Scalable Tester Architecture for Structural Test of Wafers and Packaged ICs

CJ Clark and Mike Ricchetti
Intellitech Corporation, 70 Main Street, Durham, NH 03824
cjclark@intellitech.com and miker@intellitech.com

Abstract

Advances in test technology and growing industry awareness of the benefits of structural test may change the shape of ATE for IC and wafer testers in the near future. Increased device complexity and inclusion of on-chip or in-package FLASH will continue to cause test times for ICs to increase. Structural stuck-at, path-delay and at-speed BIST have become dominant and in some cases replacing functional IC test. Structural test patterns are primarily serial in nature, the opposite of functional test patterns, which are mostly broadside parallel patterns. Memory-behind-pin tester architectures were originally developed to speed delivery of these broadside parallel functional test patterns. Companies committed to implementing DFT for structural IC test may be better served with new tester architectures designed to deliver structural test patterns rather than adapting memory-behind-pin tester architectures.

Traditional IC Design-for-Test

There are many different forms of IC DFT, and it can be implemented in many different ways. For example, when scan was first adopted it was typical to implement just a single internal scan chain, as shown in Figure 1. This was thought to be acceptable for early adopters, as the scan chain length was relatively short.

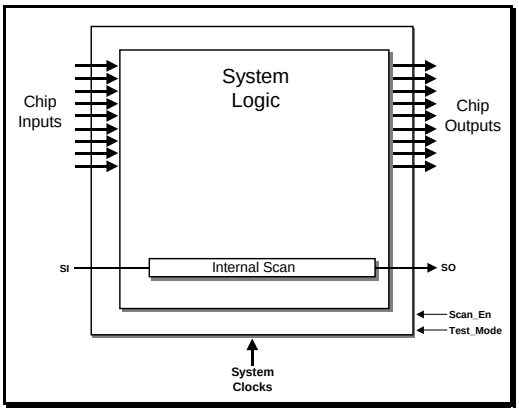

Figure 1

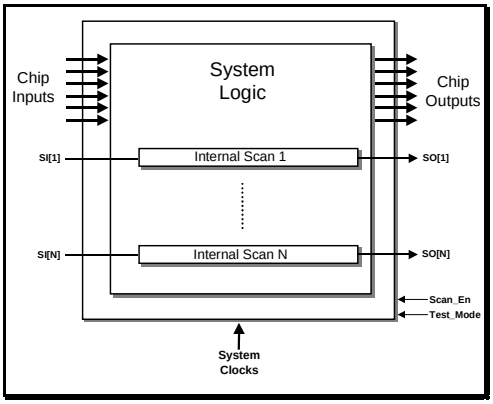

Figure 2

However, early IC testers were designed to apply functional test patterns at speed, so the tester's pin memory was configured for parallel, or broadside, application of test patterns. The scan input and scan output of the scan chain were simply accessed from this parallel pin memory. As design complexity grew, the length of the internal scan chains increased. An affect was that the amount of memory in each tester channel became considerably shallow relative to the length of the scan chain. Since the scan chain length couldn't be deeper than the memory behind the pins of the tester, additional test time was needed to stop and load a continuation of the patterns to memory. This then led to scan-chain partitioning. By implementing a number of short scan chains, which are scanned in parallel, the long single scan-chain was eliminated. This is shown in Figure 2. This then removed the need for adding deeper memory on the tester, and in addition, the scan data could be loaded faster thus reducing the IC/die test time. This was an early form of implementing design-for-test in the IC to match the tester. Scan-chains were partitioned to match the tester's memory depth and number of channels.

Figure 2 shows the scan chains with dedicated input and output pins on the IC, however as the number of parallel scan chains grew, it was unacceptable to add more pins to the package. The solution for this was to multiplex functional pins

with the scan input and scan output pins, so that a larger number of parallel chains could be accessed without additional pads or pins.

Advanced DFT Techniques

As structured DFT techniques such as scan became more widely accepted in the industry, the use of more advanced DFT methods were adopted. Test structures for memory BIST, memory BISR (Built-In Self-Repair), logic BIST and embedded core test began to be included on chip. This is shown in Figure . As BIST and path delay test techniques were developed, it was recognized that for these tests, it was not necessary to apply the test patterns to the IC pins directly and at-speed. Nor was it necessary to "shift" test data at-speed. At-speed testing was accomplished by controlling the 'launch' and 'capture' of the test data by the system clocks, which would be separate from the scan shifting frequency. On-chip control of the launch and capture clocks implemented asynchronous from shifting data into the IC is a key factor in lowering the tester cost. Traditional methods for path delay utilized a fast capture clock at the end of a shift, synchronized with scan_enable going from test mode to functional mode. This approach requires more precise tester pin electronics with very low skew. It becomes more challenging to control the scan_enable off chip in the required time frame as frequencies have increased.

Since structural tests for stuck-ats did not depend on full pin access, it was recognized that chip test could be done with a small number of pins, thus reducing the number that needed to be contacted during wafer and package test. To achieve this, a small amount of additional DFT was added to provide the ability for the scan chain to capture the logic values being driven by the outputs as well as provide input stimulus for the input pins. This technique is known as "I/O wrap". As the difference between V_{OH} and V_{OL} continues to shrink, "I/O Wrap" can provide reasonable assurances that drivers can attain proper output voltages without direct measurement of the pin. Figure shows a chip with an IEEE Std. 1149.1 TAP Controller and Boundary Scan Register [1] added to provide these capabilities. The I/O wrap logic can be incorporated into the Boundary Scan Register and the TAP Controller provides for a standard access port, with reduced contact for IC and die test.

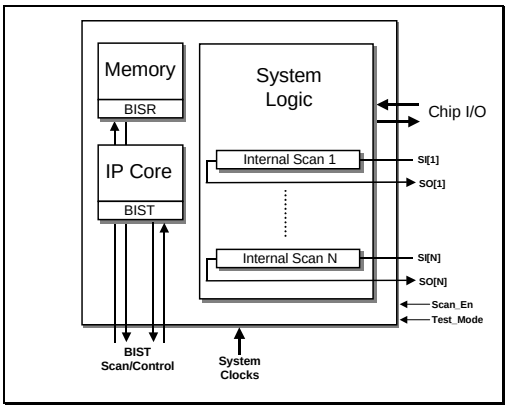

Figure 3

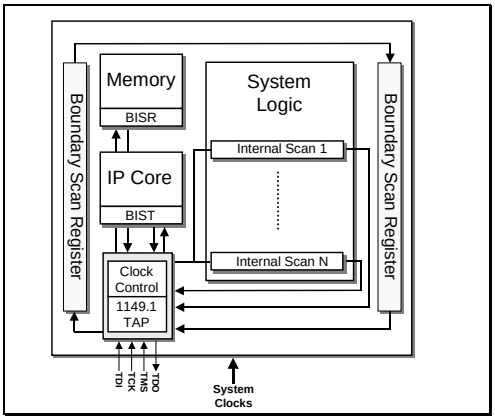

Figure 4

Standards such as IEEE 1149.4 have been architected such that $V_{OH,}$ V_{OL} and other parametrics could be measured without direct access to the IC. However, many of the measurements must be done one at a time, increasing test time. Furthermore, a commercial device has yet to be made that incorporates this new standard. The design is non-trivial. Gigabit serial I/O is becoming more commonplace in IC and SOC designs. The most likely method of testing these high-speed I/O when the I/O are likely faster than the tester channels available is through loopbacks. That is, implementing some type of BIST or BERT on the transmitter and looping back electronically or physically on the load board to an IC input with a reciprocal BIST or BERT receiver.

Low Cost DFT Testers for Structural Test

With wider acceptance of DFT more test capabilities were being built into the design. It was observed that the cost of the tester could be reduced, as it was no longer necessary to use a tester with high-speed pin-electronics for many IC product types. Consequently, low cost DFT testers are now available to take advantage of the structured, scan-based, DFT inside

the IC. The architecture of the low cost DFT tester is a scaled down version of the "big iron" tester. These DFT focused testers couple inexpensive memories with low cost FPGAs to provide around 256 tester pins, capable of applying vector data at roughly 50 MHz. The architecture must mimic the 'big iron' tester, as the low cost DFT tester is typically a platform for vector debug and validation. Once the patterns are fully debugged they are used with the traditional big iron testers in production. The DFT focused tester may only be a desktop tester for one die. Other variations can handle testing a few die in parallel, scaling in cost directly proportionately with the number of die in parallel as a 'big iron' tester would.

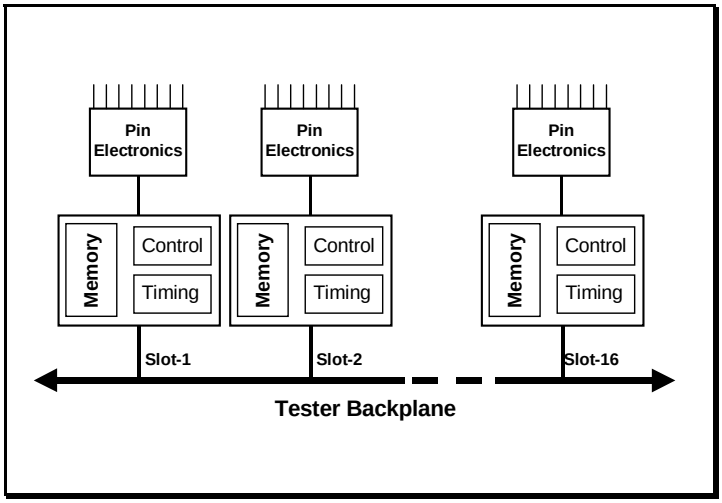

Figure 3

A traditional tester architecture with memory-behind-pin is shown in Figure 3. The architecture is designed around a tester backplane, usually proprietary, where proprietary cards can be plugged in to expand the test capabilities of the tester. A typical card would have pattern memory, a control unit, and timing formatters and generators. This then drives the pin electronics of the tester, which interface to the device or unit under test (DUT or UUT). Reduced cost versions of this architecture with less timing granularity, slower memory, smaller pin voltage ranges and granularity or less channel timing independence are typical of low-cost DFT focused testers. Efforts in the industry to create an 'open' architecture tester are underway, however, this doesn't appear to be a new architecture, just an open backplane with much of the same architecture of proprietary 'big iron' testers such as memory behind pin digital I/O.

Multi-Site Testing with Memory-Behind-Pin

In and effort to further reduce costs, and improve manufacturing throughput, methods for testing multiple UUTs simultaneously are being adopted and have existed for memories for some time. A common method is referred to as multi-site testing. Multi-site testing can be performed on traditional tester architectures, such as shown in Figure 3. For example, each backplane card in Figure 3 interfaces to 8 tester channels. With up to 16 cards this provides for 128 channels. These channels can be partitioned among multiple "sites" in order to test a number of UUTs in parallel.

Although conventional testers can be used for multi-site test, such testers are ultimately limited in the number of units that can be tested in parallel. Such limitations are inherent in their architectures, which depend on individual tester channels and the configuration of the pattern memory associated with the channels. Since the channels and memory are limited in number and capacity, there is a limit to the number of units that can be tested or configured simultaneously. In addition, the expansion method of these memory-behind-pin testers is constrained by their fixed number of PCBs that will fit the backplane and the bandwidth of the busses. Performance bottlenecks will occur as it is expanded and more resources are added. The bandwidth and number of slots of the tester backplane thus determines the expansion limit with this approach. Memory will have to be large enough on the pin electronics card to avoid re-loading the card with continuation patterns.

Consequently, these architectures are not readily scalable and their costs rise prohibitively as more parallelism is necessitated and additional resources are required. These traditional memory-behind-pin testers are not purposely designed to accommodate parallel testing. Rather, they have limits to the expansion of their resources, and so the number of UUTs to be tested when using multi-site methods is based on what is practicable, not what is optimal. These factors are determined during the design of the memory-behind-pin architecture tester. What increases the tester costs is the fact that

as more pins are designed for the tester, the more data bandwidth that is needed from the memory to the pins. With 64 pins at 100 MHz the data delivery infrastructure, the physical infrastructure, and power infrastructure must handle 6.4Gbits/sec of data. With 28 pins, all three need "12.8Gbits/sec worth" of infrastructure.

Consider the power infrastructure for multi-site testing. In multi-site testing, all ICs are going through the exact same sequence of operations and therefore requiring approximately the same amount of current for good ICs/die. ICs with faults require varying amounts of current depending on the fault. One particular attribute to consider is the "in-rush" current needed at the instant switching inside the IC occurs. A power-supply must be capable of delivering this 'instantaneous' in-rush current. As more die/ICs are tested simultaneously, the more instantaneous current needed, the more sophisticated/costly the power delivery infrastructure and the higher total tester cost.

Parallel Test Architecture

A novel parallel tester which is based on a patent-pending architecture called the Parallel Test Architecture (PTA) is described in [2], [3]. A unique aspect of the tester and parallel test bus is that the typical comparison between received data, expected data and mask is done not at the source of the test stimulus, but in small addressable controllers that are more 'local' to the UUT and in some cases embedded in the UUT itself. The addressable controllers act as tester channel 'duplicators', removing the need for true tester channels when parallel test with a small number of pins is desired. This offers a lower cost approach than the 'one tester size fits all', that is using a tester where all pins are flexible for testing a single UUT with many pins, or many smaller ICs in parallel. The authors prefer to show the patent-pending parallel test bus as implemented over the IEEE standard 1149.1 bus, however any parallel or serial bus can be used. A parallel architecture using traditional internal scan with Scan Enable, Scan In, Scan Out and a System clock is illustrated in the last figure of the paper. IEEE 1149.1 has never been seriously considered for IC test in the past, but, as more on-chip test resources such as BIST, deterministic BIST and test compression are provided, the less demanding the bandwidth needs are between the IC and the tester. Path-delay and at-speed BIST have shown that full speed access to IC all pins is not needed in order to test the IC at speed. Looking towards the future, it may be more common to test an entire digital IC with merely a high-speed clock and one or two scan-chains, perhaps under 1149.1 control. The reader should also note that, while IEEE 1149.1 at the PCB level may have been only 10 MHz in the past, today's test bus controllers are running 64mb/sec and more. There is no upper limit on the scan frequency as specified in the standard.

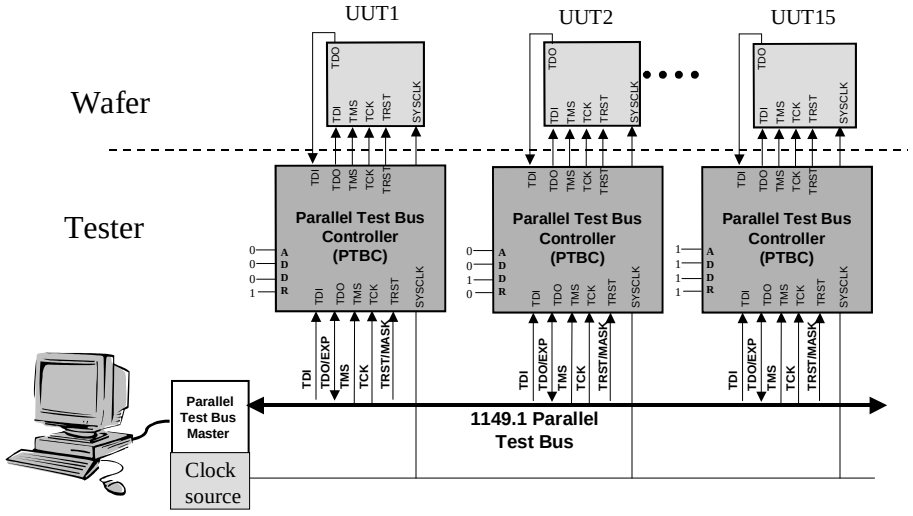

Figure 6. Basic Parallel Test Bus as implemented over IEEE 1149.1

Figure 6 shows a basic block diagram of the tester in its simplest form. The PC, high-speed clock source and Parallel Test Bus Master are typically implemented on a PXI chassis. Each PTBC Parallel Test Bus controller receives 'broadcast' instructions or targeted instructions for a PTBC on a particular address over a modified multi-drop bus. The multi-drop bus differs from previously described multi-drop busses with commercial parts such as the TI ASP and the National ScanBridge in that each PTBC receives expected and mask data and compares the received scan-out data from the UUT locally. Unlike

standard IEEE 1149.1, the TDO is bidirectional so the expected data can be sent to it during SHIFT_DR. Optionally, if the technology does not allow TDO to be bidirectional (LVDS for instance) then a separate input pin would be used to receive expected data. What was noticed is that during SHIFT_DR, the TRST and TDO in a multi-drop environment are not used, so their function can be multiplexed. The Parallel Test Bus master sends to all PTBCs simultaneously scan in data on TDI, TCK, TMS and expected data on TDO and MASK data on TRST. When a miscompare occurs in the PTBC, the internal failure bit is set and optionally an external open collector pass/fail pin (not shown in Figure 6) can indicate a failure occurred. At the end of the test each PTBC can be addressed to check its pass/fail status and the contents of a small bit failure RAM could be sent back to the tester. Time can be saved by ganging the pass/fail lines together for all PTBCs so the tester does not need to poll if all UUTs pass. Other features include performing a reset on the UUT after first failure or signaling power to shut down the individual failing UUT. With the addressability, individual access to each UUT is preserved, so, for instance, a unique serial number could be loaded into each UUT by addressing the appropriate PTBC one at a time. In addition, on-chip repair mechanisms that are not automated by BISR could also be enabled for UUTs that are determined to have failed.

Of course, as more PTBCs are added bandwidth to the UUT will be reduced due to the additional loads. To overcome this, the PTA includes a Parallel Test Bus Synchronizer as shown in Figure 7.

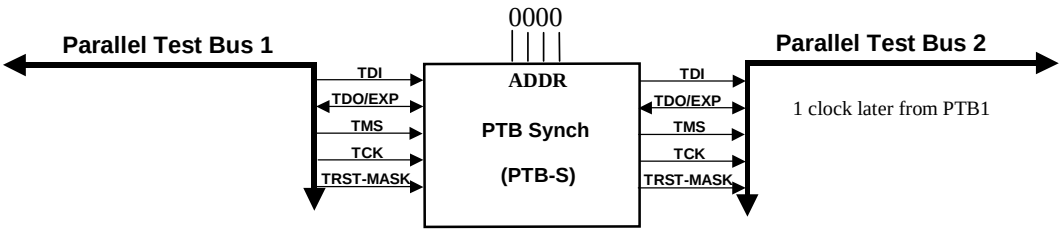

Figure 7. Parallel Test Bus Syncrhonizer

The PTB-S connects Parallel Test Bus segments together so there is no loss of bandwidth or shift rate for the test data. A PTB-S is addressable as well so during individual access to the UUT, the scan out data can be brought back to the PC and Parallel Test Bus master. The PTB-S can inject one or more clock cycles between the Parallel Test Busses. In practice, one clock is used and this can be enough to change the characteristics of the in-rush current needed by the UUTs. The total test time for the 'last' UUT is only increased by the number of UUTs that are being tested multiplied by the number of clock cycles injected by the PTB-S. In practice, one UUT is used per Parallel Test Bus segment and the PTBC and PTB-S are integrated so they share a common address. So, while the UUTs are tested 'in parallel', each is really going through a different clock cycle of the test, which intentionally reduces the instantaneous current needed for the UUTs. Figure 8, shows a typical implementation.

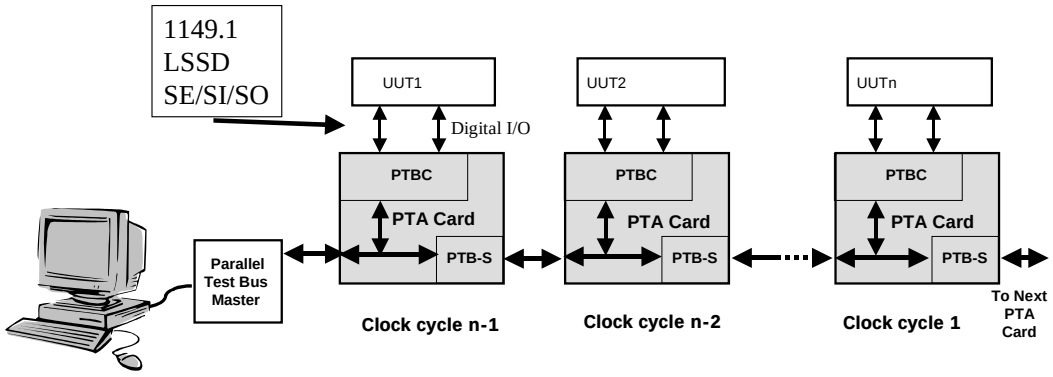

Figure 8. Scalable Tester

The PTBC and PTB-S have been combined on a stand-alone pin electronics card. Each card can be connected to the next with high-speed cable. Variable I/O voltage control to the UUT from .8V to 5V is included and controlled by the PTBC. Slow speed digital I/O pins are also provided to hold test_enable and similar near static pins. Power for the pin electronics is daisy chained as well, but can be segmented for very large collections of cards. The scan protocol from the PTBC to the UUT is also variable, supporting 1149.1, LSSD and traditional muxed scan. Since each PTA card is going through test at a slightly different clock cycle the opportunity exists for taking analog measurements across all of the UUTs from a single instrument without impacting the test time over testing a single die with a single PMU. Each PTA card can signal to a single PMU and switch fabric that the PTA is at the point in the test that requires an analog measurement. As long as the analog measurement can be done in just a few shift clock cycles, then it can be done without impacting the test time. Also, it should be noted that the data bandwidth needed by the PC and the PXI bus is fixed, it does not need to increase as more PTA pin electronics cards are added.

The technique described above contrasts typical IC test today. In order to keep test times short, many parallel internal scan-chains are used. This partitions the problem so the parallelism is in the IC itself. In the PTA architecture, the parallelism is in the tester across multiple die. While it may take longer to shift through a single scan chain of 100,000 bits than it would for 10 parallel scan-chains of 10,000 bits, the Parallel Test Bus allows 10 die to be tested simultaneously in the same time. In many cases the scan-chain 'shifting' is not the limiting factor for test time of ICs with embedded RAMs and embedded FLASH. Advances in on-chip compression/decompression enable many 'virtual' internal scan-chains. Looking to the future, it may not be necessary to use physical parallel internal scan-chains that require more tester pins and contact to the die.

If the power pins and system clocks are considered the same between the two approaches, the parallel tester would require more pins, fifty versus twenty-one for the traditional approach. If the die/IC also has 1149.1 boundary-scan, as many do today, then the traditional approach needs twenty-five pins to exercise the TAP too. Since the PTBC is a test channel 'duplicator' the cost of the pins would be far less for equivalent traditional 50 MHz memory-behind-pin channels. Where appropriate, it is possible to run multiple concurrent Parallel Test Busses for certain ICs with multiple internal scan-chains in parallel. However, it is best to keep the internal scan-chains to one or two, otherwise the complexity of the tester, the pin contact and eventually the data bandwidth in the PC will all need to increase.

The industry mantra of 'shortest test time' will lower product test costs at the exclusion of all other factors, should be re-evaluated for certain product classes. For some ICs, longer test times on lower cost test platforms with many IC tested in parallel may prove to provide a lower overall product cost.

Issues

Much of the proposed method will draw concerns for the experienced IC test engineer. Obviously only certain IC types could benefit from this test approach. There are still technical hurdles to overcome in order to make this approach a preferred test method. As more die are tested in parallel then more contact is needed from the wafer probe. Even with an ultra low pin count tester (ULPCT) of ten pins or so, the number of power and grounds needed by all of the die is beyond traditional wafer probe techniques. However, recent advances have been made in probe technology. As of the writing of this paper, a few companies are claiming full wafer contact probe cards. Looking towards the future, it would appear that this obstacle could be overcome. Since this technique is using ULPCT, it may be possible to make the test pads larger to improve contact over the full wafer.

As more die are tested in parallel, more total current is needed. Even reducing the in-rush current would not help if the die under test were a microprocessor requiring 50 watts or more during testing. Increases in cost of handling the power and the cooling requirements for the wafer may offset the advantages of the lower cost tester. Perhaps it is possible that those increases in cost could be offset by a lower handler cost. With all die tested in parallel or nearly all die, the indexing speed of the prober, and expectedly the cost, could be reduced. There may be other ways of testing multiple die in parallel. Emerging 'cartridge' technologies for handling multiple wafers may offer a way to test just a few die in parallel on 10 or 12 wafers simultaneously. This may alleviate exotic cooling requirements for high power die.

Even if only one or two die could be tested on a wafer, there are no real physical limits on how far apart a PTBC can be from another PTBC. Current implementations of the PTA card allow it to be cabled forty feet (13 meters) apart from each other with no loss of shift speed. This could enable a scenario of a single computer with parallel test bus master, the

'tester', with distributed pin electronics cards over multiple handlers. This would alleviate the need for full wafer contact and the number of 'wires' needed to interface from the tester to the probe card at a cost of having multiple handlers.

Problems also exist in getting the proper DFT on-chip for even multiple parallel scan-chains. Commercial ICs exist with multiple BIST engines, at-speed test and I/O wrap enabling testing with a ULPCT type tester and single high-speed clock; however, they appear to be the exception, not the rule. Companies appear to want to use DFT structured testers due to their low cost, however, the advanced DFT is not being implemented. Many of these same companies are producing designs with twelve independent scan-chains each needing an independent shift clock, slightly out of phase, fast capture and needing full pin access. DFT insertion tools need to be easier to use, especially on multiple clock domain ICs. I/O wrap insertion also needs to be more automated to make it mainstream. Designers personal review process by managers may need to have less focus on time-to-tape out deadlines and include elements such as measuring the success of time-to-market, yield and profitability of the design.

Other possibilities

There may be some advantages to implementing the PTBC into the IC itself or in the probe card. The PTBC is only several hundred gates and can be combined with an 1149.1 TAP controller. The same is true for the PTB-S. The PTBC of Figure 6 could be moved inside the IC. This could provide added parallel test capabilities for current ATE that have limited tester pins. If currently owned depreciated capital equipment could be enabled with additional multi-site capability without adding more tester channels, it may be the lowest cost approach to using the Parallel Test Architecture. Implementing the entire PTA all on chip and on-wafer has the traditional 'catch-22, that a possible defect could prevent the entire bus from working. One approach to overcome this would be to implement the PTA architecture portion of the IC in a larger technology such as .35um that would be less susceptible to defects compared to 90nm. Another approach would be to add 'bypass' switches for each die under test in the probe card or possibly in the wafer etch as shown in Figure 9. In this scenario, the IC is designed with the PTBC *and* PTB-S. The pins needed for the address, PTB bus in and PTB bus out can all be muxed with functional pins except for the SE (Scan_Enable) and the System Clock.

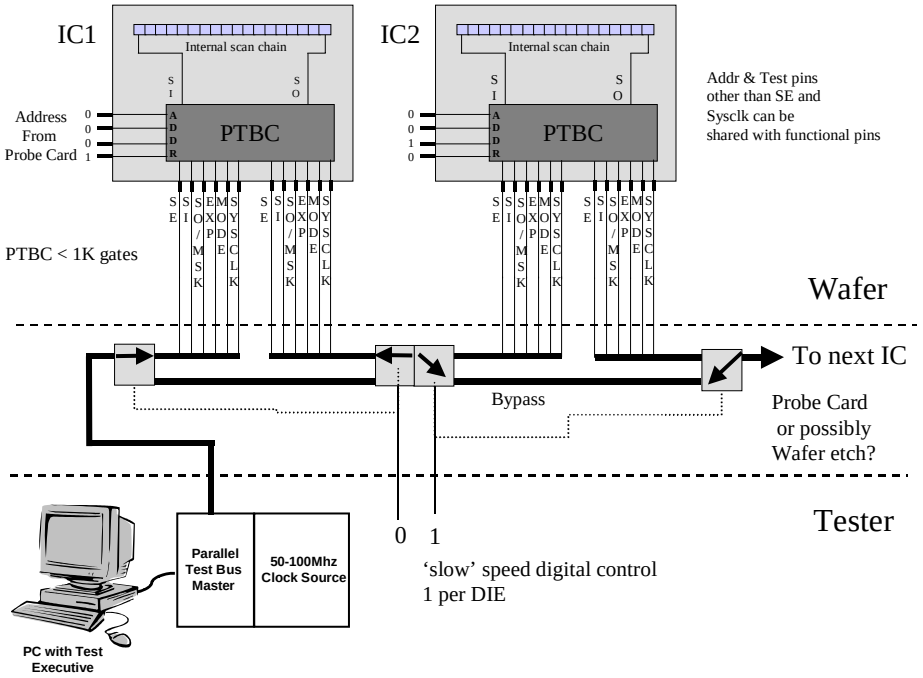

Figure 9 On-chip/On-wafer test architecture

Since a bad IC could prevent the PTA from working, the two 'switches' per die enable the die to be taken off the Parallel Test bus. At the start of wafer test, the Parallel Test Bus master with access to one slow speed digital control signal per die could quickly check the integrity of the PTB, enabling and disabling die as needed. This approach provides the benefits of

the PTA, but with a reduced number of wires needed from the ‘tester’ to the probe card. Since the switches are small, it may be possible to implement them in the wafer etch. Doing this would reduce the number of contacts to the wafer as well.

Conclusions

We have described an implementation of the novel Parallel Test Architecture for testing multiple ICs or die in parallel. An example of a scan-based tester using the architecture was shown. This parallel tester reduces costs over that of traditional memory-behind-pin, and tester-per-channel architectures. The PTA tester is designed to be readily expandable, in order to support parallel testing, and is targeted toward structured DFT-based methods and designs to take full advantage of reduced cost of test.

References

[1] IEEE Std 1149.1b-1994, “IEEE Standard Test Access Port and Boundary-Scan Architecture”, *Institute of Electrical and Electronic Engineers, Inc*., New York, NY, USA.

[2] Ricchetti, M. Clark, CJ, "Method and Apparatus for Optimized Parallel Testing and Access of Electronic Circuits", *US Patent Application 2003009715*, US Patent and Trademark Office, Washington, D.C., July 5, 2001.

[3] Clark, CJ, Ricchetti, M., "Method and Apparatus for Optimized Parallel Testing and Access of Electronic Circuits", *PCT Patent Application WO03005050*, World Intellectual Property Organization, Geneva, Switzerland, July 5, 2001.

[4] Clark, CJ, Ricchetti, Mike, “Infrastructure IP for Configuration and Test of Boards and Systems”, IEEE Design & Test of Computers, vol. 20, no. 3, May-June 2003, pp. 78-87.