usage to monitor aging and perform in-field adjustments and/or send results back to IC suppliers to adjust their test limits.
11. **Adaptive Diagnostics (In-situ, Feed-forward):** Test results drive advanced diagnostic data collection.
    - For example, on-chip BIST (built-in self-test) circuitry can be programmed on-the-fly to localize and characterize specific types of failures. But these methods must only be selectively applied to ensure reasonable test time/cost.

- Many emerging chips have programmable on-chip test controllers that can interpret test results on-chip – and take action (test, diagnostics, characterization) without requiring extensive data collection being transmitted to/from the test equipment.

Adaptive Test Architecture & Infrastructure

Adaptive Test makes decisions throughout the manufacturing process based upon data from multiple sources and using data of varying detail and completeness. Before actionable decisions can be made with multiple-sourced data, new data integration requirements are needed. Some integration requirements are unique to Adaptive Test and are different from data requirements used at any of the originating sources.

Figure 2 displays a model of the entire End-to-End flow of parts under test and Adaptive Test applications. Note that in the flows shown in Figure 2; feed-forward, in-situ, feed-back and post-test dispositioning opportunities can occur at each test step. Although Figure 2 shows the total data store as a single database, the actual data structure would probably consist of multiple distributed database hierarchies each with unique capacity, latency and accessibility characteristics.

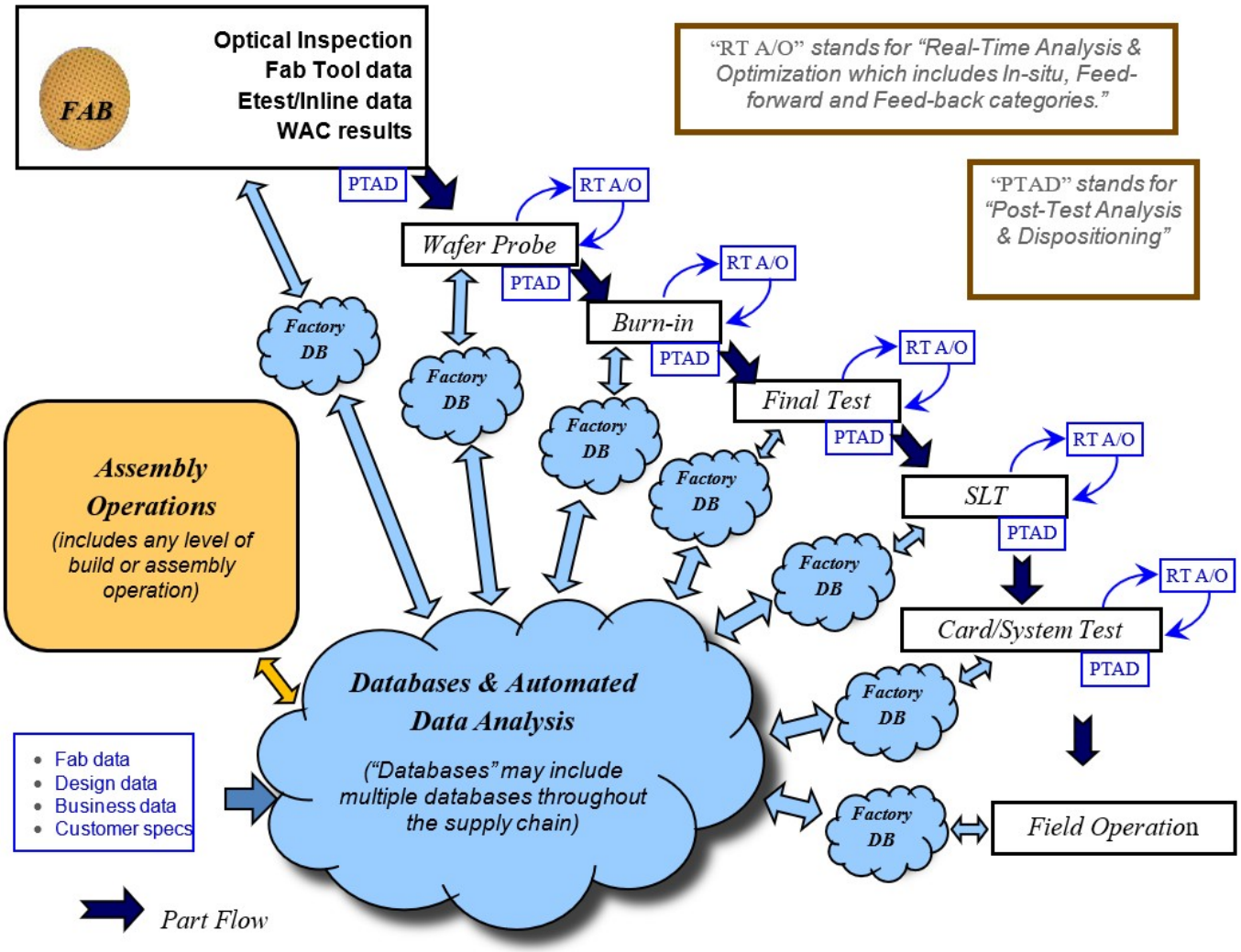


Figure 2: The architecture of Adaptive Test organizes each insertion's test data into one or more databases. A waterfall of manufactured parts may insert, join or query databases for test flow decision-making.

Overall Data Model Requirements for Adaptive Test

Making the decisions to adapt the test attributes listed above first involves collecting the proper data and then organizing the data into a structured data model so that the right data can be accessed when and where it is needed. At the appropriate time, data of the proper scope – that is data from a particular test run or data from a particular part, wafer, or lot – is accessed from the data model and processed by the applicable decision algorithms. Similarly, the

test variables, such as limits, conditions, flow and content, must be changed at the right time to complete the adaptation decision.

The data model can exist entirely in an off-line database apart from the tester, or be distributed between servers and the tester depending on the latency requirements and convenience. To branch a test flow for a particular part (a real-time decision) latency must be short, i.e. there can be no significant impact to test time. To support low-latency requirements, the data needs to either be stored on the tester or be rapidly pulled into the tester. To make an outlier decision such as to re-bin already tested parts, longer latencies are tolerated such as from the time of test until wafer maps are uploaded or the parts are dispositioned. Longer latencies mean an off-line database can be used.

Decisions to adapt a test are often based on comparing the variation observed on the sample of parts in question to a model of the expected variation. In outlier detection, parametric test limits are adapted to track expected variation so as to only discard parts with unexpected variation. Tests can be temporarily dropped from a test flow when their control charts show the manufactured material and test process is in control and within specification. If and when monitoring based on sample testing shows the material or test process has changed and gone out of control, the tests are reinstated on every part. Similarly, diagnostic data collection tests can be added to a test flow when certain types of failures occur more frequently than expected.

Generally, more adaptability means more frequent decision-making in the test flow with the goal of improving the trade-off between shipped product defect level and yield/cost. Adaptability follows a bottom-up progression from the conventional static limit, to a static parameter variance model (static PAT), to a variance model with variable parameters (dynamic PAT), to choosing variance model equations based upon well-grounded principles. Moving up this progression requires not only more data but also a better understanding of the processes that cause the test response to vary. This progression also means the decision-making generally moves from an off-line human activity to an on-line machine activity.

Data requirements unique to Adaptive Test center on the database policies of latency, access and retention period. Latency measures the time between the request for a data item and the availability of the requested item. Access refers to the scope of the user community that can store, retrieve, and act on a data item. Retention period measures the time the data item is electronically available within the required access time period.

- **Local processing in the test cell requires low latency.** For example, access latency should be in a few milliseconds if data is to be retrieved on a per device level – Real-Time Analysis & Optimization (RT A/O). Post-Test Analysis & Dispositioning (PTAD) applications may have latency requirements of a few seconds to a few minutes. Normally data volumes for these steps would be relatively small. Processing in a central database (e.g., "The Cloud") has more relaxed timing constraints (minutes, hours), but typically deals with much larger data volumes.
- **Data access requirements are influenced by the diversity of users "touching" the data.** Applications that bridge assembly or test companies, fabless design companies, and test developers require robust access mechanisms to ensure the correct people and processes access only the appropriate data.
- **Data retention requirements depend on the specific market requirement.** Some requirements drive data retention of all shipped products to 10 years or more.

Many areas of IC manufacturing are increasingly more comfortable with using data from the cloud, but a notable exception is the test cell. Test cell integration of Adaptive Test algorithms is one of the most challenging applications. For example, local test cell actions (such as "clean probe-card now") were the sole responsibility of the specific test floor and were designed to guarantee the test cell integrity and test cell-to-test cell correlation. Adaptive Test changes this paradigm in a number of ways:

- **Algorithms will be owned by multiple involved parties,** including wafer fab, design house and test floor. Some algorithms may originate from commercial providers, others from the involved parties themselves. They all need be executed synchronously in a real-time environment.
- **Data collection as well as data access** (e.g., to upstream data in case of data feed-forward) **becomes a mission-critical task of a test operation as well as the entire supply chain.** This challenges the reliability of the underlying infrastructure, which likely spans multiple companies, geographic areas and cultures.
- **It is required to simulate the impact of algorithms on historical databases** to understand how to maximize the value of Adaptive Test without creating adverse side effects. This requires the exact same algorithm to be executed in as diverse environments as a cloud database and a test cell measuring real-time data.

As a consequence, the industry needs to develop:

- **Provenance models** to allow disparate data sources and users to access and trust data.
- **Data exchange formats** which are flexible, compact and standardized so that only minimal extraction and translation effort is required. A common set of indices is required, such that data remains identifiable and traceable even across heterogeneous supply chains.
- **Recipe management systems** which can handle a diverse set of recipe origins, check for (likely unintended) interactions and maintain consistency across non-synchronized update cycles from the various origins. Version control systems for these recipes are also required.
- **Monitoring and rule-checking systems** must be enabled to monitor the health of Adaptive Test algorithms (are basic assumptions met?) and the health of the distributed data, and escalate errors to the right entities in an actionable format.
- **Distributed data distribution infrastructure** which can handle a diverse set of data origins, monitor data changes, and maintain consistency across non-synchronized updates from the various sources. Data version control systems are also required, as are logging systems able to trace changes.
- **Shared analysis** may be the only way to jointly discover problems and opportunities. The large quantities of proprietary data may require that analysis be performed in a distributed manner with intermediate results shared.

Global Adaptive Test Infrastructure Requirements *(data exchange and archiving, reference Figure 2)*

This section describes the data storage and exchange requirements across the supply chain.

Data requirements that are different but not unique to Adaptive Test include date and time stamping, test naming, and date recording methods. For example, Adaptive Test data stamps should be consistent across all insertions and between companies. Current date stamping practices are ad-hoc with some companies using different date formats and date references at different test insertions. Database standards exist for date stamping such as Coordinated Universal Time. A time and date stamp requirement policy eases integrating test data when some units are retested and simplifies merging two (or more) data sets in an unambiguous time order. Similar issues arise in recording floating point results of the same test from different insertions with different precisions and formats.

The following are important attributes of a global infrastructure:

- **Latency.** Global latencies can be quite long, only needing to satisfy the needs of the typical uses. Data at this level is packaged in some lot-related group so access time in the order of lot processing times are considered adequate. These times (for non-local data) could be in days.
- **Volume.** While data volumes have increased and storage times extended so has the ability to store this data. The real issue is not the storage but the locating and use of this data. Data volumes are in the order of 2-20 Gbytes per tester-week with history storage spanning 4 to 20 years with the longer times being for automotive, medical and some industrial applications.
- **Provenance.** All characteristics of data must be trusted. Who created it, has it been changed, and is it accurate are all facets which must be known.
- **Security.** Adaptive test requires the sharing of data but it must control that sharing to those who have the need and right to use it.
- **Reliability.** Testing cannot stop due to infrastructure equipment or communications failures. Checks are also required to ensure that the required data is collected and stored.

Production Test Floor Infrastructure and ATE (and component/socketed SLT) Requirements

This section describes the data infrastructure requirements local to the Test Floor.

The test cell is expected to deliver a cost-effective means to screen defects for quality, classify devices for performance and collect data for learning. The rate of product complexity is increasing with more clock domains, voltage planes, IOs and configurable fuses followed by the introduction of parallel testing of multiple, dissimilar devices with 2.5D and 3D stack packaging. In parallel, the business demand for higher quality and reduced product cost severely challenges the ability of the test cell to continue to provide an effective test solution and still continue to reduce the cost of test. Adaptive Test methods provide levers to address these additional demands but not without disruptions of their own.

Adaptive Test requires the test cell to be able to accept input from external and internal sources, apply device-specific models to determine test conditions and to evaluate results for flow control and device binning. This materially changes the setup, execution and categorization requirements of the test cell and affects both low-level software capability such as firmware as well as high-level software such as the executable test program. Of particular challenge is the relationship of flow control and binning when the test flow becomes a function of non-deterministic evaluations influenced by dynamic test points, limits and device configuration.

The following are important attributes of a test floor infrastructure:

- **Latency.** Adding Adaptive Test should not impact the throughput for a test cell. Data collection must have minimal impact. In addition, the various equipment in a cell must respond quickly to changes driven by an adaptive test rule. For real-time control, the response time should be in the low milliseconds for each part. Some applications require require that data taken from a part can be used for binning that same part in real time. There is also lot feed-forward which is more of a lot-based time scale, usually hours.
- **Volume.** Required data volume to enable Adaptive Test varies by Adaptive Test method. Incremental data for real time decisions is small whereas data for feed-back applications can be much larger. Data volumes in the range of 2-20 Gbytes of data per tester per week (while archiving this data monthly) are not unusual. The increase in data volume also brings challenges to the network infrastructure supporting the tester with increased demand on both reliability and bandwidth.
- **Provenance.** All data generated from a test cell must be fully trusted. Data should be signed and possibly encrypted.
- **Security.** Within a test floor, this is not usually a concern. With the advent of keys and other sensitive data being stored into parts, it is becoming necessary to comprehend security methods such as selective encryption of data from specific parts.
- **Reliability.** Testing cannot stop due to infrastructure equipment or communications failures. Each test cell must continue at least the current lot without floor communications. This implies that each cell has sufficient data resources (data bases and storage) to continue.
- **Legacy Support.** Support for legacy testers, handlers, probers and products.

In addition, future test cells will have to support the following:

- Per-device test flows (and per-device limits, test conditions and content) based on external inputs, the device itself, and dynamic business rules.
- A move from being the entire cell controller to a test engine with a standard API.
- Asynchronous and distributed (multi-insertion) test flows.

Much work is being done today outside of the semiconductor industry to address the needs for machines to communicate and for the data generated by these interactions to be collected, analyzed and acted on. The semiconductor industry should take advantage of these as appropriate. Groups like SEMI CAST (www.semi.org/en/collaborative-alliance-semiconductor-test) have been organized to stimulate discussion within the industry to standardize around available technologies.

- **IoT (Internet of Things) and MTM (Machine-to-Machine).** Communications protocols, e.g., MQTT (Message Queue Telemetry Transport) must be both secure and extendable.
- **Distributed file systems.** Systems like IPFS (Interplanetary File System) are proposed to handle the distribution of asynchronous data while maintaining trust.
- **Streaming data analysis.** While data mining of large collections of data is a popular topic today, the concept of treating data as a stream is more appropriate to the needs of the semiconductor industry.
- **Replay/simulation capability.** Data systems must have the capability to evaluate the impact of different test rules and flow. For example, there must be a "replay" capability where a user can change test rules – and evaluate the quality, cost and other impacts.

Test results driving reconfiguration of "Adaptive Designs"

More and more designs are being reconfigured during testing. Examples include partial goods (on-chip redundancy), VDD/frequency adjustment per die, 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".

Testing resilient features

Resilient features are on-chip structures that can be used to configure the product to work around hard defects or to tolerate latent defects. These structures span a wide range of circuits and architectures, including fuses, redundant memory elements, architectures capable of operating on reduced numbers of logic elements like CPU cores or graphics components, error-detection and retry schemes for hard and soft errors, and the sensing and control circuitry for adaptive designs. Like every other circuit, these structures must be themselves tested and characterized, though these circuits present unique testing challenges beyond standard logic and memory elements, including temporary configurations (for fuses), soft-repair vs. hard-repair validation (for memories), combinatorial explosion of down-cored variants of redundant features, the need for error injection to test recovery circuits, and analog stimulus for sensors (such as voltage or aging monitors).

While resilient features are widely used for memories, it is currently less frequently applied for logic cores. This could change if an automated approach were available to help chip designers employ partial-good die on their chips; this will need to include power-off means for bad cores and functional operation in the presence of non-functional core instances. This is in addition to DFT to isolate such cores and fuses to disable the bad ones once identified by testing. EDA companies need to pursue means to help designers add this to their chips.

Non-deterministic device behavior: Test and run-time availability

Non-determinism is incompatible with traditional cycle-accurate automated test equipment, but is nonetheless becoming typical on modern SOCs. Several new I/O protocols are non-deterministic, as are the standard methods to avoid metastability in clock-domain crossings (which are commonplace in highly integrated devices). Fine-grained power gating and power-state management can change the configuration of a device and its execution profile during test and normal operation. Adaptive designs take this notion even further with architectural features which can perform state rollback and pipeline retry based on events at arbitrary times during execution. The result is that test patterns, particularly functional patterns which execute in mission mode, must either prevent or be tolerant of non-deterministic response. The former raises coverage questions, the latter pattern and ATE interface challenges.

Testing Adaptive Designs

Adaptive designs bring the complexity of dealing with advanced power management such as power gating, variable configuration of IP (such as IO and arrays), self-defining performance bucketing and part-specific reconfiguration (such as redundancy, repair and harvesting) to a test environment traditionally characterized by a linear test flow measuring to fixed corners to verify device operability. Instead, on-chip sensors are used to detect the workload, voltage, temperature, and timing margin of the chip as it operates and dynamically adjusts power supplies, clock frequencies, thermal control, and even the instruction stream. The adaptive features of a design make it much harder to define (and thus characterize) both typical and worst-case use models, which in turn makes it more difficult to test appropriately. Additionally, the removal of excess margin represented by traditional guard-banding increases the risk of exposure to subtle defects, necessitating both higher coverage and better correlation between structural and functional test modes.

An emerging direction is to apply Adaptive Test techniques (which modify the parameters or content of a test program based on information collected about the device under test from the current or previous test insertions) to adaptive designs (which modify their own operating point or execution flows based on internally generated feedback). The proclivity of an adaptive design to compensate for the environment in which it is being (functionally) tested will present challenges for data gathering by the Adaptive Test process beyond opening control loops to test at fixed conditions. A means to record and store the conditions to which the device is tested, organized in a manner for ease of retrieval and consumption, is required.

Online Testing for Automotive ICs

Automotive standards such as ISO26262 are driving the need to perform testing in real-time while the system is in use. Common online testing methods include Logic Built-in Self Test (LBIST) and memory BIST and repair. Results from these online tests will drive reconfiguration – for example, to work around logic blocks that fail the online LBIST. It is clear that, as online testing is used more widely, the industry will need to develop methods to validate these BIST methods and ensure they are defect free. Also, it will be required to verify reconfiguration capability.

Adaptive Test for Fab Process Feedback & Control

There is a long history of defect-related fault information identified through test for use in yield improvement by the fab, but parametric tuning feedback related to performance has been much more limited. The use of data collection and statistical models applied through Adaptive Test methods is showing promise and an increasing level of interest in providing product performance related feedback to the wafer fabs. Example applications include optimizing the "sweet spot" for N and P device targets to meet customer demand distributions as well as providing direction in optimizing process modules for transistor-limited or R/C-limited performance improvement.

Adaptive Manufacturing (Post Fab)

An emerging direction is using test results to drive other IC production steps such as packaging multi-chip products. For example, the card/board assembly operation may require that specific dies or types of dies be used on specific boards based on previous test results. Given the emergence of multi-chip packages (such as 3D ICs) and power constraints, specific bare dies will need to be selected for assembly based on parametric data collected at test such as IDD or power/performance measurements. This opportunity is broader than just focusing on die build, and should include the entire post-fab supply chain including board, card, module and system manufacturing.

Key challenges of End-to-End data feed-forward for assembly operations include:

- Cross-company data management
- Build logistics for selecting specific dies/packages
- Robust data availability
- Data security
- Full traceability
- Data format standardization

Adaptive Test for Card/System/Field

Adaptive Test methodologies described in this section for IC-level testing can be equally extended and applied to board and system testing and even field usage. While ICs have traditionally been tested standalone in an almost 'noise-free' ATE environment and/or tested with limited structural tests to see whether they perform to their specifications, the board/system environment can be quite different in terms of noise, timing margin, voltage and functional test trigger conditions that structural tests were unable to produce. Improved board yield and IC DPM can be improved significantly where adaptive test that includes the board/system level performance is able drive enhanced screening both at the IC suppliers test and/or board/system manufacturing test.

The four types of Adaptive Test described in Section 10 (In-situ, Feed-forward, Feed-back and Post-test) can all be extended to include board and system manufacturing.

One of the difficulties in extending the chip-level adaptive test to board/system or even in-field test is to track their test trigger conditions and be able to convert between them. For example, chip-level scan-based logic gate test may not always be applicable for board/system/in-field tests due to the difficulties or impossibilities of controlling the scan chain data, clock pulse, non-stoppable in-field online function executions, etc. Similarly, a functional execution, which can be treated as a functional test, may be hard to convert to a chip-level ATE test because the function execution could involve memory contents, their transactions, logic and I/O data flow, etc. Therefore, tracking the test/failure conditions and the capability to convert between them is the key for adaptive test extension to board/system level.

A key emerging requirement is to enable full end-to-end correlation analysis – e.g., from fab data, through product die test, board/card/system test and field operation. The infrastructure to enable this capability broadly is one of the industry's key challenges.

Extending Adaptive Test applications to the board and system level requires extensive data infrastructure, analysis, exchange and security. Companies providing ICs, board design and test need to openly collaborate on a technical and business level to be successful.

Adaptive Test Challenges and Directions

This section highlights the key challenges that the industry must address to fully exploit Adaptive Testing across the supply chain.

The color scheme of the table below is:

- White – Manufacturing solutions exist today.
- Yellow – solutions are known – but there are still adoption hurdles
  - Solutions may be widely accepted or solutions may only be company-specific.
- Red – Research & development is needed to develop solutions
  - Or some solutions may only be proprietary and not generally available.

Challenge	Status	Needs
Recipe (Input variables, data treatments, output variables/ responses)	Many good outlier recipes exist. Spatial interpolation over a wafer is becoming popular for sample testing. Opportunity to branch test flow for only fault coverage required by current defect rate	Guidance on best measurements to make
		Guidance on which recipes to apply
		Fault or defect coverage metrics
		Higher level of adaptability where best variance model is automatically discovered instead of chosen beforehand
Decision Rule (Define actions resulting from recipe output)	Actions defined for gross outliers; loosely defined for less extreme events such as downgrade or reconfigure or escalation/ de-escalation of decisions (such as test sampling)	Where to set outlier thresholds
		How to combine the results of multiple outlier definitions (e.g., develop metrics for Quality or Reliability)
		Criteria for rejecting vs. downgrading or reconfiguring
Infrastructure (Foundation to enable execution of recipes and decision rules)	Part traceability enables feed-forward, feed-back but robust environment for data transport, storage and providence is lacking (no commercial solutions currently exist).	Standard data formats amenable to adaptive test
		Move from working with files to working with databases
		Ability to feed data and decisions both forward (with the parts to future stages) and backward (future parts at given stage) in test & assembly manufacturing flow
		Full traceability of adaptive test parameters for each part: limits, content, flows, decision rules, model parameters
		Full part configuration as tested (e.g., redundant cores, partial goods, on-chip tuning, multiple die such as 2.5D/3D in a package)
		Real-time communication among test cell machines and data storage and analysis engines
Evaluation (Execution of test cases to prove viability and benefit)	Receiver-operator curve concept understood by most practitioners but standard methods for experimental definition and ROI interpretation do not exist.	Clear evaluation criteria to build trust in adaptive test methods
		"Gold standard" against which to compare outliers (including all variations of adaptive test flows & settings)
		Good metrics for continuous monitoring of recipe effectiveness.
		Replay capability to evaluate new rules/recipes.
		Quantification of cost of shipping a bad part
Deployment (Implementation and release into production use)	Company-specific implementations currently utilized. But current deployments vary widely in their capabilities and there are still significant implementation hurdles to apply at all Test steps.	Commercial adaptive test platform into which methods can be plugged and recipes specified
		Connections to Manufacturing Execution Systems & Product Data Management systems
		Real-time (unit level) decision making that requires decisions based on off-tester analysis using broad set of data. (and update die result in real-time on the tester)
		Complete visibility across supply chain: fab, test, assembly both internal and external
		Supply chain data integration and processes which automatically detect supply chain issues and implement corrective actions in near real-time

Summary

Adaptive Testing has the opportunity to improve product quality & reliability, reduce cost and improve product yield beyond today's capabilities. Almost all companies are starting to use some forms of Adaptive Testing, but there is not a sequential roadmap for implementation and many applications are created in an ad-hoc way.

Adaptive Test methods are evolving over time as new technologies (<10nm, SOI), design methodologies (multichip/stack packaging) and supply chain support models are introduced.

There are a number of challenges that are today limiting the industry's ability to fully exploit Adaptive Testing across the supply chain. These are highlighted in the table in a previous section.

References and Industry Links

Below is a list of references that provide information about methods and applications of Adaptive Testing.

Part Average Testing (PAT) is described in the AEC-Q001 documents at http://www.aecouncil.com/AECDocuments.html.

K. Huang, N. Kupp, C. Xanthopoulos, J.M. Carulli, Y. Makris, "Low-Cost Analog/RF IC Testing Through Combined Intra- and Inter-Die Correlation Models," Design & Test, Volume 32, Issue 1, pp. 53-60, 2015.

A. Ahmadi, K. Huang, A. Nahar, B. Orr, M. Pas, J.M. Carulli Jr., Y. Makris, "Yield Prognosis for Fab-to-Fab Product Migration," IEEE VLSI Test Symposium, Session 3B, p. 3.3, 2015.

S. Zhang, X. Li, R.D. Blanton, J. Machado da Silva, J.M. Carulli, K.M. Butler, "Bayesian Model Fusion: Enabling Test Cost Reduction of Analog/RF Circuits via Wafer-level Spatial Variation Modeling," International Test Conference Proceedings, p. 19.1, 2014.

U. Guin, K. Huang, D. DiMase, J.M. Carulli Jr., M. Tehranipoor, Y. Makris, "Counterfeit Integrated Circuits: A Rising Threat in the Global Semiconductor Supply Chain," Proceedings of the IEEE, Volume 102, Issue 8, 2014.

Xanthopoulos, K. Huang, A. Poonawala, A. Nahar, B. Orr, J.M. Carulli Jr., Y. Makris, "IC Laser Trimming Speed-Up through Wafer-Level Spatial Correlation Modeling," International Test Conference Proceedings, p. 19.2, 2014.

Ahmadi, K. Huang, S. Natarajan, J.M. Carulli Jr., Y. Makris, "Spatio-Temporal Wafer-Level Correlation Modeling with Progressive Sampling: A Pathway to HVM Yield Estimation," International Test Conference Proceedings, p. 18.1, 2014.

Tikkanen, J.; Siatkowski, S.; Sumikawa, N.; Wang, L.-C.; Abadir, M.S., "Yield optimization using advanced statistical correlation methods," Test Conference (ITC), 2014 IEEE International , vol., no., pp.1,10, 20-23 Oct. 2014

K. Huang, N. Kupp, J. Carulli and Y. Makris "Handling discontinuous effects in modeling spatial correlation of wafer-level analog/RF tests", Proc. Design Autom. Test Eur. Conf., pp.553 -558 2013

K. Huang, N. Kupp, J. Carulli and Y. Makris "On combining alternate test with spatial correlation modeling in analog/RF ICs", Proc. IEEE Eur. Test Symp., 2013

S.K. Devarakond, S. Bhattacharya, J. McCoy, A. Nahar, J.M. Carulli, A. Chatterjee, "Predicting Die-to-Die Process Variations from Wafer Test Data: A Feasibility Study," IEEE Latin American Test Workshop, April 2013.

K. Huang, N. Kupp, J.M. Carulli, Y. Makris, "Process Monitoring through Wafer-level Spatial Variation Decomposition," International Test Conference Proceedings, p. 5.3, 2013.

C-K Hsu, F. Lin, K-T Cheng, W. Zhang, X. Li, J.M. Carulli Jr., K.M.. Butler, "Test Data Analytics - Exploring Spatial and Test-Item Correlations in Production Test Data," International Test Conference Proceedings, p. 5.2, 2013.

K. Huang, J.M. Carulli, Y. Makris, "Parametric Counterfeit IC Detection via Support Vector Machines," International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems, p. 1.1, 2012.

N. Kupp, K. Huang, J. Carulli and Y. Makris "Spatial correlation modeling for probe test cost reduction", Proc. IEEE/ACM Int. Conf. Comput.-Aided Design, pp.23 -29 2012

N. Kupp, K. Huang, J.M. Carulli, Y. Makris, "Spatial Estimation of Wafer Measurement Parameters Using Gaussian Process Models," International Test Conference Proceedings, p. 5.1, 2012.

K. R. Gotkhindikar, W.R. Daasch, K.M. Butler, J. M. Carulli Jr., A. Nahar, "Die-level Adaptive Test: Real-Time Test Reordering and Elimination," International Test Conference Proceedings, p. 15.1, 2011.

Sumikawa, N.; Drmanac, D.G.; Wang, Li.-C.; Winemberg, L.; Abadir, M.S., "Forward prediction based on wafer sort data — A case study," Test Conference (ITC), 2011 IEEE International , vol., no., pp.1,10, 20-22 Sept. 2011

Sumikawa, N.; Drmanac, D.; Li-C Wang; Winemberg, L.; Abadir, M.S., "Important test selection for screening potential customer returns," VLSI Design, Automation and Test (VLSI-DAT), 2011 International Symposium on , vol., no., pp.1,4, 25-28 April 2011

Drmanac, D.; Sumikawa, N.; Winemberg, L.; Wang, L.-C.; Abadir, M.S., "Multidimensional parametric test set optimization of wafer probe data for predicting in field failures and setting tighter test limits," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011 , vol., no., pp.1,6, 14-18 March 2011

Sumikawa, N.; Drmanac, D.; Wang, Li.-C.; Winemberg, L.; Abadir, M.S., "Understanding customer returns from a test perspective," VLSI Test Symposium (VTS), 2011 IEEE 29th , vol., no., pp.2,7, 1-5 May 2011

W. Zhang "Virtual probe: A statistical framework for low-cost silicon characterization of nanoscale integrated circuits", IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 30, no. 12, pp.1814 -1827 2011

Drmanac, D.; Laisne, M.; Wang, L.C., "Wafer probe test cost reduction of an RF/A device by automatic testset minimization — A case study," Test Conference (ITC), 2011 IEEE International , vol., no., pp.1,10, 20-22 Sept. 2011

S. Reda and S. R. Nassif "Accurate spatial estimation and decomposition techniques for variability characterization", IEEE Trans. Semicond. Manuf., vol. 23, no. 3, pp.345 -357 2010

E. Marinissen, A. Singh, D. Glotter, M. Esposito, J. Carulli, A. Nahar, K. Butler, D. Appello, and C. Portelli, "Adapting to adaptive testing," Design, Automation and Test in Europe Conference & Exhibition (DATE), pp. 556-561, 2010.

Chen, J.; Jing Zeng; Wang, L.-C.; Mateja, M., "Correlating system test fmax with structural test fmax and process monitoring measurements," Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific , vol., no., pp.419,424, 18-21 Jan. 2010

Chen, J.; Jing Zeng; Wang, Li.-C.; Mateja, M.; Rearick, J., "Predicting multi-core system Fmax by data-learning methodology," VLSI Design Automation and Test (VLSI-DAT), 2010 International Symposium on , vol., no., pp.220,223, 26-29 April 2010

D. Mannath "Structural approach for built-in tests in RF devices", Proc. Int. Test Conf., 2010

Chen, J.; Wang, Li.-C.; Po-Hsien Chang; Jing Zeng; Yu, S.; Mateja, M., "Data learning techniques and methodology for Fmax prediction," Test Conference, 2009. ITC 2009. International , vol., no., pp.1,10, 1-6 Nov. 2009

Drmanac, D.; Bolin, B.; Wang, L.-C.; Abadir, M.S., "Minimizing outlier delay test cost in the presence of systematic variability," Test Conference, 2009. ITC 2009. International , vol., no., pp.1,10, 1-6 Nov. 2009

K.M. Butler, J.M. Carulli, J. Saxena, A. Nahar, R. Daasch, "Multidimensional Test Escape Rate Modeling," Design & Test of Computers, Volume 26, Issue 5, pp. 74-82, 2009.

A. Nahar, K. Butler, J. Carulli, and C. Weinberger, "Quality improvement and cost reduction using statistical outlier methods," IEEE International Conference on Computer Design, pp. 64-69, 2009.

A. Nahar, K.M. Butler, J.M. Carulli, C. Weinberger, "Statistical Outlier Method Applications", IEEE International Workshop on Defect and Data-Driven Testing, 2009.

Wu, S.H.; Drmanac, D.; Wang, Li.-C., "A Study of Outlier Analysis Techniques for Delay Testing," Test Conference, 2008. ITC 2008. IEEE International , vol., no., pp.1,10, 28-30 Oct. 2008

R. Aitken, "Defect or Variation? Characterizing Standard Cell Behavior at 90 nm and Below," IEEE Transactions on Semiconductor Manufacturing, vol. 21, pp. 46-54, 2008.

H.-G. Stratigopoulos and Y. Makris "Error moderation in low-cost machine-learning-based analog/RF testing", IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 27, no. 2, pp.339 -351 2008

F. Ferhani, N. Saxena, E. McCluskey, and P. Nigh, "How Many Test Patterns are Useless?" VLSI Test Symposium, pp. 23-28, 2008.

K.M. Butler, J.M. Carulli, J. Saxena, "Modeling Test Escape Rate as a Function of Multiple Coverages," International Test Conference Proceedings, p. 21.2, 2008.

F. Liu "A general framework for spatial correlation modeling in VLSI design", Proc. Design Autom. Conf., pp.817 -822 2007

L. Fang, M. Lemnawar, and Y. Xing. Cost Effective Outliers Screening with Moving Limits and Correlation Testing for Analogue ICs. In Proceedings IEEE International Test Conference (ITC), October 2006. Paper 31.2.

T.J. Anderson, J.M. Carulli, "Modeling & Monitoring of Product DPPM with Multiple Fail Modes," International Reliability Physics Symposium, p. 545, 2006.

K. M. Butler, S. Subramaniam, A. Nahar, J. M. Carulli Jr., T. J. Anderson, W. R. Daasch, "Successful Development and Implementation of Statistical Outlier Techniques on 90nm and 65nm Process Driver Devices," Proc. 2006 Int. Reliability Physics Symp., pp. 552-559, Oct. 2006.

J.M. Carulli, T.J. Anderson, "The impact of multiple failure modes on estimating product field reliability," Design & Test of Computers, Volume 23, Issue 2, p. 118, 2006.

A. Nahar, R. Daasch, and S. Subramaniam, "Burn-in Reduction Using Principle Component Analysis," Proc. 2005 IEEE Int. Test Conf., Paper 7.2, Nov. 2005.

L. Ning, A. Nahar, R. Daasch, K. M. Butler, J. M. Carulli Jr., S. Subramaniam, "Burn-in Reduction using Robust Canonical Correlation Analysis," Proc. SRC TECHCON 2005, Paper 9.1, Oct. 2005.

Z. Yuan and Y. Yang "Combining linear regression models: When and how?", J. Amer. Stat. Assoc., vol. 100, no. 472, pp.1202 -1214 2005

R. Turakhia, B. Benware, R. Madge, T. Shannon, R. Daasch, "Defect Screening Using Independent Component Analysis on IDDQ," Proc 2005 IEEE VLSI Test Symp., pp. 427-432, May 2005.

W. C. Riordan, R. Miller, E. R. St. Pierre, "Reliability Improvement and Burn In Optimization Through the use of Die Level Predictive Modeling," Proc. 2005 IEEE Int. Rel Physics Symp., pp. 435-445, Apr. 2005.

R. Madge, B. Benware, R. Turakhia, R. Daasch, C. Schuermyer, and J. Ruffler, "In search of the optimum test set - adaptive test methods for maximum defect coverage and lowest test cost," Proceedings International Test Conference, pp. 203-212, 2004.

S.S. Sabade, D. M. H. Walker, "CROWNE: Current Ratio Outliers with Neighbor Estimate", DFT 2003, pp 132-139.

C. Schuermyer, B. Benware, K. Cota, R. Madge, R. Daasch, L. Ning, "Screening VDSM Outliers Using Nominal and Subthreshold Supply Voltage IDDQ," Proceedings IEEE Int. Test Conference, 2003, pp. 565-573, Sep 2003.
Automotive Electronics Council, "Guidelines for Part Average Testing (PAT)," AEC Q001, rev. C, 18 July 2003, http://www.aecauncil. comAECDocuments.html.
P. Buxton, P. Tabor "Outlier Detection for DPPM Reduction," Proceedings IEEE Int. Test Conference, 2003, pp. 818-827, Sep-Oct 2003.
T.S. Barnett and A.D. Singh. Relating Yield Models to Burn-In Fall-Out in Time. In Proceedings IEEE International Test Conference (ITC), pages 77-84, October 2003.
S.S. Sabade, D. M. Walker, "Evaluation of Effectiveness of Median of Absolute Deviations Outlier Rejection-based IDDQ Testing Burn-in Reduction" Proceedings 2002 VLSI Test Symp., pp. 81-86, Apr-May 2002.
D. I. Bergman, H. Engler, "Improved IDDQ Testing with Empirical Linear Prediction", Proc. 2002 IEEE Int. Test Conf, page 954, Nov. 2002.
W.R. Daasch, J. McNames, R. Madge, and K. Cota. Neighborhood Selection for IDDQ Outlier Screening at Wafer Sort. IEEE Design & Test of Computers, 19(5):74-81, September-October 2002.
W. R. Daasch, J. McNames, R. Madge and K. Cota "outlier screening at wafer sort", IEEE Design Test Comput., vol. 19, no. 5, pp.74 -81 2002
P. N. Variyam, S. Cherubal and A. Chatterjee "Prediction of analog performance parameters using fast transient testing", IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 21, no. 3, pp.349 -361 2002
T.S. Barnett, A.D. Singh, M. Grady, and K. Purdy. Redundancy Implications for Early-Life Reliability: Experimental Verification of an Integrated Yield-Reliability Model. In Proceedings IEEE International Test Conference (ITC), pages 693-699, October 2002.
J. Saxena, K. M. Butler, J. Gatt, Raghuraman R., S. P. Kumar, S. Basu, D. J. Campbell, J. Berech, "Scan-Based Transition Fault Testing: Implementation Challenges," Proc. 2002 IEEE Int. Test Conf., pp. 1120-1129, Oct. 2002.
R. Madge et al. Screening MinVDD Outliers Using Feed-Forward Voltage Testing. In Proceedings IEEE International Test Conference (ITC), pages 673-682, October 2002.
R. Madge, M. Rehani, K. Cota, and W.R. Daasch. Statistical Post-Processing at Wafer Sort - An Alternative to Burn-In and a Manufacturable Solution to Test Limit Setting for Sub-Micron Technologies. In Proceedings IEEE VLSI Test Symposium (VTS), pages 69-74, May 2002.
T.S. Barnett, A.D. Singh, M. Grady, and K. Purdy. Yield-Reliability Modeling: Experimental Verification and Application to Burn-In Reduction. In Proceedings IEEE VLSI Test Symposium (VTS), pages 75-80, May 2002.
S.S. Sabade, D.M.H. Walker, "Improved wafer-level spatial analysis for IDDQ limit setting, Proc. 2001 IEEE Int. Test Conf., pp 82-91, Nov. 2001.
R. Daasch, K. Cota, J. McNames, R. Madge, "Neighbor Selection for Variance Reduction in IDDQ and Other Parametric Data" Proceedings IEEE Int. Test Conference, 2001, pp. 92-100, Oct 2001.
S. Benner and O. Boroffice, "Optimal production test times through adaptive test programming", Proceedings International Test Conference, pp. 908-915, 2001.
R. B. Miller, W. C. Riordan, "Unit Level Predicted Yield: A Method of Identifying High Defect Density Die at Wafer Sort," Proc 2001 IEEE Int. Test Conf., pp. 1118-1127, Oct.-Nov. 2001.
K. Butler and J. Saxena, "An empirical study on the effects of test type ordering on overall test efficiency," Proceedings International Test Conference, pp. 408-416, 2000.
E. Riess, "Automatic Monitoring of Electrical Parameters in the Semiconductor Industry Based on ROC," IEEE Trans. On Sys., Man, and Cybernetics-Part A: Sys. And Humans, vol. 30, No. 6, pp. 853-857, Nov. 2000.
C. Thibeault, "Diagnosis Method Using AIDDQ Probabilistic Signatures: Theory and Results", Journal of Electronic Testing: Theory and Applications 2000, Vol 4., pp 339-353.
A. Kinra, H. Balachandran, R. Thomas, J.M. Carulli, “Logic mapping on a microprocessor,” International Test Conference Proceedings, p. 701, 2000.
W. R. Daasch, J. McNames, D. Bockelman, K.Cota, R. Madge, "Variance Reduction Using Wafer Patterns in IDDQ Data" Proceedings IEEE Int. Test Conference, 2000, pp. 189-198, Oct 2000.
J.M. Carulli Jr., D.C. Wrobbel, A. Mehta, K.E. Krause Jr., B.E. Campbell, F.A. Valente, “Frequency Distribution Modeling for High-Speed Microprocessors Using On-chip Ring-oscillators,” SPIE Proceedings, Inline Methods & Monitors for Process & Yield Improvement, p. 146, 1999.
A. E. Gattiker, W. Maly, "Current Signatures: Application," Proc. 1998 IEEE Int. Test Conf, pp. 1168-1177, Oct. 1998.
B. Stine, D. Boning, and J. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," IEEE Transactions on Semiconductor Manufacturing, vol. 10, pp. 24-41, 1997.
A. Singh, P. Nigh, and C. M. Krishna, "Screening for known good die (KGD) based on defect clustering: An experimental study," in Proceedings International Test Conference, 1997, pp. 362-371.
A.D. Singh, P. Nigh, and C.M. Krishna. Screening for Known Good Die based on Defect Clustering: An Experimental Study. In Proceedings IEEE International Test Conference (ITC), pages 324-331, November 1997.

T. W. Williams, R. H. Dennard, R. Kapur, M. R. Mercer, W. Maly, "IDDQ Test: Sensitivity Analysis of Scaling," Proc. 1996 IEEE Int. Test Conf, pp. 786-792, Oct. 1996.
A. Singh and C. Krishna, "On the effect of defect clustering on test transparency and IC test optimization," IEEE Transactions on Computers, vol. 45, no. 6, pp. 753-757, Jun. 1996.
A.D. Singh and C.M. Krishna. On Optimizing VLSI Testing for Product Quality Using Die Yield Prediction. IEEE Transactions on Computer-Aided Design, 12(5):693-709, May 1993.
J. M. Soden, C. F. Hawkins, R. K. Gulati, and W. Mao, "IDDQ Testing: A Review," J. Electronic Testing:Theory and Applic. (JETTA), vol.3, no.4, pp. 291-303, Dec. 1992.
C. Crapuchettes, "Testing CMOS IDD on Large Devices," Proc. 1987 IEEE Int. Test Conf., pp. 310-315, Sep. 1987.

Industry Links
CAST www.semi.org/en/collaborative-alliance-semiconductor-test
DR Yield dryield.com
Optimal+ www.optimalplus.com
PDF Solutions www.pdf.com
Mentor Graphics (previously Galaxy) www.mentor.com/products/silicon-yield/quantix/
Qualtera www.Qualtera.com

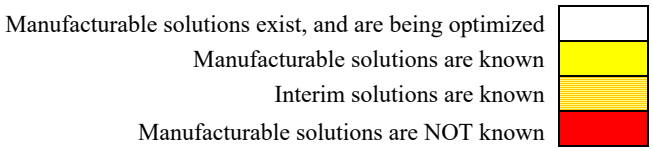
Section 10: DFT, Concurrent, and SOC Testing

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 includes targeting the interfaces between the cores and the top-level glue logic (i.e. a logic not placed within a core) in addition to what is within each core instance. 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, it is indispensable to improve the SOC test technology to handle a progression of design technologies accelerated by the evolving applications. The well-organized roadmap and the potential solutions that reflect these design intents should be reviewed by the readers. For example, 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 this issue is strongly required. In product area requiring high reliability, in particular automotive devices, recently in-system self-test of digital circuits are required and logic BIST has come into use for this purpose. Furthermore, wide adoption of FinFET transistor technology possibly brings new and elusive defects on the silicon. Additional tests on logic and memory must be captured by introducing the corresponding test structure as extensions of existing DFT.

The quantitative trends and requirements of a consumer logic chip are shown in Logic section, compared with a MPU chip. Table 1 introduces the guideline for DFT design and the requirements for EDA tools.

Table 1: DFT Requirements



Definitions for DFT Requirements Table 1 (next page):

[1] STIL (Test Interface Language, IEEE1450.x) is an example. I/F should include not only test vectors, but also parametric factors.

[2] A method to obtain overall test quality measure of SoC considering all cores; logic, memory and analog.

[3] Growing number of row & column spares, and both divided and shared spares for segments in the future.

[4] The current BISR for two dimensional repair is limited to a few row and column spares.

[5] IEEE1500 and IEEE P1687 (IJTAG) are examples.

[6] ATE software analyzes power and noise at testing and schedules concurrent test from IP/chip information.

Year of Production	2018	2019	2020	2021	2026	2031
DFT Methodology for SOC						
Hierarchical DFT Techniques This includes use of test compression within blocks/cores and top level tie-in of compression structures to chip scan interface resources. (FAL: Full Automation Limited use, GA: Generally Applied to any SOC)	GA	GA	GA	GA	GA	GA
3D Stacked Die DFT Techniques for chips with TSVs This includes DFT for wafer test for middle or top die in stack which may have only TSVs and micro-bumps to contact neighboring die in stack; it also includes DFT for stack testing. (PA: Partially automated, FAL: Full Automation with limited industry use, GA: Generally Applicable to most SoCs)	PA	PA	FAL	FAL	GA	GA
Logic Core Integration						
Supported Fault Models by ATPG for Overall Test (DBFM: Defect-based Fault, Model SDX: Extended Small Delay)	+DBFM	+DBFM	+SDX	+SDX	+SDX	+SDX
Standardization of DFT-ATE I/F [1] (LF: Limited Use of Full Information, F: General Use of Full Information)	LF	F	F	F	F	F
SoC Level Fault Coverage [2] (AH: Adhoc, L: Logic, M: Memory, IO: I/O, A: Analog)	L+M	+IO	+IO	+A	+A	+A
Inter-Core/Core-Interface Test (PA: Partially Automated ; FA: Fully Automated)	FA	FA	FA	FA	FA	FA
Embedded cores: Memory						
Repairing Mechanism of Memory Cells to improve Yield [3] (RC: BISR/BISD for a few Row & Col R/D, RCM: for more Row & Col R/D, M: for More Sophisticated R/D)	RCM	RCM	M	M	M	M
Area Investment of BIST/BISR/BISD [4] (Kgates/Mbits)	35	35	35	35	35	37
AMS Core Integration						
AMS BIST with digital interface; covers PLLs, High Speed SERDES, DA & AD, other AMS cores with BIST. Should include coverage estimate for Core. (PA: Partial Automation, FAL: Full Automation Limited use, GA: Generally Application to any SOC)	PA	PA	FAL	FAL	GA	GA
AMS non-BIST; covers any AMS cores that require functional tests using analog stimulus and/or response. Should include coverage estimate for Core. (PA: Partial Automation, FAL: Full Automation Limited use, GA: Generally Application to any SOC)	PA	PA	FAL	FAL	GA	GA
DFT in Manufacturing						
Systematic Hierarchical Diagnosis (L: Logic, M: Memory, I: Interface, A:Analog)	+I	+I	+A	+A	+A	+A
Supported Defect Type for Fault Diagnosis (C: Conventional (SAF, TF, BF), D: Delay Fault Model Considering Defective Delay Size, CT: Cross-talk, TRF: Transient Response Fault)	+CT	+CT	+CT	+TRF	+TRF	+TRF
Standardized Diagnosis Interface/Data in the diagnosis flow (ATE: Tester Log, DFT: DFT Method, PFA: Physical Failure Analysis)	+PFA	+PFA	+PFA	+PFA	+PFA	+PFA
Volume Diagnosis Database (SI: Collection and Storing Defect Information (B: Bad sample, G: Good sample), AD: Automated SoC Diagnosis)	+AD	+AD	+AD	+AD	+AD	+AD
Concurrent Testing						
Automated DFT environment for Concurrent Testing; integrates efficient interfaces for test of core itself and core test access [5]. (D: Digital , A: Analog)	D	D	D+A	D+A	D+A	D+A
ATE for Concurrent Testing There are some items to be carefully considered when Test scheduling [6] (R: pin Resource T: Test time P: Power consumption N: Noise)	R+T+P	+N	+N	+N	+N	+N
Standardized IP core access interface for Concurrent Testing (L: Logic , M : Memory , HV: high-voltage I : high-speed Interface A : Analog)	+HV	+HV	+A	+A	+I	+I
Test time reduction ratio by concurrent test (%) (L: Logic , M : Memory , HV: high-voltage I : high-speed Interface A : Analog)	95 (L+M)	90 (+HV)	75 (+A)	60 (+I)	60	60

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, compression logic, core wrappers 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 1 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:

- **Flat**: 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. Virtually no one does this anymore, but it is useful for comparison purposes with the more appropriate approaches briefly described below.
- **Hierarchical**: The EDA tools can consider the hierarchal design elements to achieve an on-die parallel test setup. Parallel test would be applied to instances of wrapped cores to enable multiple instances to be tested in parallel.
- **Compression**: The EDA tools can imbed compression and decompression circuitry around the scan chains, allowing for many times more chains internally without increase of ATE scan pin resources, resulting in less data being required to be stored on the ATE for stimulus or output comparison purposes.
- **Hierarchical and Compression**: The EDA tools can implement a combination of 2 and 3 for a hierarchal compressed approach. This would involve cores being wrapped for isolation and including compression within the cores. Further compression can be obtained by testing multiple instances with the same set of scan-in pins, considered scan pin sharing, to allow testing of multiple instances of cores in parallel. The test data/scan outputs from each core instance can be observed independently or further compressed together and sent to a common set of chip scan-out pins, possibly resulting in more chip scan pin sharing.

The approach used to apply tests to embedded cores will have a large impact on test time and probably also test data volume. One traditional approach is to test a core in isolation and route its stimulus and expected responses up to the SOC pins to avoid running 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 test compression approach is to test multiple cores in parallel and not put them into complete isolation from other cores. If compression can be used inside cores, it can also be used in the upper hierarchy of the cores to send the scan stimulus to multiple cores in parallel and to compact the output from several cores before sending it off-chip.

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, cell-aware and other defect-based faults to achieve a higher-level of test quality. Scan 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 reduced over the roadmap time period. Therefore, various approaches, such as test pattern reduction, scan chain length reduction and scalable speed-up of scan shift frequency, should be investigated. The acceleration of scan shift speed increases the power consumption during scan shift cycles and it might possibly make the test power problem more serious. Therefore, some DFT and ATPG approaches to solve the problem are required. Power consumption during the scan capture cycle is also an important issue and several approaches to relax this issue have been proposed. However, 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 is shown with a 20% test data volume premium in the low-power rows. This will be too optimistic for cases where very low (e.g. less than 15%) switching is required since that could easily result in a doubling of the pattern count for the same coverage.

Another problem caused by the increase of test patterns is the test data volume. Even assuming tester memory size will be doubled every three years, high test data compression ratios will be required in the near future; therefore, test data reduction will remain a serious issue that must be tackled. One possible solution to reduce test application time and test data volume at a time is simultaneous test of repeatedly used IP cores in a design that can share a

common set of chip scan pins. By broadcasting the same scan-in stimulus to all such core instances, we reduce the bandwidth of data being sent from the ATE onto the chip and need less storage for that data on the ATE. Observing the outputs from each instance independently can aid in diagnosing failures, but by compressing the core instance outputs together and observing them at a common set of chip pins further increases the effectiveness of compression inside each core.

The increase of power domains may require some additional test patterns. Since the increase of test patterns will be linear to the number of power domains, it won't have severe impact on overall test pattern counts. Nevertheless, the increase of power domains or restrictions on test power possibly prevents maximum simultaneous test of identical IP cores. The impact of this effect should be investigated for future editions of the roadmap.

The issue of power consumption during test mentioned above is one cause for the increase of test patterns which will increase test data volume. Therefore, requirements on test data reduction also take account of this issue.

Year of Production	2016	2017	2018	2019	2020	2025	2030
Worst Case (Flat) Data Volume (Gb)							
MPU-HP - High Performance MPU (Server)	3673	4998	6802	9256	11366	31737	88623
MPU-CP - Consumer MPU (Laptop/Desktop)	2230	3035	4130	5620	6901	19272	53811
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	1150	1565	2133	2907	3964	18662	85923
Best-Case Test Data Volume (Hierarchal & Compression) (Gb)							
MPU-HP - High Performance MPU (Server)	7.5	8.9	10.3	12.0	12.6	15.1	16.7
MPU-CP - Consumer MPU (Laptop/Desktop)	6.2	7.2	8.5	9.8	10.2	12.1	13.1
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	4.9	5.8	7.0	8.4	9.6	20.6	41.1
Best-Case Compression Factor (Hierarchal & Compression) (Gb)							
MPU-HP - High Performance MPU (Server)	487	564	658	770	905	2106	5319
MPU-CP - Consumer MPU (Laptop/Desktop)	362	420	488	571	675	1595	4118
SOC-CP - Consumer SOC (Consumer SOC, APU, Mobile Processor)	237	269	303	347	414	905	2089

Figure 2: Scan Test Data Compression Factors (Flat with No Compression = 1)

Figure 2 shows the impact of hierarchical and compression scan test techniques on the test data increase. The current compression technologies 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 factors of more than 100X compression are often achieved. However, even a 500X compression won't be enough for SoC (as shown in Table 3 in the Logic Device Testing section); therefore, more sophisticated technologies will be required in the future. Figure 2 shows the level of compression anticipated. The similarity of test vectors applied on scan chains will allow a chance of achieving higher compression ratios. The similarity of test vectors applied in time-space possibly also allows further compression. Thus, utilizing multi-dimensional similarity will be a potential solution.

As shown earlier (Table 2 in the Logic Device Testing section: Logic Test Data Volume), the external scan pins count for SOC-CP is comparatively small and it implies the necessity for scan input/output pins sharing among multiple embedded IP cores. More percentage of scan pins sharing increases the number of cores that can be tested in parallel, and it provides lower data volume and better compression factors.

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. Estimations of these important parameters are shown in the SOC and MPU sections of the previous table (Table 2: Logic Test Data Volume). 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 shorter test time and less memory per pin in the test system. This assumes the scan chain lengths are proportionately reduced.
- Clocking the scan chains at a faster speed also results in a shorter test time but doesn't reduce 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 a given device. This question is impacted by many parameters including device size, personal preference and time to market constraints. As such, the analysis (in Table 2: Logic Test Data Volume) 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 scan data shifting frequency impacts the test time necessary to drive and receive this large volume of test data. As cost effective testers with higher performance are deployed, apparently it seems that scan data shifting can be accelerated and then it can reduce the test time per device. The analysis calculates this impact and suggests that test times will be dropping over time due to this faster scan shifting. 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 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 is a potential solution of the test time reduction. ATPG/DFT level reduction technologies should be developed in the 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 required. There is some research regarding this methodology; however, more research will be required for industrial use.

High-level design languages are being used to improve design efficiency, and it is preferable for DFT to be applied at the high-level design phase. DFT design rule checking, testability analysis and fault coverage estimation are already available to some extent. Those features, including non-scan-design approaches and DFT synthesis in high-level design, are required in the next stage. Furthermore yield-loss is a great concern. As test patterns excite all possible faults on the DUT, it will lead to excessive transistor switching activity, which does not occur in normal functional operations. This will cause excessive power consumption which makes the functional operation unstable, and eventually makes 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 leak current of test circuit itself should also be considered as a part of power consumption.

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.

Requirements for Embedded Memory Cores

As process technology advances, and due to some special application needs, both the number of memory instances and the total capacity of memory bits increase and will cause an increase in area investment for BIST, repair and diagnostic circuitry for memories. As the density and operating frequency of memory cores grow, memory DFT technologies as follows are implemented on SOCs and are factors of area investment increase:

- To cover new types of defects that appear in the advanced process technologies, dedicated optimal algorithms must be applied for a given memory design and defect set. In some cases, a highly programmable BIST that enables flexible composition of the testing algorithms is adopted.
- Practical embedded repair technologies, such as built-in redundancy allocation (BIRA) which analyzes the BIST results and allocates redundancy elements, and built-in self-repair (BISR) which performs the actual reconfiguration (hard-repairing) on-chip, are implemented for manufacturing yield improvement.
- On-line acquisition of failure information is essential for yield learning. A built-in self-diagnostic (BISD) technology distinguishes failure types such as bit, row, and column failures or combinations of them on-chip without dumping a large quantity of test 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.

The embedded memory test, repair and diagnostic logic size was estimated to be up to 35k gates per million bits in 2013. 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 SOCs; 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 a smaller number of memory blocks can reduce memory DFT area investment dramatically. 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.

Requirements for Integration of SOC

Reuse of IP cores 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, IEEE Std 1500); and this format must be preserved and extended to other core types, such as analog cores. The DFT-ATE interface is going to be standardized (for example, IEEE Std 1450), and it should include not only test vectors but also parametric factors. An automatic design and test development environment is required to construct an SOC-level test logic structure and generate tester patterns from the test design information and test data of each IP core. This environment should realize concurrent testing described below.

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 in the near future. SOC-level diagnosis requires a systematic hierarchical diagnosis platform that is available for learning the limiting factors in a design or 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 failure analysis equipment 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 crucial for efficient yield learning, but it is a know-how issue now.

Concurrent Testing

For SOC test time reduction, concurrent testing, which performs tests of a number of IP (non-identical) cores concurrently, is a promising technology. For instance, the long test time of high-speed IO can be mitigated if other tests could be performed at the same time, which would decrease the total test time dramatically. To realize the concurrent testing concept, there are items that must be carefully considered in the product design process. These items include the number of test pins, power consumption during test, and restrictions of the test process. These items are classified as either DFT or ATE required features in Figures 3 and 4 below. IP cores should have a concurrent test capability that reduces the number of test pins (Reduced Pin Count Test: RPCT) without a test time increase, and a DFT methodology which enables concurrent testing for various types of cores. As these requirements differ corresponding to the core types on a chip, a standardized integration method of RF, MEMS and optical devices into a single SOC with conventional CMOS devices can be developed. It includes unification and standardization of test specifications which are used as interfaces by IP vendors, designers, DFT engineers and ATE engineers that can be combined with breakthroughs on analog-mixed signal/RF DFT methodologies (e.g. integrated efficient interfaces for test of the core itself and core test access, or wide adoption of IEEE Std 1500, and its extension to analog etc.)

DFT and ATE must cooperatively consider concurrent testing requirements and restrictions. This may not be an easy task as there are multiple challenges to enable concurrent testing. For instance, ATE software needs to be able to perform concurrent test scheduling after analyzing the power and noise expected during testing based upon design and test information specified for each IP core and chip architecture by the designer.

Figure 3: Required Concurrent Testing DFT Features

Features	Contents
External test pin sharing	Each JTAG-enabled IP core must use the 5 JTAG interfaces (TRST, TMS, TCK, TDI, TDO). Cores that have non-JTAG interfaces must be able to share external test pins with other cores.
Design for concurrent testing	The test structure of an IP core must be operationally independent from that of all other IP cores.
Identification of concurrent test restrictions	The presence of any test restrictions for each IP core must be identified to the scheduler (e.g. some IP cores are not testable at the same time due to noise, measurement precision, etc.).
Dynamic test configuration	Test structures/engines that can change the order of test and the combination of the simultaneous test to each IP core.
Test Data Volume	The test data volume of all IP cores must be able to be stored in the tester memory.
Test scheduling	Critical information on each IP must be available to the test scheduler: a) Test time of each IP core. b) Peak current and average power consumption for each IP core. c) Test Frequency for each IP core.
Common core interface	The test access interface of IP cores must be common among all IP cores (e.g. IJTAG)
Defective IP identification	There must be a mechanism to identify defective IP cores prior to and during test.

Figure 4: Required Concurrent Testing ATE Features

Features	Contents
Numerous Tester Channels with Frequency Flexibility	A large Number of Test channels that cover a wide range of frequencies will enable efficient concurrent testing. Test channels must provide test data such as clocks, resets, data, or control signals to many corresponding IP blocks. Testing can be more flexible if channel assignments are dynamically changeable.
Mixed Data type support	Capability of loading/unloading test data that is a mixed combination of digital, analog, and high-speed I/O data is required.
IP block measurement accuracy	Measuring accuracy of testing (e.g. high-speed I/O test) should be preserved in concurrent testing to match the specifications.
Test Data Handling Efficiency	Test data loadable to each divided test channel should closely match memory usage efficiency as that of non-concurrent test.
Power supply capability	A large number of capable power supply pins will enable large number of IP blocks to be simultaneously tested.
Multi-Site Testing capability	Capability to perform both multi-site testing and IP-level concurrent testing at a time will enable efficient testing.
Capable Software	Automated test scheduling software that can decide test scheduling configurations while considering many constraints is required.

Figure 5: Comparison between Multisite and Concurrent

Pin Count	Production Volume	Efficiency	
		Multisite Testing	Concurrent Testing
Many	Large	Medium	High
	Small	Low	High
Few	Large	High	Medium
	Small	Low	Low
Consideration		Cost of Jig (Initial Cost) - Probe Card, Test Board, etc. Cost of Tester - Pin Count, Power Supply, etc.	Reduction of Test Pins (RPCT) Cost of Chip - Impact on area, etc. Cost of Design

Multi-site testing is another approach to reduce effective test time per die or chip. The effect of cost reduction for each approach depends mainly on the number of test pins and the production volume, as shown in Figure 5 – Comparison between Multisite and Concurrent. Larger number of test pins will make the number of multi-sites

smaller, and higher production volume will get larger profit by the cost reduction. To estimate an accurate profit, cost of jigs and expense on engineering and designing should be also considered.

DFT for Low-Power Design Components

Low power consumption design is indispensable for battery-powered devices to enhance system performance and reliability. The design includes multiple power domains which are independently controlled by PMU (Power Management Unit), and some special cells used for controlling power at the physical level, such as level shifters, isolators, power switches, state retention registers, and low-power SRAMs.

However, the design raises new requirements in testing, which are some dedicated functions in test. For example, an isolator should be tested in both active and inactive mode to fully test its functionality, and a state retention cell requires a specific sequence of control signals to check if it satisfies a specification on power shut-off and turn-on. Please refer to the Figure 6 – Low-power Cell Test – for more low-power cells. Some of the defects on the special low-power cells can possibly be detected in an ordinary test flow but it is usually not enough to assure entire low power features of a design. These functions have not been treated in the historical scan test that only focuses on the structure of circuits. Therefore, a full support of these dedicated test functions for special low power cells is strongly required.

Figure 6: Low-Power Cell Test

#	**Component**	**Test Contents**
1	Isolator	Generate patterns controlling power-on/off of the power domain
2	Level Shifter	Include the cell faults in the ATPG fault list
3	Retention F/F	Generate patterns to confirm saved data after RESTORE operation
4	LP SRAM	Generate patterns which activate peripheral circuit inside the macro during the sleep mode and confirm cell data retention
5	Power Switch	Generate patterns to measure IDDQ with domains power on/off

Summary

SOC test difficulty will rise depending on the complexity and size of the chip. Adoption of new process technology or new devices such as MEMS, and high-quality test requirements for automotive or other application areas, also will introduces new test challenges which must be considered comprehensively.

For logic cores, since test pattern size will significantly increase for new types of defects and high-quality test requirements, hierarchical test logic structures and higher scan compression ratios will be essential. For memory cores, more sophisticated built-in features of test, repair and diagnostics are required. Introducing new types of embedded memory devices requires studies on necessary test sets. While test and DFT methodologies for logic and memory cores have been basically established, more studies are required for other types of cores such as analog, RF, etc.

Test cost reduction using concurrent testing necessitates standardized test structures and tool supports for them. Low power design trends introduce various design methodologies which make testing more complicated. Consistent automated design flow based on a standardized power format is required.

Section 11: 2.5D & 3D Device Testing

This section will address six key test challenges, based on the evolution of 2.5D/3D, from complex die stacks through SiP. These test challenges include: test flows, cost and resources; test access; testing heterogeneous die individually and in a single stack/package; debug and diagnosis of failing stacks/die; DfX (Design for Test, Yield, Cost); and power. It is important to note that 2.5D/3D is not yet a mature and mainstream technology, and, because of that, it is difficult at this time to make any predictions regarding 2.5D/3D test flows. 2.5D and 3D technologies are characteristic of a system, and because of that, should be tested like a system: testing the complete package at an application level and diagnosing failures at the die and interconnect level.

Memory die stacks (Wide I/O, High Bandwidth Memory and Hybrid Memory Cube) were precursors to 2.5D and 3D. Both of these technologies have provided insights to requirements and challenges associated with 3D and 2.5D test. The best that can be gleaned from these technologies at this time is that reliance on BIST and boundary-scan based technologies, and use of fault tolerance with simple configurations, tends to produce relatively high yields at the stack level. As these adjacent technologies become more mature and as more 2.5D/3D-TSV applications emerge, more and better data will enable better predictions and decision making, with respect to 2.5D/3D-TSV test processes.

Executive Summary

2.5D and 3D/TSV are the next evolution beyond SiP/SoC. There have been significant advances from both academics (research) and industry (standards and working models/test chips) to identify and resolve challenges to testing 3D/TSV devices. In the medium to long term, as TSV-based die stacking becomes more prevalent and more complex/exotic die stacks appear, test challenges will also become 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 discuss challenges for testing, including cost (dollars, resources and yields); design for test; test access; debug and diagnosis; and heterogeneous device testing.

Difficult Challenges for Test: 5 years

While the current state of 2.5D/3D seems to be maturing, new enabling and supporting technologies will require advances in test access, capabilities and costs. These emerging technologies will provide significant challenges for testing 2.5D and 3D technologies, in addition to supportive technologies. The challenges below represent potential impacts to test, including increased costs, test times, and reduced yields and reliability.

2.5D Test Challenges (Short Term)

- **Known Good Die Test**: While logic blocks in the die can be partitioned and effectively tested, testing interactions between the logic blocks requires an application-based test.
- **System level test/diagnosis/repair**: Emulating a system-level test environment can be considerably costly and time consuming. Complex integration can produce new and challenging defects for test.
- **Access to the individual die in the 2.5D assembly**: Access may include multiple protocols: IEEE-1149 based (including IEEE 1687.1 and IEEE P1838 proposed standards), photonics, various component and system-level protocols, including I2C and SPI.
- **Interposer testing**: Point to Point testing can be accomplished primarily by probing (see the probing section). Multipoint interposer testing requires significantly more probing (more time and higher costs) and requires embedded logic to coordinate point-to-multipoint connections. Academic research from Duke University[1] has proposed solutions to point-to-multipoint connections on the interposer.
- **High speed interconnects (photonics)/signal integrity**: while testing photonics interconnects is mature, the cost of test is significant, primarily due to equipment and knowledge.
- **Discrete components**: more specific to high speed interconnects through the interposer. Point to point, high speed interconnects may compromise probe-based connectivity test.
- **Impact of emerging technologies with respect to test**: New technologies can impose new defects on the 2.5D. A table presented by Li Li at ECTC describes the relationship to the new technologies (Die thinning, TSV, C2C connection, BS-RDL, micro-bumps, Large volume of copper, increased power density and large thermal gradients, removal of IO structure) and the failure/defect mechanisms. See Table 1 below.

- **Innovations in Wafer Level Packaging/Wafer Level Fan Out**: similar to the statements above: advances in Wafer Level Packaging open the door for new defects and new requirements for test. Aspects of wafer level packaging impinge on interposer technologies.

3D Test Challenges (Short Term)

- **Known Good Die test**: Known good die testing becomes more important in a stack integration, where a defective logic die could compromise the entire die stack. System/application level testing at the die level has significant challenges. Significant Design-for-Test and Built-in Self-Test features are required to complement and simplify system/application level test. Cost of test ($$ and time) are significant.
- **System level test/diagnosis/repair**: System and application level test of a 3D die stack may require significant test/tester resources in order to emulate the component/test environment. These resources may consume significant time and money.
- **Diagnosis of system level failures** on the stack will have significant challenges. On a positive note, a diagnostic die may be added to the stack without significant effort. Redundant die may help as well.
- **TSV/interconnect testing**: TSV testing can be done pre-bond, microvoid and pinhole defects. Mid-bond testing becomes quite challenging. Some probing could help. A boundary-scan stack architecture may be a possibility. An IEEE standard is in the works (IEEE P1838 – this will be covered later as part of "technical issues"). Application and/or system test most likely will not be feasible or practical with part of the stack still missing. Test results may not be informative.
- **Probing 3D die stacks**: This area seems to be maturing. Companies like Tezzaron have probing figured out, since microbumps also seem to be maturing.
- **Stack repair**: While stack repair is now possible, cost for stack repair and re-test would be prohibitive. Redundant die may simplify stack repair.
- **High speed signaling/signal integrity/interconnects**: Signal integrity measurements from die to die are challenging. Design for Signal Integrity Test, at the edge of the die, may be a solution for SI measurement. Costs could be significant at the die level and higher at the stack level. There may be impact to stack level power.
- **KGD/KGS**: Known Good Die is critical from a supply chain perspective. Known Good Stack also requires a stack level system/application test.
- **Wafer Level Packaging** – uncertain if there will be an impact to test.

Difficult Challenges for Test: 10, 15, 25 years

Over the next 10 to 25 years, it is expected that requirements for speed and power will be significantly higher than the current state. Significant advancement to the integration of massive, high performance, low power die will be required for the future versions of application specific integrated devices. Test will be significantly challenged.

2.5D Test Challenges (Long term)

- **Known Good Die Test**: From ITRS predictions, the number of flops per KGD are significantly higher (exponential) over time. Speed, power and thermal scale with logic. Device cooling during test will become critical. Timing may be an issue through the stack. Test time may be significant.
- **System level test/diagnosis/repair**: System test requires higher speeds. Longer test times due to significantly more logic in the die. IEEE 1149 protocols will be obsolete or antiquated at best. Optical protocols using IEEE P1838 test access protocols will replace current test protocols.
- **Access to the 2.5D assembly**: Photonic protocols, more in line with P1838. Protocol objective is access to interposer of multiple-die stacks. Multiple-die stacks create a significant, interconnected logic pool.
- **Interposer testing**: Testing becomes logic based. Probing becomes more challenging, from Point-to-Point to Large Multipoint.
- **Cost of test** (equipment/resources/time): Cost of test increases as high-speed interconnects pass through the optical interposer. Signal integrity adds cost and time.
- **Testing MEMS- and Sensor-based die**: technology is somewhat analog. Testing is not conventional and requires an active motion, light, sound source. Significant setup and test times while income/prices are low and volumes are extremely high.

3D Test Challenges (Long term)

- **Known Good Die(s):** similar to the 2.5D section, logic per die is increasing over time. Speed, power and thermal are scaling with increasing logic/flops. Heterogeneous die become more and more exotic, making test more challenging, and determining when the die becomes "known good".
- **System level test/diagnosis/repair**: Test equipment cost increases significantly with respect to dollars and time. High speed signaling increases test costs and the possibility for errors/failures. Potential impact on the supply chain.
- **Carbon Nanotubes** replace TSV, CNT test replaces TSV Test: Carbon nanotubes help to sustain logic, speed, power, thermal characteristics. CNT is still academic; however, more prognostication exists for short- to mid-term replacement of TSVs with Carbon Nanotubes. External interfaces convert protocols to CNT interface.
- **Mega Stack testing**: Mega Stacks address the need for high performance data processing. Stacking multiple die becomes feasible as advances in die to die bonding minimize the stack size, while adding more die to the stack. Future Mega stacks may potentially have the same structural integrity as initial die stacks. Multiple redundant die may be necessary for more probable defective die in the stack, or mis-connections between die in the stack.
- **Probing die and stacks**: Advanced Probing technology will be needed as TSVs and micro-bumps become significantly smaller.
- **Stack repair**: Primary repair will come from redundant die. Dis-assembly of the die stack becomes significantly challenging and could potentially destroy the pre-bonded die.
- **Testing MEMS- and Sensor-based die**: technology is somewhat analog. Testing is not conventional and requires an active motion, light, sound source. Significant setup and test times while income/prices are low and volumes are extremely high.

Discussion of Key 3D Test Technical Issues

Future of 3D Integration (Test and DfT focus, Academic Perspective)

The semiconductor industry has been able to meet the demand for high-performance integrated circuits (ICs) with added functionality by relentlessly scaling device sizes. However, it is becoming increasingly difficult to sustain device scaling in an economically viable manner.

A promising way to achieve high-performance ICs with more functionality and reduced die footprint is through 3D integration. Today's 3D integration process is primarily based on die/wafer stacking, as it does not require substantial changes to the existing fabrication flow. In this process, separately manufactured dies/wafers are integrated onto the same package, and through-silicon-vias (TSVs) are used to connect dies to each other. Considerable research efforts have therefore been directed toward the development of TSV-based 3D stacking technology, and products based on this technology have been successfully introduced into the market, e.g., the AMD Fiji chip. However, the keep-out-zone (KOZ) required for TSVs and limitations on the die alignment precision impose limits on the device integration density that can be achieved using TSV-based 3D stacking. A minimum KOZ of 3 μm is required for ICs fabricated at the 20 nm technology node [Kannan et al. 2015], and the die alignment precision is currently limited to 0.5μm.

Monolithic three-dimensional (M3D) integration is receiving considerable interest as a technology for the future, as it has the potential to achieve higher device density compared to TSV-based 3D stacking. In this technology, transistors are processed layer by layer on the same wafer. Sequential integration of transistor layers enables high-density vertical interconnects, known as the interlayer vias (ILVs). Typically, the size and pitch of an ILV is one to two orders of magnitude smaller than those of a TSV [Batude et al. 2012]. To realize such high-density vertical interconnects, the interlayer dielectric (ILD) thickness is being aggressively scaled [Batude et al. 2012; Lee and Lim 2013], and such scaling has been shown to lead to electrostatic coupling between device layers. This is a challenge for test researchers.

Researchers have recently analyzed electrostatic coupling between device layers in M3D ICs and quantified its impact on circuit timing [Koneru 2017]. Device simulations have been carried out to understand the impact of coupling on the threshold voltage of a top-layer transistor for both transistor- and gate-level integration. To realize a new silicon layer over the bottom layer without damaging the underlying interconnects and degrading the properties

of the bottom-layer transistors, several layer-transfer techniques are being explored [Batude et al. 2015; Ishihara et al. 2012].

Low-temperature wafer bonding is a key processing step in these techniques. The condition of the bonding surfaces plays a crucial role in achieving a defect-free bond. Oxide layers are the prime candidates for bonding surfaces due to the presence of hydroxyl (OH) groups that lead to high bond strengths. These oxide layers also act as the ILD. Therefore, defects that arise during the wafer-bonding step can impact the top-layer transistors, as well as the ILVs. It is important to understand and analyze wafer-bonding defects, and develop methods to test for these defects. Research is needed to study the impact of bond defects on the threshold voltage of a top-layer transistor and on the ILVs. In addition, advances in test access and debug/diagnosis for M3D will also be of growing interest as this technology advances.

It has been shown thus far that the impact of coupling and wafer-bonding defects on the threshold voltage of a top-layer transistor is significant, and cannot be ignored, when thickness of the ILD is less than 100 nm. In such scenarios, the paths through the top layer in a gate-level-integrated M3D IC can change depending on the size of the defect and the voltage on the metal lines in the bottom layer. The presence of defects at the bond interface can lead to a change in resistance of an ILV and in some cases lead to an open in the ILV or a short between two ILVs. A resistive open in an ILV or a resistive short between two ILVs can have a significant impact on the path delays. Due to these challenges, existing test-generation methods for small-delay defects are of limited effectiveness when the ILD is less than 100nm.

There is also a paper which includes a discussion regarding supply-chain capability requirements for test and reliability: see [Alfano].

References

[Kannan 2015] S. Kannan, R. Agarwal, A. Bousquet, G. Aluri, and H. S. Chang. 2015. Device performance analysis on 20nm technology thin wafers in a 3D package. In Proceedings of the 2015 IEEE International Reliability Physics Symposium. 4C.4.1–4C.4.5. DOI:http://dx.doi.org/10.1109/IRPS.2015.7112735

[Batude 2012] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P. E. Gaillardon. 2012. 3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS. IEEE Journal on Emerging and Selected Topics in Circuits and Systems 2, 4, 714–722. DOI:http://dx.doi.org/ 10.1109/JETCAS.2012.2223593

[Lee and Lim 2013] Y.-J. Lee and S. K. Lim. 2013. Ultrahigh density logic designs using monolithic 3-D integration. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 32, 12, 1892–1905.

[Koneru 2017] A. Koneru, S. Kannan and K. Chakrabarty, "Impact of electrostatic coupling and wafer-bonding defects on delay testing of monolithic 3D integrated circuits", ACM Journal on Emerging Technologies in Computing Systems, vol. 13, no. 4, Article 54, July 2017.

Alfano, Black, Rearick, Siegel, Su, and Din “Unleashing Fury: A New Paradigm in 3D Design and Testing”

Section 12: Burn-In and Reliability Testing

The objective of reliability solutions and burn-in is to eliminate latent defects in ICs that will cause early-life failures and screen them out before the product is shipped to the customer. Reliability screens are critical to achieving the low failure rates required by high-reliability applications, such as automobiles. The visibility of automotive applications is increasing dramatically, as the electronics in cars includes not only engine and brake control, but also communications, internet access, entertainment, GPS, collision avoidance, and many other nascent applications. Reliability requirements in other mobile and mass storage applications is also increasing in importance, as the number of ICs and their transistor density increases. Latent defects are typically removed by process improvements, design improvements and accelerated stress methods during the test process. 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 meet the reliability specification and the needs of the end customer while providing the best value for the reliability dollar spent.

Burn-In and Reliability Testing Goals

In reliability circles, customer satisfaction is measured by the field failure rate or failures in time (FITs). The cost of reliability screening 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 three fundamental components – burn in duration, BIB/socket cost 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.

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 escapes from the manufacturing stress process.

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. The best way 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 and functional test at stress conditions. 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 may operate fine at application conditions. Although these parts may contain no reliability defects, in situ stress will fail these perfectly functional parts – hence over-kill. Determining the proper test method and interpretation of the test results are key ingredients of a successful in-situ burn-in strategy to ensure that the latent defects are identified and overkill is minimized. These same parts with "marginal margin" are the target of 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. 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 electromigration for C4 and BEOL wiring requires careful scrutiny when confronted with the radical currents and powers conjured up by stress conditions.

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 consume 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.

Burn-In and Reliability Testing 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. In addition to burn-in, several alternate techniques such as IDDQ, high voltage stress, and wafer mapping are being used to try to improve device reliability.

Burn-in system technology must continue to evolve with device technology. 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. Without continued innovation by the burn-in system manufacturers in cooperation with the IC manufacturers, all these trends tend to increase the cost of burn-in systems and sockets.

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 1). In certain applications, system-level burn-in complements or replaces traditional device-level burn-in, but this typically involves a significantly increased cost, since the burn-in system, socketing solution and burn-in time tend to increase. Wafer-level burn-in technology continues to be developed, but has made only limited inroads against traditional package-level burn-in. The challenge here is to use techniques such as scan/logic and memory BIST (MBIST) to improve the technical feasibility of wafer-level burn-in.

Table 1: Burn-in Requirements

Year of Production	2018	2019	2020	2021	2026	2030
Packaged Part Burn-in						
Clock input frequency (MHz)	400	400	400	400	400	400
Off-chip data frequency (MHz)	75	75	75	75	75	75
Power dissipation (W per DUT)	600	600	600	600	600	600
Power Supply Voltage Range (V)						
High-performance ASIC / microprocessor / graphics processor	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5	0.4–2.5
Low-end microcontroller	0.5–10	0.5–10	0.5–10	0.5–10	0.5–10	0.5–10
Mixed-signal	0.5–1000	0.5–1000	0.5–1000	0.5–1000	0.5–1000	0.5–1000
Memory	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5	0.5-12.5
Maximum Number of Signal I/O						
High-performance ASIC	384	384	384	384	384	384
High-performance microprocessor / graphics processor / mixed-signal	128	128	128	128	128	128
Commodity memory	72	72	72	72	72	72
Maximum Current (A)						
High-performance microprocessor	450	450	450	450	450	450
High-performance graphics processor	200	200	200	200	200	200
Mixed-signal	30	30	30	30	30	30
Memory	10	10	10	10	20	20
Vector memory depth (M vectors – DFT/BIST SOC *2)	256	256	256	256	256	256
Maximum burn-in temperature (ºC)	175±3	175±3	200±3	200±3	200±3	200±3
Burn-in Socket						
Pin count	3000	3000	3000	3000	3000	3000
Pitch (mm)	0.08	0.08	0.08	0.08	0.08	0.08
Power consumption (A/Pin)	6	6	6	6	6	6
Wafer Level Burn-In						
Maximum burn-in temperature (ºC)	175±3	175±3	200±3	200±3	200±3	200±3
Pad Layout – See Probe Table						
Power consumption (KW/wafer)						
Low-end microcontroller, DFT/BIST SOC *2)	30	30	30	30	30	30
Memory	5	5	5	5	8	9
Maximum number of Signal I/O (Commodity memory)	45	45	45	45	45	45

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

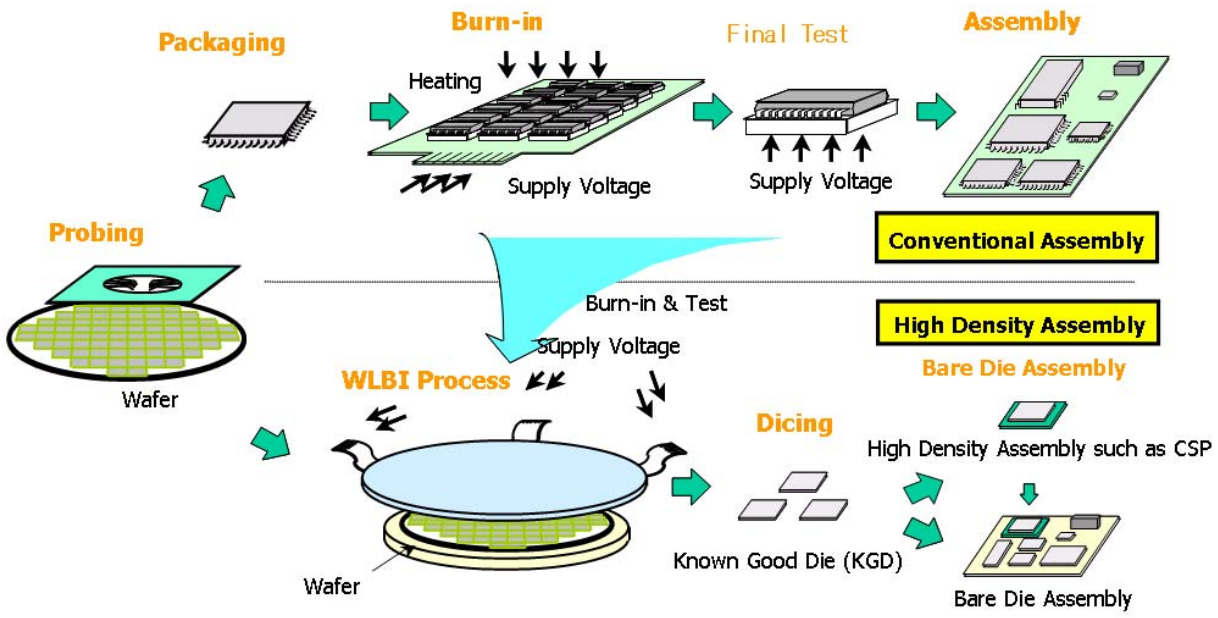


Figure 1: The Production Process with WLBI Compared with Package Burn-in.

Flash

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 just 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, especially stacked die in mobile and mass storage applications. Reliability failures of the packaged part increase exponentially with the number of die in a multi-die package, so the need for reliable die before packaging is increasing substantially in importance. 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.

There are two methods of performing a wafer-level burn-in. 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. This is typically referred to as wafer-level stress, as it applies a stress voltage for a short time, but does not apply the high temperature of burn-in. Actual 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. DFT functions such as scan or BIST are enablers for WLBI.

The challenge for DRAM for example, 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 does not increase the cost of the final part. Leveraging the time spent in burn-in by using the burn-in environment as a massively parallel testing opportunity can effectively lower the overall cost-of-test.

Probing Technology for Wafer Level Burn-in

Full-wafer probing is a significant challenge, both technically and economically. The cost of a full-wafer probe tends to increase as the number of pads on the wafer increase and the pitch of the pads decreases. Contacting all the pads on a state-of-the-art wafer can require contacting in excess of 250,000 pads across a 300 mm wafer at a pitch of 60 microns or less over a wide temperature range. Intelligent use of DFT and pad placement rules by the semiconductor manufacturer can make this challenge less daunting. A WLBI 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. Other vertical pin contactors operate in a similar manner. The probe pitch is technology dependent.

For a pitch less than 70 μm, MEMS technology by use of photolithography is an option. This technology, however, is very challenging 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.

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.

Other WLBI Technology Considerations

The current consumption of a wafer is increased 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 in certain applications. 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 can 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 JTAG1, the number of tested pins per device and total cost per device must be decreased and performance of the WLBI technology must be improved.

References

1. IEEE standard 1149, Boundary Scan

Section 13: Test and Yield Learning

In the normal sorting function, test provides the essential feedback loop for yield-learning. Product-based diagnostics, product-like test chips and parametric-sensitive test structures all play a key role.

Key Cost of Test Trends

Value derived from diagnostics of actual product hardware is driven by systematic defect mechanisms that are now increasingly complex functions of neighboring shapes, local pattern densities, etc. As a result, some failure mechanisms may be visible only on product. In addition, product-based diagnostics automatically places focus on key yield-limiting failure mechanisms. Volume-based diagnostics are important since individual occurrence of any given systematic defect mechanism may be rare. The pooling of data across many failing die can be important to identify true systematic defect mechanisms.

Product-like test chips can provide some of the same insight for yield-learning, but have the advantage of being available earlier, even when design is on-going. Specifically, rapidly designable, scalable and 100% testable & diagnosable test chips, with and without embedded memory and other key IP blocks, including fast automated design methodologies, are required to accelerate yield ramps and first-time yield success of complex SOCs. Such test chips should play the role of "send-aheads" and be designed on early foundry testsites even while the product design is ongoing. The test chip should be scalable, in that a complete SOC-style (optionally, timing-closed) design is possible with tens or hundreds of standard cells and with a small or large compiled memory and other IP blocks. The test chip should enable both logical and physical layout diversity in order to capture layout topologies found on real product chips. Finally, the test chip must be able to maintain a stable test and diagnosis infrastructure, meaning the same set of ATPG, diagnosis and failure analysis capabilities should be enabled whether the test chip is tiny ($<1mm^2$) or huge ($>100mm^2$).

In addition, parametric-related feedback is needed for (1) device and interconnect 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) increases the negative impact of scribe-line-to-chip offsets. Moreover, test structures are limited by the number of configurations they can cover. Variations in configurations include both physical variations and electrical variations, such as different gate types and differences in load characteristics. As circuit parametrics are increasingly affected by such configurations, including within-standard-cell and transistor-layout configurations, it becomes necessary to base learning on product test or test of product-like layout configurations. Embedded, distributed monitor circuits such as thermal and VDD sensors, process-monitoring ring oscillators and critical path proxies 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, V_t, source-drain resistance, etc.) The spatial component includes both die-to-die and within-die components.

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 sufficiently sensitive for analog/RF are needed. Moreover, 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.

Top challenges for test-based yield-learning include:

- Better resolution for cell-internal defects. Latest advances in structural testing and scan-based logic/layout-aware diagnosis methods are adequately addressing interconnect and via defects. Statistical approaches built into volume-based diagnostics are able to predict interconnect-related defect modes without an over-dependence on Physical Failure Analysis (PFA). Innovation is required, however, for cell-internal-defect-targeted diagnostics to be able to identify systematic fail modes inside standard cells. Observations derived from production silicon suggest a shift toward a larger percentage of the defect distribution being cell-internal defects, as opposed to interconnect-related. Current best methods for cell-internal defect diagnostics are cell truth-table and gate-exhaustive model-based, with the truth tables established via SPICE simulations of modeled cell-internal parasitics. These methods suffer from aliasing issues and over-reliance on potentially inaccurate modeling of cell-internal defects used in SPICE simulations. In addition, diagnosis resolution needs to be better due to limitations in the PFA process.

- Managing design data for yield learning. A tremendous amount of design data can be brought to bear for yield learning purposes, but it is often not organized effectively for this purpose. In addition, with hierarchical design and DFT flows, the overall management of this data at most companies today is ad-hoc, limiting its effective use.
- Inadequacy of LEF/DEF as the basis of layout-aware diagnosis. LEF/DEF suffices for the purposes of layout-aware ATPG but is too early in the design cycle to be used effectively for layout-aware diagnosis. LEF/DEF is less likely to closely resemble the final mask shapes due to complex OPC, boolean and retargeting steps.
- Yield-Learning in an OSAT/Fabless/Foundry environment. Yield-learning capabilities must be cognizant of the environment that has become the dominant model for our industry. If the technology cannot deal with the security and logistical concerns of this environment, it cannot be effective. 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 for scan fail data, and infrastructure to support their transmittal are needed to support automation and sharing of data. Specifically, data exchange standards are needed between the Fabless and the OSAT/Foundries to share system-level test feedback and correlation to wafer-level test and measurement data to (1) improve IC quality and reliability, (2) correlate process variations and parametric variability, and (3) reduce overkill. Distributed design, manufacture and test also creates an emerging role for methodologies and tools to help determine which areas that problems reside, e.g., design house, foundry or OSAT.
- Test for ZERO DPPM/Automotive in advanced node technologies. A change of mindset away from structural test coverage only is required to guarantee functional safety for automotive ICs. Mission-mode in-situ MBIST and LBIST and Design Failure Mode and Effects Analysis [DFMEA] are already part of ISO26262 (an automotive-specific set of standards for designing and testing electronics that focuses on safety critical components), but test architecture research is required to minimize the die-footprint increase due to added circuit redundancy. Moreover, Automotive may require root cause reports to be produced quickly for field failures. This requirement is another driver for rapid diagnosis and root cause analysis.
- Test and data-collection time increases due to longer scan chains. These increases drive a need for focus on LBIST methodologies and scan compression for both test and diagnosis.
- Faster Memory BIST bitmapping.
- Guidance for trading off test and data collection time against improved failure diagnostics.

Section 14: Cost of Test

Minimizing costs are a key part of the semiconductor manufacturing process. Test is no exception, although steady improvements in efficiencies over the last 15 years have lowered the typical cost of test as a percentage of IC revenue to less than 2-3%. The primary drivers of increased efficiency have been reductions in capital costs per resource and test times, coupled with increases in parallelism and Built-In Self-Test (BIST) capability. Most SOC device are tested 2 to 16 at a time, and memory devices can have more than 1,000 devices tested at once. Measured as the cost to use capital equipment for test (in terms of cost per hour per device), these decreases in test cost will continue at a relatively consistent rate per year. The figure below shows the historical rate of capital investment in test, interface (consumables) and handling equipment.

It is notable that, in 2015 for the first time, the cost of consumable material has become the leading capital expenditure relative to ATE-based test. This has to do with the increased cost of interface material (primarily influenced by probe cards and relative items) and the decreasing depreciation period for materials utilized for the production of devices used in the mobile device space where devices have a shorter life span. In this case, material is typically discarded not because it has ceased to function, but rather because the devices it is used to test are replaced by newer versions for end devices like mobile phones.

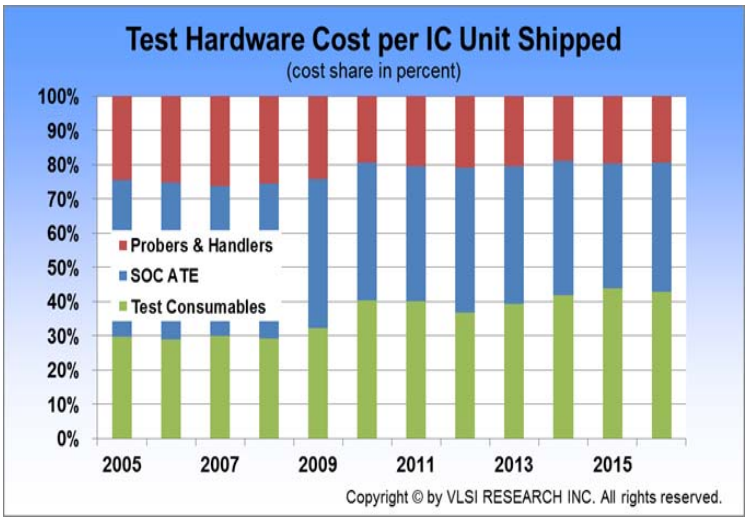


Figure 1: Test Costs as a percentage of device ASP (Used by permission of VLSI Research)

Key Cost of Test Trends

Looking forward, there are several trends which will counterbalance equipment efficiency and serve to cause cost increases:

- Increases in transistor count that outstrip compression technology will increase the amount of external data which must be supplied to the Device Under Test (DUT). Coupled with scan shift rates that are limited by power and thermal concerns, the overall effect will be longer test times. This will be addressed primarily with increased parallelism.
- Device configuration and one-time programming during test is causing more time to be spent during test to perform initial device calibrations or to reconfigure devices based on defects or electrical performance. As silicon geometries shrink and defect densities drive circuit redundancy, repair functions will also add to test costs.
- The eventual drive to multi-die packages will add a requirement for more System Level ("mission mode") testing owing to lack of access to individual die. Without significant Design For Test (DFT) improvements, this type of testing can take much longer than conventional structural test. This will also drive more exhaustive test processes at wafer probe in order to improve the yield of multi-die packages.
- Site count increases at probe test are not able to increase owing to the attendant increase in the cost of consumable material (discussed above) and the limitations of Touch-Down Efficiency (TDE). This is discussed in more detail below.

- An increasing reluctance on the part of IC manufacturers to dedicate silicon area and power to circuitry used exclusively for test.
- The continuing increase of silicon content in automotive applications, especially for safety systems, which drives additional test insertions for fault coverage and temperature-related test.

Even though continuous improvement in equipment efficiency will be offset by new device test requirements, the overall cost of test will continue to decrease. The major contributors to that cost are described below.

Cost of Test as a Part of Overall Manufacturing Cost

While the cost to own and operate test equipment has been reducing, other semiconductor manufacturing costs have been significantly increasing with new silicon technology. Specifically, fab costs for leading-edge processes have increased to about 70-80% of the overall cost of producing a large-scale SOC device. It now costs far more to fab a device than to test it, and that trend will accelerate as new fabrication technologies are deployed.

The figure below represents third-party analysis of the capital and service costs of equipment used in device fabrication, packaging and test.

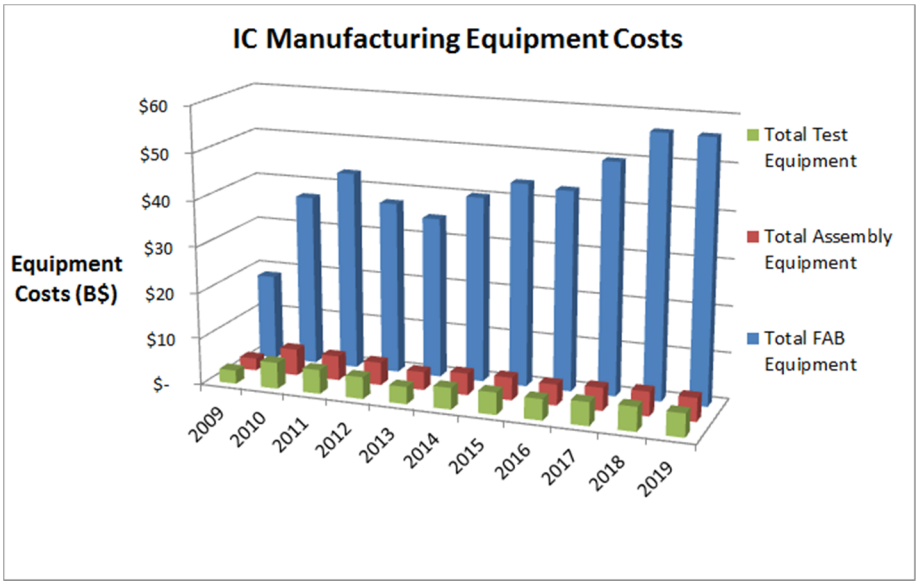


Figure 2: Relative cost of Fab, Packaging and Test Equipment

While it is helpful to focus on the cost of test itself, the overall contribution to a manufacturer's profitability from lower test costs will be very small since test is a small part of the device cost overall. The highest avoidable costs in test are devices that are good but are rejected at test for some reason.

Consider the following, simplified example.

- A device costs $1.00 to manufacture, including Fabrication, packaging, etc.
- Test constitutes 5% of that cost, or $0.05

Reducing the cost of test by 10%, will reduce overall costs by $0.05 X 10% = $0.005 per device

Improving yield by 1% reduces overall cost by $1.00 * 1% = $0.01 per device

While the 10% Cost of Test reduction is good, the yield improvement is better.

The figure below shows the effect on cost of test of traditional cost reduction techniques:

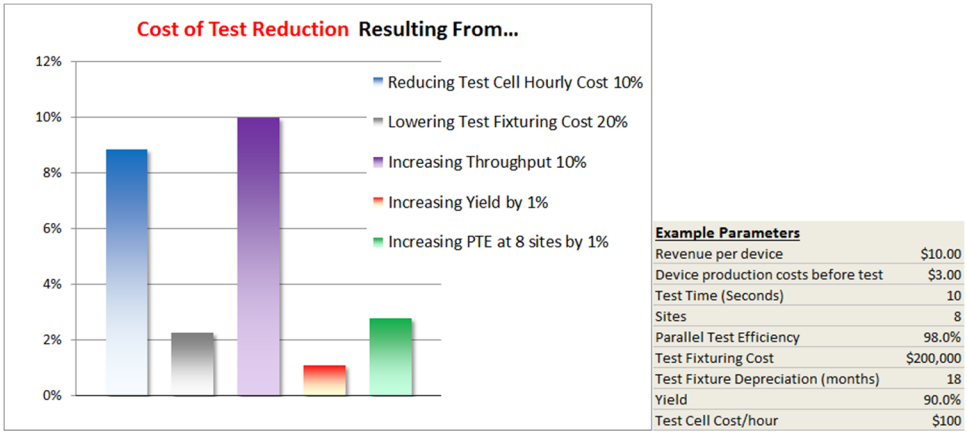


Figure 3: Cost of Test Reduction realized by traditional cost reduction techniques

If one considers the effect on total manufacturing costs, including the cost to scrap devices that are actually good, the cost savings due to improved yield becomes far more significant.

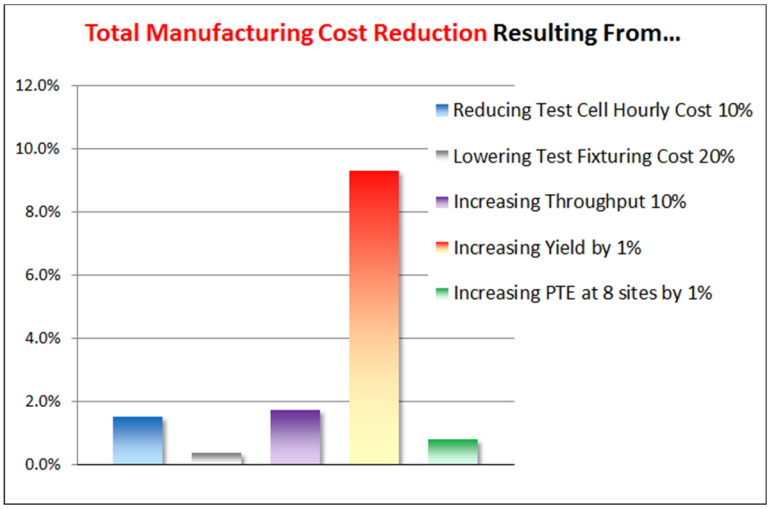


Figure 4: Total Cost of Manufacturing Reduction realized by traditional cost reduction techniques

The risk of yield loss is increasing over time for several reasons:

- Trends such as the reduction of power supply voltages and more complex RF modulation standards will drive higher accuracy requirements for test equipment. Test equipment accuracy is typically added as a "guardband" in testing, reducing the range of acceptable measurements. If measured DC and AC values become smaller and there is no improvement in test accuracy, this guardband will cause more marginal (but good) devices to be scrapped.
- As noted earlier, many devices, especially for mobile applications, require some sort of calibration or trim during the test process to improve DC and AC accuracy. This dramatically increases both the number of measurements made and the accuracy required of the test equipment. The requirements increase the chance of discarding devices that would otherwise have been good.
- Faster production ramps and short IC product life cycles will reduce the amount of time available to optimize measurements for the majority of devices produced.

The remainder of this section will examine Costs associated with owning and operating test equipment. It must be stressed that reducing these costs must be done in the context of the overall cost to produce devices and balance reduction in test costs with potential reductions in product yield.

Test Cost Models and Cost Improvement Techniques

The cost of semiconductor test has many drivers, which is further complicated for multi-die SiP precuts as shown in Figure 5.

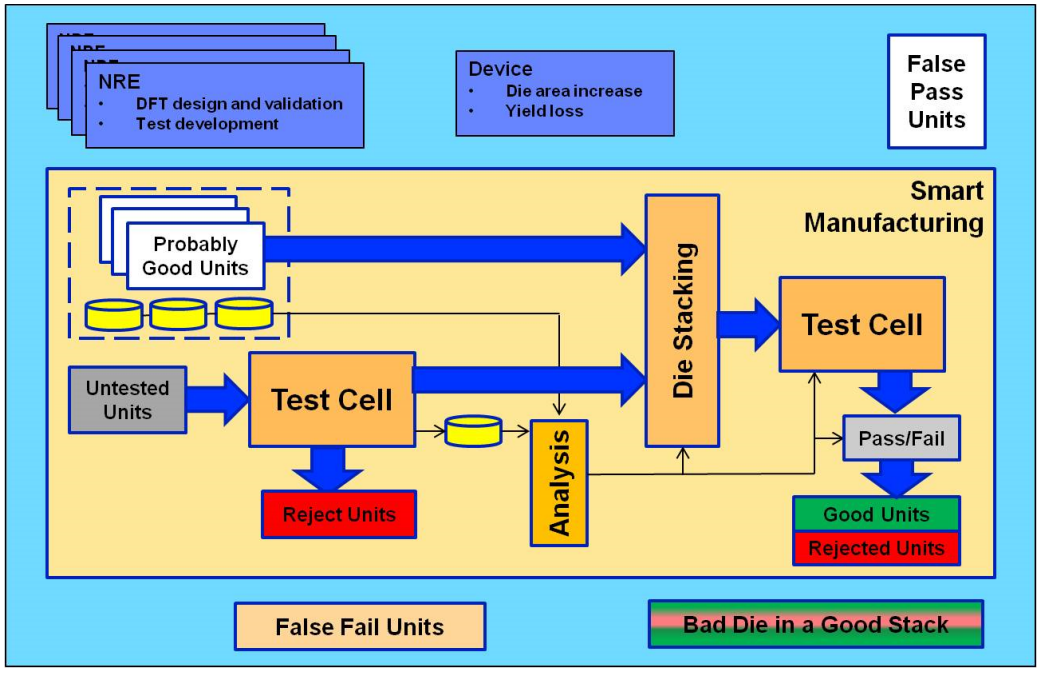


Figure 5: Multi-die Flow

Current Top Cost Drivers

The traditional drivers of Test Costs typically include (In rough order of impact to Cost)

- Device Yield
- Test Time, site count and Parallel Test Efficiency (PTE)
- Overall Equipment Utilization
- ATE Capital & Interface Expenditures
- Facility/Labor costs
- Cost of Test Program Development
- Cost of die space used for Test-only functions

Future Cost Drivers

- Increased test time due to larger scan patterns
- Increased testing at wafer to produce Known Good Die (KGD)
- Addition of system-level testing
- Increased cost of handling equipment to support high site count or singulated die
- Increasing use of device calibration/trimming at test or device repair with redundant components

Currently Deployed Cost Reduction Techniques

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

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

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%.

Typical site counts for various device types are shown in the ITRS "Site Count Table 2017". We also looked at the roadmap plans from 2013 and compared them to 2017 ("Site Count Comparison 2013 to 2017"). This clearly shows how increased device complexity as well as device interface complexity and costs have constrained efforts to expand site counts as quickly as desired.

Minimizing costs are a key part of the semiconductor manufacturing process. Test is no exception, although steady improvements in efficiencies over the last 15 years have lowered the typical cost of test as a percentage of IC revenue to less than 2-3%. The primary drivers of increased efficiency have been reductions in capital costs per resource and test times, coupled with increases in parallelism and Built-In Self-Test (BIST) capability. Most SOC devices are tested 2 to 16 at a time, and memory devices can have more than 1,000 devices tested at once. Measured as the cost to use capital equipment for test (in terms of cost per hour per device), these decreases in test cost will continue at a relatively consistent rate per year. The figure below shows the historical rate of capital investment in test, interface (consumables) and handling equipment.

It is notable that, in 2016, for the first time, the cost of consumable material had become the leading capital expenditure relative to ATE-based test. This has to do with the increased cost of interface material (primarily influenced by probe cards and relative items) and the decreasing depreciation period for materials utilized for the production of devices used in the mobile device space where devices have a shorter life span. In this case, material is typically discarded not because it has ceased to function, but rather because the devices it is used to test are replaced by newer versions.

Figure 6: Importance of Multi-Site Efficiency in Massive Parallel Test

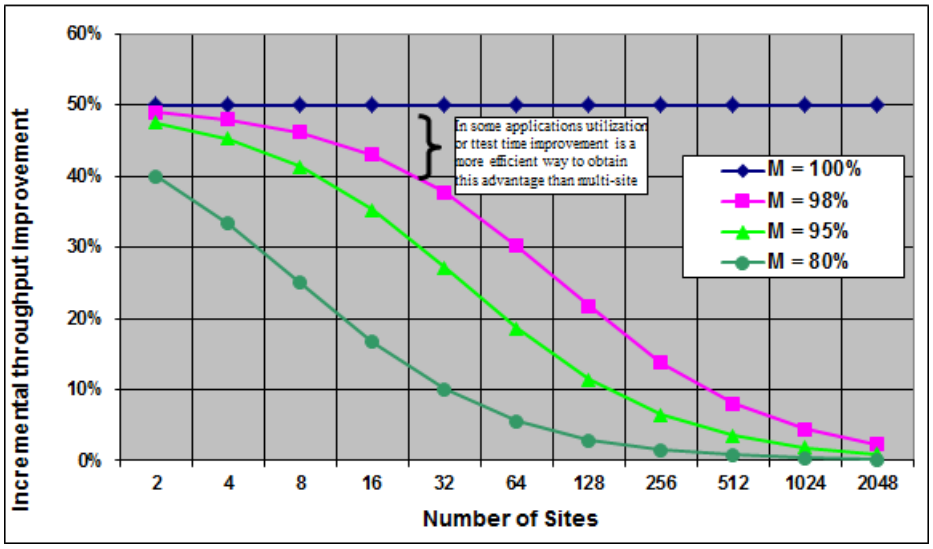


Table 1: Multi-site Test for Product Segments

		2018	2019	2020	2021	2026	2030	Drivers
High Performance MPU, ASIC (1)								
Wafer test	Number of sites	8	8	8	8	8	8	
Package test	Number of sites	16	16	16	16	16		MPU
SoC (2)								
Wafer test	Number of sites	8	8	8	16	32	32	SoC
Package test	Number of sites	16	16	16	32	64		SoC
Low Performance - MCU, MPU, ASIC (3)								
Wafer test	Number of sites	64	64	128	128	256	256	MCU
Package test	Number of sites	16	16	32	32	64	64	MCU
Mixed-signal, & Communications								
Wafer test	Number of sites	16	16	16	16	32	32	Mixed
Packaged Test	Number of sites	16	16	16	16	32	32	Mixed
DRAM Memory								
Wafer test [note 4]	Number of sites	3000	3000	3000	3000	3000	3000	DRAM
Packaged Test	Number of sites	1024	2048	2048	2048	2048		DRAM
At Speed DRAM Memory								
Wafer Test Parallelism	Number of sites	128	128	128	128	128	128	DRAM
3D Stacked Memory (Wide I/O, HBM, HMC)								
Wafer test	Number of sites	3000	3000	3000	3000	3000	3000	DRAM
Packaged Test	Number of sites	1024	2048	2048	2048	2048		DRAM
Commodity Flash Memory (NAND)								
Wafer test	Number of sites	2048	2048	2048	2048	2048		NAND
Packaged Test	Number of sites	2048	2048	2048	2048	2048		NAND
Stack test [note 6]	Number of sites	4	4	4	4	4	4	NAND
LCD Driver								
Wafer test (Small panel) (5)	Number of sites	6	6	8	8	8	8	LCD
Wafer test (Large panel) (5)	Number of sites	12	12	16	16	16	16	LCD
RF								
Wafer & Packaged test [7]	Number of sites	**32**	**32**	**32**	**32**	64	64	RF
CIS								
Wafer test	Number of sites	64	96	96	128	256	512	CIS
MEMS - Inertial Sensor (Consumer)								
Wafer test	Number of sites	64	64	64	128	512	1024	MEMS
Final test	Number of sites	98	128	256	256	512	1024	MEMS
MEMS - Inertial Sensor (Automotive & Industrial)								
Wafer test	Number of sites	4	4	8	8	16	32	MEMS
Final test	Number of sites	8	8	8	8	16	32	MEMS
MEMS - Microphone								
Wafer test	Number of sites	16	16	16	32	64	128	MEMS
Final test	Number of sites	49	144	144	144	256	512	MEMS

Manufacturable solutions exist, and are being optimized

Manufacturable solutions are known

Interim solutions are known

Manufacturable solutions are NOT known

Notes for Table 1:

1. Assumes I/O count of 250 for MPU and 1000 for ASIC
2. Assumes I/O count of 300
3. Assumes I/O count of 100
4. Wafer test uses Reduced Pin Interface
5. Assumes define Small panel as hand-held display application with one LCD device per each set and Large panel as TV display application with multiplex LCD devices per set
6. Engineering Testing
7. Maximum according to # active RF ports/device

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 as shown in Figure 6. There are more efficient ways to reduce overall cost of test than going to the next setup with more sites in certain cases. Especially for high-mix, low-volume applications, there are many tester utilization challenges. In these setups, frequently, lower degrees of multi-site is preferable because test time improvement of techniques to improve utilization have a higher impact on the overall cost of test.

Touch-Down Efficiency (TDE) is defined as the number of wafer touch-downs required to test all devices on a wafer, relative to the theoretical minimum. TDE is influenced for the most part by the die size (and therefore the number of die per wafer) and the pattern used to probe. For example, if a device is tested 10 sites at a time, and there are 1,000 die per wafer, then ideally a probe card would have to touch down 100 times in order to tester the wafer and be 100% efficient. If, due to the mismatch between the round shape of the wafer and the linear or rectangular pattern of the probe card, the probe card must touch down 110 times to test the 1,000 devices, then the TDE is closer to 90%. This is shown in the figures below.

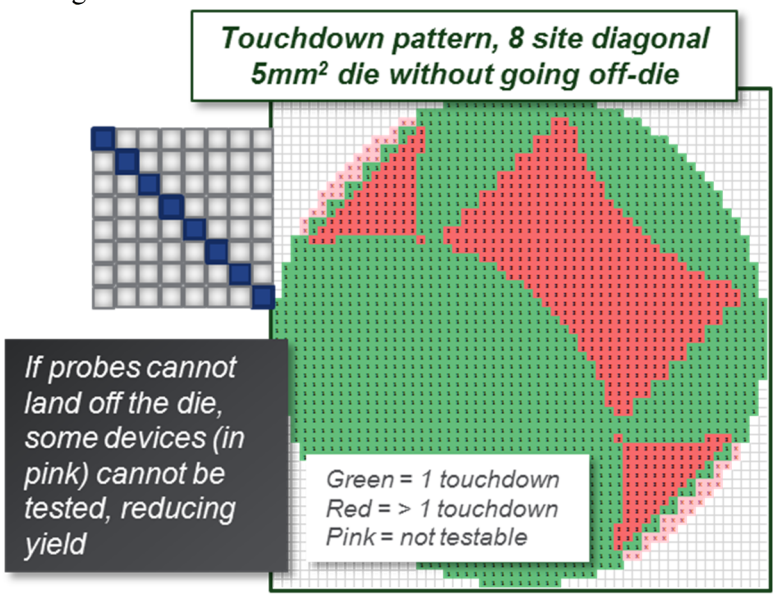


Figure 7: Probe Pattern of 5mm2 die using 8-site probe pattern

As die size of complex device increases, the TDE will continue to degrade as shown below in Figure 8. This degradation of efficiency will negate any advantages of increased site count and will eventually increase Cost of Test as shown in the example below. In this case, there are gaps in the probe pattern to allow for the inclusion of electrical components on the probe card required for the proper operation of the Device Under Test.

TDE inefficiencies will primarily be address by the development of singulated die testing technology. There is significant work underway to allow die to be reassembled in silicon panels that have a rectangular shape as opposed to the round shape of the original silicon wafer. The deployment of this technology will re-start the increase in site count at probe that is currently stalled due to interface costs and TDE limitations.

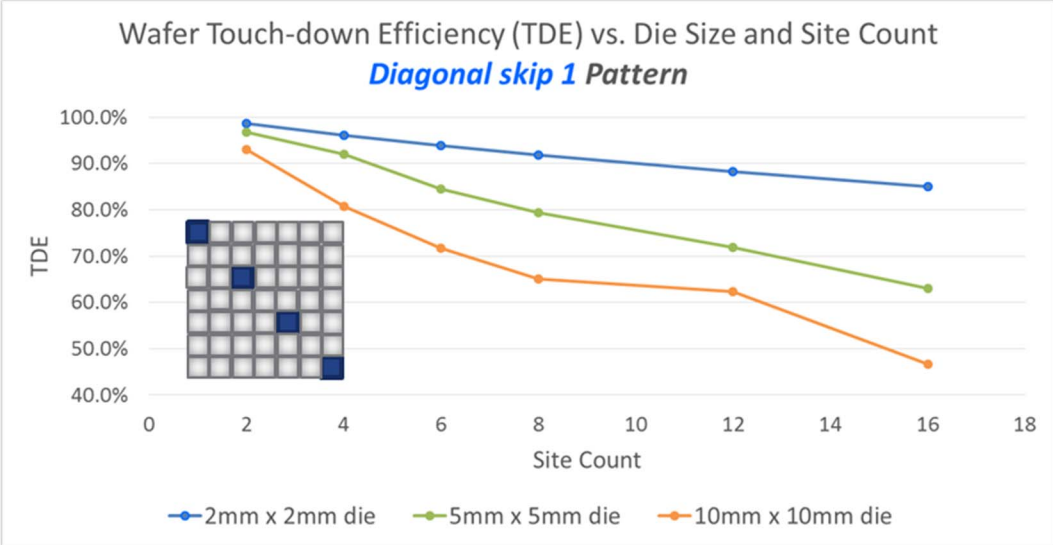


Figure 8: Touch-Down Efficiency as function of die size using a 4-site probe pattern

Summary

Major conclusions are:

- Cost of test has been declining for some time, but the rate of reduction has slowed down.
- Major reason for the slower rate of cost reduction are:
  - Packaging trends that drive more test at the wafer probe insertion where site counts are lower.
  - Increased cost of consumable material, which now dominates tester capital cost in terms of test cell costs.
  - Desire for higher yield, which has a much larger impact on overall device production costs than test costs alone.
  - Desire for higher device quality, especially for automotive applications, which necessitates more test.
- Potential solutions to decrease test costs are:
  - New probing technology which allows test of singulated die.
  - New PCB and Interposer technology to lower the cost and complexity of consumable material.
  - Factory automation.
  - Cost reduction of system-level testing.

Edited by Paul Wesling