Power Aware Embedded Test

Xijiang Lin, Elham Moghaddam, Nilanjan Mukherjee
Benoit Nadeau-Dostie, Janusz Rajski
Mentor Graphics Corporation
Wilsonville, OR 97070, USA

Jerzy Tyszer
Poznań University of Technology
60-965 Poznań, Poland

***Abstract*—In this paper we examine several embedded low power test schemes that we have proposed over the last few years. These solutions are aimed at reducing the switching activity during all scan-based test operations, particularly including those developed for BIST or deployed to perform on-chip test data compression.**

I. Introduction

Embedded test resources are being viewed increasingly as essential to reduce test cost. In particular, scan-based DFT schemes have gained broad acceptance as reliable test solutions. When it comes to energy consumption, however, scan-based test operations can dissipate much more power than low-power circuits were designed to function under. With overstressing devices beyond the mission mode, reductions in the operating power of ICs in a test mode have been recurring themes for years. A full-toggle scan pattern may draw several times the typical functional mode power, and this trend continues to rise, particularly over the mission mode's peak power. This so-called power-induced over-test may result in thermal issues, voltage noise, power droop, or excessive peak power over multiple cycles, which, in turn, cause a yield loss due to instant device damage, severe decrease in chip reliability, shorter product lifetime, or a device malfunction because of, for example, timing failures following a significant circuit delay increase.

Abnormal switching activity may also cause fully functional chips to fail during testing because of phenomena such as IR-drop, crosstalk, or di/dt problem. IR-drop refers to the power-rail voltage decrease due to higher current and the resistance of the devices between the rail and a circuit's node. Its large value increases path delays, and this degradation only worsens with elevated test frequencies or low supply voltages. Therefore, a device passing a slow speed scan test at nominal voltage may fail at-speed scan test where either the capture frequency is raised, the supply voltage is lowered, or both.

The underlying mechanism in crosstalk is related to capacitive coupling between IC's neighboring nodes. If the aggressor and victim nodes switch in the same direction, a delay on the victim node decreases, whereas it tends to increase, if they switch in opposite directions. Depending on the extent of coupling and a switching pattern during at-speed test, a failure can only be observed during testing, not in the functional mode.

The di/dt problem occurs in circuits using at-speed scan tests and is due to the sudden current changes within a few nanoseconds when applying the capture clock burst (typically a pair of launch and capture cycles) after a long pause where all clocks are inactive. Hence, the power distribution network may not be able to maintain the supply voltage in desired limits (the supply lines become unstable and reach levels below the ones encountered in a mission mode). The circuit might appear to be faster or slower than it actually is. It seems to be faster, if the clock stretching effect is dominant, and slower, if the combinational slowdown is forceful due to the supply variations.

The problem of excessive power dissipation during scan testing is typically split into two separate domains dealing with power reduction in shift and capture modes, respectively. Accordingly, numerous techniques for test power reduction and management have been proposed, focusing on test scheduling and reordering, partitioning and modifications of scan chains, scan cell reordering, as well as transition blocking and various forms of clock gating. Similar schemes have also been proposed for built-in self-test (BIST), including low-transition pseudorandom test pattern generators and gated scan cells.

Other solutions tailor patterns to the requirements of tests with the lowered switching activity, for instance, by reassigning certain non-random values to unspecified positions in test cubes causing power violations. It minimizes the number of transitions during scan loading, as done in power-aware ATPG. Deterministic test vectors can yield patterns leading to a reduced switching activity because of their usually low fill rates. Instead of random fill, don't care bits may assume (in the process of X-filling) values minimizing the amount of transitions during scan-in shifting. Other forms of X-filling decrease capture power by assigning particular values to unspecified bits so that the number of transitions at the outputs of scan cells in the capture mode is minimized.

On-chip test compression is facing similar test power problems. As on-chip decompressors expand and subsequently load test cubes into scan chains, don't care bits are typically filled with random values, and therefore the amount of flip-flop toggling during test may result in a power droop condition. The bulk of test power consumption can also be attributed to unloading test responses. Several low power test data encoding schemes were presented. Some of them rest on static LFSR reseeding techniques with certain extensions that allow reducing the scan-in transition probability. Other methods reduce test power in dynamic reseeding by using available encoding capacity to limit transitions in scan chains. These techniques freeze a decompressor in certain states by providing control data through external channels. It allows loading scan chains with decompressed patterns having low transition counts, and thus reduced scan-in power dissipation.

1081-7735/11 $26.00 © 2011 IEEE
DOI 10.1109/ATS.2011.49

A comprehensive coverage of all aspects related to power-aware testing can be found, for example, in [6]. This paper examines schemes that we have proposed over the last few years to conduct various low power embedded tests.

II. Controlling Toggle Rate in Shift

Loading scan chains dissipates power dependent directly on the number of transitions that occur in the scan chains and other parts of the CUT. One can estimate the resultant switching activity by counting the number of invoked transitions in scan cells and taking into account their relative positions for all scan chains and all test patterns. Typically, a degree of switching is determined by a random fill, which is a side effect of feeding scan chains with various types of patterns, as done in the EDT-based decompression where a compressed test vector goes through the decompressor.

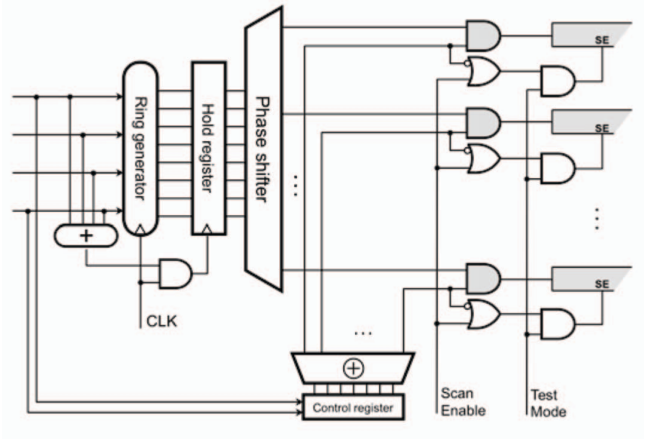

Figure 1. Low power decompressor

A. Low power decompressor

Since deterministic tests have typically only a small fraction of bits specified with their sites confined to a few scan chains, one can feed such scan chains directly through a low power decompressor that replaces all don't cares with a low toggling fill while delivering a constant 0 or/and 1 to all remaining chains. It significantly reduces the number of transitions during scan-in shifting without compromising test coverage. Fig. 1 shows a solution [2], [3], [4] where input channels first deliver information to a control block, and then provide compressed test patterns to a low power decompressor. The control block comprises XOR logic driven by compressed data stored in a register, and it outputs gating signals to gray AND gates such that either the decompressor or a constant value of zero feeds scan chains on a per pattern basis. The control circuitry selects gating signals for scan chains with specified bits such that the AND gate has 1 on its input. It allows the decompressor to drive these scan chains directly. The selection process continues until power limits are reached. As for the remaining scan chains, one can determine gating signals that were not encoded yet. Approximately, the decompressor can still feed 50% of such scan chains, while the others receive a constant 0.

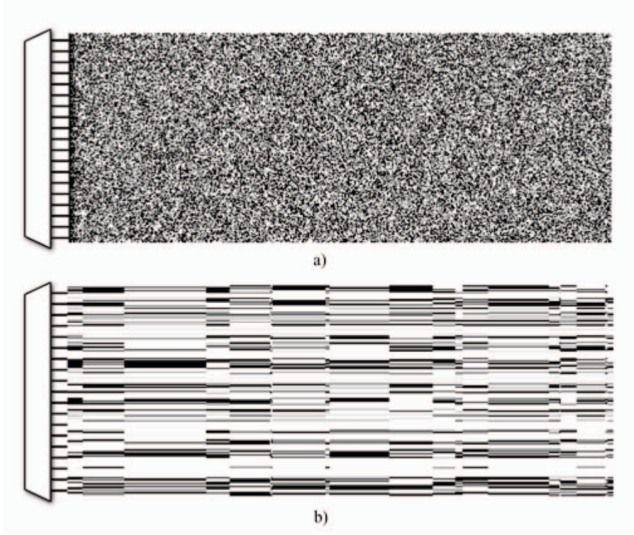

Figure 2. Random (a) and low power (b) fill in scan chains

In addition to shutting down a number of scan chains, the same scheme employs a power aware decompressor, as shown in Fig. 1. It provides data to the scan chains for a number of shift cycles through a *hold register* placed between the ring generator and the phase shifter. The hold register sustains, for a number of cycles, a desired state of the ring generator, while the generator itself keeps advancing. It visibly reduces a transition count since majority of chains can be loaded as long as there are no conflicts between next specified bits and data used in the previous cycles. The decompressor input channels facilitate also the operation of the hold register. Every shift cycle, a new control bit indicates whether the hold register should be reloaded with the current content of the ring generator. This information is retrieved from seed variables as their parity bit for a given cycle. If the parity is odd, then the hold register is reloaded before new seed variables enter the ring generator. Otherwise, the content of the register remains unchanged.

The corresponding compression procedure partitions a given test cube into several blocks comprising a certain number of consecutive time frames in such a way that there are no transitions inside the blocks. This feature allows repeating a given decompressor state many times in succession by using the hold register storing a state that the ring generator entered at the beginning of a block. The ability to encode specified bits within boundaries of the block determines its size. The encoding process begins with a block and the corresponding state of the ring generator, which should be applied first, and it gradually moves towards the other end of the scan chains. As long as the solver can compress a test cube, the algorithm repeatedly increases the size of the block, adds new equations (if needed), and invokes the solver again. If there is no solution anymore, a new block is formed, and the procedure continues to arrive with the smallest number of blocks that cover the entire test cube. Since the same test data drives the decompressor and the hold register, the algorithm works with equations in the seed variables representing both the specified bits and control data.

Consider an industrial design with a random switching profile shown in Fig. 2a. When both the low power control-

ler and the power-aware decompressor are employed to reduce the amount of transitions in scan chains, the resultant switching activity is significantly lower than that of Fig 2a, as clearly visualized in Fig. 2b, where white (0s) and black (1s) stripes make up the low power-toggling pattern. It is worth noting that similar results can also be obtained by freezing decompressors in so-called self-loop states [5]. This purely software approach does not require any modifications to the already complex core and the corresponding test logic.

B. Selective clock gating

One can reduce test data volume and test power by independently gating clocks associated with different scan chains [8], and thus allowing each scan chain to be enabled autonomously. XOR logic and its control register can gate the clock signals (in a manner similar to that of Fig. 1) prior to driving the clock inputs to scan cells. For each test pattern, only a predetermined fraction of scan chains becomes active. These chains are loaded with new test data, while a test response yielded by the previous test pattern is shifted out. Since only selected scan chains are operable during test application, scan shift transitions together with the resulting switching activity in the combinational logic are both reduced. Clearly, this technique is paired with a compression method capable of generating and merging test cubes in a custom way. In particular, it freezes scan chains that can be loaded only once for several consecutive test patterns as their specified bits are shared by these vectors. The remaining scan chains, which host patterns with conflicting values on the corresponding positions, are reloaded individually for each test pattern. The fraction of scan chains, which can hold their contents between loads, determines the effectiveness of the scheme. It is worth noting that by encoding a limited number of specified bits, this approach not only reduces scan shift power consumption, but it also achieves a considerable reduction of test data volume. Furthermore, similarly to segmented scan designs that achieve higher delay fault coverage when using multiple capture cycles after loading a test pattern, the proposed scan architecture can deliver tests of a comparable quality.

III. Controlling Toggle Rate in Capture

A. Dependence of capture on load

Consider now scan operations during capture of test responses and shifting them out. Fig. 1 shows the solution where the same XOR network as before acts as a capture control circuitry gating the scan enable signals. Once all control variables are known, the status of those scan chains that were not the subject of encoding is determined so that approximately 50% of them may remain in the shift mode, that is, they do not record test results. As a result, the power dissipation during capture cycles is reduced. Furthermore, many of the scan chains that do not capture test responses maintain their low toggle test patterns loaded earlier. Hence, when the entire circuit is placed in the scan mode again, that preserved scan content facilitates suppression of flip-flop toggling during shift out operations.

It is worth noting that selecting different constant values loaded into the scan chains fed by the constant test stimulus may have a different impact on the switching activity during capture. A preferred fill of [13] uses signal probability to determine the filling value during test generation to reduce the capture switching activity. This strategy can be adopted here to select a constant stimulus in such a way that the constant 0 is chosen anytime the number of scan cells with the preferred value 0 is greater than or equal to the number of scan cells with the preferred value 1.

B. Scan segmentation and clock gating

A solution with no impact on the design flow is to leverage the embedded clock gaters available in many designs to reduce the power consumption. The clock gaters are simple devices inserted between a clock line and the corresponding state elements. They prevent clock pulses from reaching selected flip-flops. As a result, one can reduce the capture power by forcing those memory elements to hold their states. Fig. 3 shows a typical clock gater. During a shift, the test enable (TE) pin is asserted. As a result, the clock gater becomes transparent and allows CLK to pass through. During a capture, TE is de-asserted, and the functional enable (FE) pin takes over. To reduce the capture power, ATPG can de-assert FE if the state elements controlled by the clock gater are not used to activate, propagate, and/or observe faults.

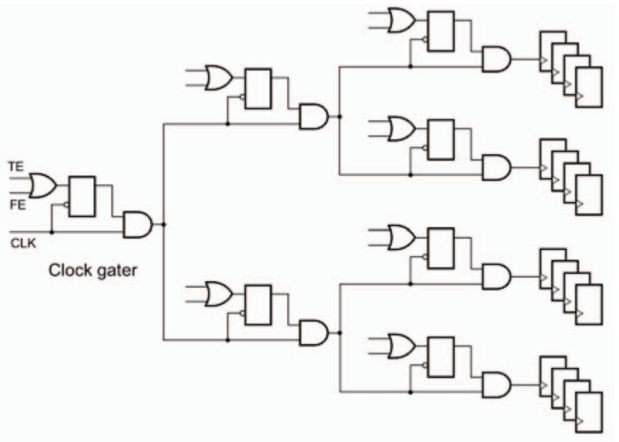

Figure 3. Hierarchical clock gating

In a mission mode, functional enable logic fed by scan cells controls the clock gaters. To support scan shift operations, clock gaters are forced on when a test mode or scan enable are asserted. ATPG can justify the functional enable logic to its off state, thus preventing many downstream flip-flops from toggling. In EDT environment, it can be accomplished by using a relatively small amount of test data provided by scan chains in such a way that, during capture, scan values control the clock gaters. Clearly, the primary objective is to meet the power constraints by disabling flip-flops with as few additional specified bits as possible to minimize impact on compression. Furthermore, many designs may feature a hierarchy of clock gaters, to achieve both fine and coarse levels of control granularity, as shown in Fig. 3. If no flip-flops in a specific area of the design observe faults, one

may turn off a higher-level clock gater by using fewer specified bits. The method for enhancing ATPG to reduce the capture power by controlling clock gaters is as follows [2].

First, a structural analysis identifies all clock gaters for each available clock. Then, ATPG targets the identified clock gaters, prior to pattern generation, to obtain clock control cubes (CCC) disabling the targeted clock gater. Due to a possible overlap in the control logic of clock gaters, a CCC may additionally enable or disable other clock gaters. Given the CCCs for a clock, the cubes are sorted in decreasing order with respect to the number of state elements whose clocks each cube turns off. After dynamic compaction, compatible CCCs are merged in the aforementioned order to maximize gating off the flip-flops that do not need to be clocked, while the number of additional specified bits that must be provided by scan is minimized. Note that during dynamic compaction, it is necessary to check routinely that there remains enough encoding capacity to merge CCCs while meeting the power constraints.

C. Recirculation

Test vectors that mimic, particularly from a toggling perspective, a mission mode can also mitigate test power dissipated during capture cycles. A method to generate pseudo functional test patterns with switching activity close to that of functional patterns in EDT environment is presented in [9]. It uses a controller and 2-input multiplexers placed in the front of scan chains, as shown in Fig. 4. The multiplexers route test data from two sources: (1) the basic EDT decompressor, (2) serial outputs of the corresponding scan chains which essentially recirculate the captured response from the previous vector to be used as a background fill for the unspecified scan cells of the next test vector. The actual source is determined by applying select signals to the multiplexers. An XOR-logic-based controller with a biasing circuitry provides select signals based on compressed data received through EDT input channels that it shares with the decompressor. The controller's configuration is reloaded once per pattern.

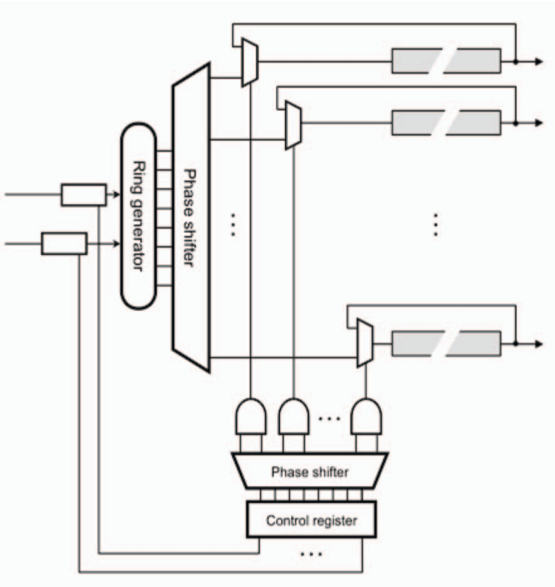

Figure 4. Mimicking functional toggling in test mode

To generate pseudo functional test patterns, the compression method of [9] first initializes a circuit with a pseudo functional pattern, and then applies a test cube. Since it has typically a very low fill rate, the resultant switching activity will be close to the switching activity of initial pseudo functional pattern. Consequently, one can consider the resulting test vector as a pseudo functional pattern and use its functional response in launch-of-capture (LOC)-based at-speed scan testing as pseudo functional background for the next test cube.

D. Blocking scan cell outputs

The capture power can be further reduced by deploying gates blocking scan cell outputs. These gates can be partitioned into multiple groups, where each group is controlled by a dedicated signal initialized through scan loading. During test generation, ATPG dynamically makes the majority of blocking gates hold their values, while the remaining ones work in the transparent mode. This approach is suitable for compression environments. The controller that generates dedicated control signals and a test generation procedure are presented in [7].

IV. Programmable Capture Clocking

Techniques presented in two subsequent sections were primarily proposed and implemented for various BIST schemes. In particular, the application of at-speed structural tests has to consider numerous design aspects that may impact the circuit behavior. Test-session related activities may influence the average current, the supply voltage, and the local temperature. These factors, in turn, affect the circuit timing. The requirement to apply two consecutive high-speed clock pulses (a functional frequency rate) during the capture phase and after shifting test patterns into scan chains at a much slower speed puts significant strains on the power supply. It was reported [12] that the clock period of a design varies by more than 15% due to a voltage droop (the phenomenon also known as a clock stretching) resulting from sudden changes in circuit activity (di/dt effect – see Section I) upon application of two high-speed clock cycles. It is therefore necessary that during at-speed clocking sufficient care be taken to control supply variations.

Numerous techniques have been proposed to address the clock stretching effect when applying LOC patterns during at-speed tests. It was suggested [12] that the number of capture pulses should be increased to allow enough time to stabilize the power supply. In fact, the di/dt effects can be reduced by an order of magnitude when three launch cycles are used instead of one before the capture cycle [1]. However, these techniques tend to increase both the ATPG complexity and the number of at-speed LOC test patterns. A *BurstMode* clocking methodology [11] was therefore proposed to address the issues related to power supply variations. It deploys a programmable capture clocking scheme to control the frequency as well as the duration of clock pulses.

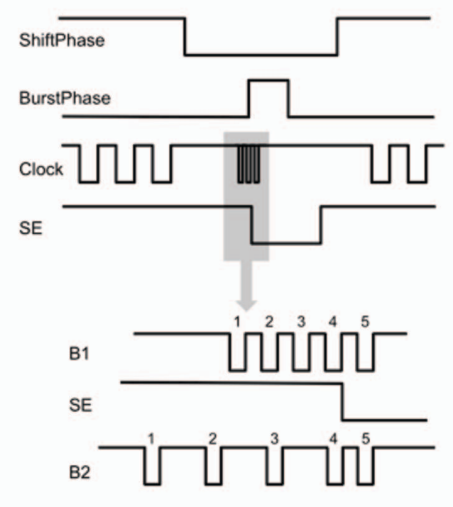

Figure 5. Example clock burst waveforms

In addition, launch-of-shift (LOS) test patterns are employed to reduce the complexity of pattern generation.

Fig. 5 shows examples of burst-mode clock waveforms. In the BurstMode, typically five consecutive clock cycles are applied. The first four cycles serve shifting purposes whereas the last one is designated for capture. The objective is to stabilize the power supply before the last shift and capture pulses are applied, which are critical for at-speed tests. In order to reduce the voltage droop related to a higher circuit activity, a burst clock controller (BCC) is used to slow down some of the shift cycles. It allows a gradual increase of the circuit activity, thereby reducing the di/dt effect. The BCC can gate the shift clocks to allow a pulse every 1, 2, 4, or 8 cycles, depending on the needs for gradually warming up the circuit. In Fig. 5, waveform B2 is an example of a burst waveform where the first three shift cycles are applied at half the speed of the system clock frequency.

In the BurstMode clocking scheme, the scan enable (SE) signal has to be kept high until the last clock cycle. This is because LOS patterns are applied for at-speed tests. A shift register with a slow speed *ShiftPhase* signal as its input generates the SE signal. It has to propagate to all flip-flops within a single system clock cycle. This is relatively easy to accomplish as most commercial layout tools can connect the shift register to an early branch of the clock tree, thus increasing the setup time margin for destination flip-flops. Furthermore, the last stage of the shift register can be replicated to reduce the fan-out wherever needed. The circuit to generate the SE signals for the LOS scheme is illustrated in Fig. 6. Typically, only one or two clock enable signals are used within a specific clock domain. These signals are pipelined because they have to switch within a single system clock cycle. However, their fan-out is small compared to the fan-out of the associated SE signal. The clock enable signals are identified by their multiplicity, i.e., CE_2, CE_3, CE_4, while the output CEF is used for false paths.

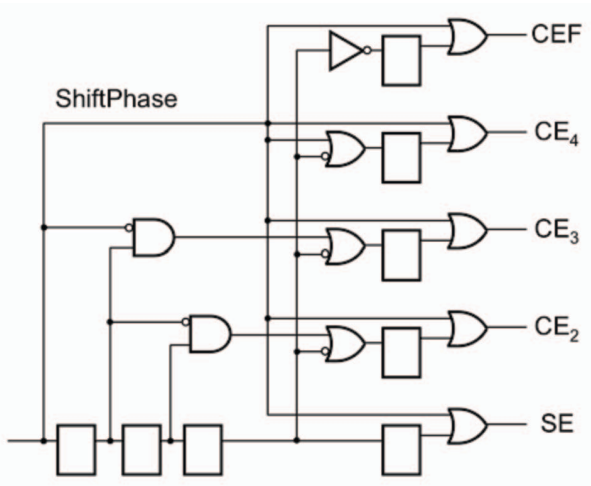

Figure 6. Generation of scan and clock enable signals

Another challenge the proposed architecture faces is how to handle scan chains during the burst phase. Since the burst can happen at the functional speed, it is impossible to drive the scan chains from either the ATE or the BIST controller. Consequently, the scan chains are arranged into circular segments based on clock domains, and several such segments are concatenated to form a scan chain. The segments behave as standard scan chains when data is shifted in from the ATE or the BIST controller during the regular shift cycles. During the burst phase, however, the scan segments are put into rotation mode, when the output of a scan segment goes back to the input, and the data is rotated within the scan segment at a very high speed. Since there are no data exchanges between the clock domains as well as between the clock domains and the BIST controller or the ATE, the exact timing during the rotation phase is not critical. Only the final shift (launch) and the capture cycles are critical and have to be placed accurately. Since the regular shift cycles are independent of the burst mode, it is possible to use a shift clock to minimize the test application time without burdening timing constraints for the scan path and without excessive power dissipation during shift.

V. Hierarchical Test Approach

Another approach to manage test power is to employ a hierarchical test methodology for the entire design. According to this scheme, a circuit is partitioned into several cores, typically corresponding to the physical regions (Fig. 7). Depending on design styles, the top-level can be clean with only instantiations of the cores, the test controller, and interconnects. On the other hand, the top-level can also contain quite a bit of glue logic, depending on a design as well as how the core isolation has been achieved.

A hierarchical test approach has been proposed in [10]. It can be implemented for core-based designs. Each core comes with a wrapper serial port (WSP) in accordance with the IEEE 1500 standard, and it is controlled by a top-level test access port (TAP). The WSP oversees all the test controllers within a core that implements either BIST, test compression, or helps in functional debug. The cores are equipped with an isolation mechanism, which is essentially a wrapper boundary register (WBR) isolating the core from external logic. The WBR controls the core inputs and observes the core outputs

during internal test, whereas it controls the core outputs and observes the core inputs during external test. A star configuration is employed to administer the WSPs and test controllers at all levels of the hierarchy. A star configuration was chosen because the cores can be operated and verified independently or in parallel. This provides an added advantage in a concurrent design environment, where the cores are being designed in parallel, but they can be validated and tested by plugging them into the design as they become available.

One of the key ideas behind implementing a hierarchical scheme is the ability to isolate a core during test. It can be done by using two types of cells – dedicated wrapper cells and shared wrapper cells. Inserting a dedicated wrapper cell mandates adding a sequential element to the core I/O just for test purposes. These cells are active during test, but they are bypassed during functional mode of operation. On the other hand, shared isolation enables sharing an existing sequential element for bounding purposes. This minimizes the area overhead as well as it does not introduce any logic on the functional paths. The choice between a shared versus a dedicated wrapper cell depends on the levels of logic and fan-out associated with each pin, so that the amount of logic that needs to be tested during external test mode does not become excessively large.

As the cores are isolated and can be tested independently of each other, the architecture of [10] allows testing any number of blocks in parallel, without exceeding the power budget. The TAP and the WSP are programmed to enable test controllers during the manufacturing test. When a particular core is being tested, the remaining cores are kept in a quiet mode. There are various ways for doing this. For example, if there are separate clock gaters for individual cores, the clock gaters for inactive cores can be shut off. Similarly, if there is not enough resolution for the clock gaters, the inactive cores can be supplied with all zeros (or ones), which reduces the switching activity for those cores. Furthermore, the SE signals for the cores can be gated in such a way that scan chains in the inactive cores can be always kept in a shift mode. It prevents a capture event and eliminates any power dissipation associated with the capture cycle. The ability to choose a subset of cores that can be tested in parallel provides the necessary flexibility to control test power dissipation. Implementing such hierarchical schemes (in addition to controlling the shift and capture switching activities as discussed in the earlier sections) gives the ultimate control on how test time can be optimized by making sure the power budgets are not exceeded.

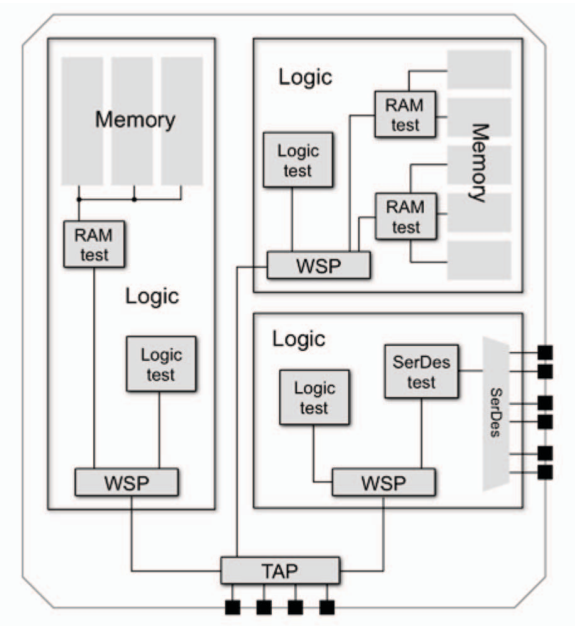

Figure 7. Hierarchical test architecture (SerDes: serializer and deserializer; TAP: test access port; WSP: wrapper serial port)

VI. Conclusion

While many techniques have evolved to address power minimization during functional operations, it is also imperative to judiciously manage power during the test mode. Circuit activity is substantially higher during test than that in the mission mode, and the resulting excessive power consumption can increase chip reliability failures leading to yield lost. The paper reviews some of our recent works devoted to scan test application schemes encompassing efficient techniques to reduce power dissipation in the embedded test environment during scan loading, capture, and unloading. The presented methods allow one to operate within specified power consumption budgets without compromising test quality, with no overheating and no risk of reliability degradation.

References

[1] K. Arabi, R. Saleh, and X. Meng, "Power supply noise in SoCs: metrics, management, and measurement," *IEEE Design & Test*, vol. 24, No. 3, May-June 2007, pp. 236-244.

[2] D. Czysz, M. Kassab, X. Lin, G. Mrugalski, J. Rajski, and J. Tyszer, "Low power scan operation in test data compression environment," *IEEE Trans. CAD*, vol. 28, No. 11, Nov. 2009, pp. 1742-1755.

[3] D. Czysz, G. Mrugalski, N. Mukherjee, J. Rajski, P. Szczerbicki, and J. Tyszer, "Low power compression of incompatible test cubes," *Proc. ITC*, 2010, paper 24.1.

[4] D. Czysz, G. Mrugalski, J. Rajski, and J. Tyszer, "New test data decompressor for low power applications," *Proc. DAC*, 2007, pp. 539-544.

[5] D. Czysz, G. Mrugalski, J. Rajski, and J. Tyszer, "Low power embedded deterministic test," *Proc. VTS*, 2007, pp. 75-83.

[6] P. Girard, N. Nicolici, X. Wen (ed.), *Power-Aware Testing and Test Strategies for Low Power Devices*, Springer, New York 2010.

[7] X. Lin and J. Rajski, "Test power reduction by blocking scan cell outputs," *Proc. ATS*., 2008, pp. 329-336.

[8] E.K. Moghaddam, J. Rajski, and S.M. Reddy "Low power compression utilizing clock-gating," *Proc ITC,* 2011, paper 7.1.

[9] E.K. Moghaddam, J. Rajski, S.M. Reddy, X. Lin, N. Mukherjee, and M. Kassab, "Low capture power at-speed test in EDT environment," *Proc. ITC*, 2010, paper 24.2.

[10] B. Nadeau-Dostie, S. Adham, and R. Abbott, "Improved core isolation and access for hierarchical embedded test," *IEEE Design & Test*, vol. 26, No. 1, Jan. 2009, pp. 18-25.

[11] B. Nadeau-Dostie, K. Takeshita, and J.-F Cote, "Power-aware at-speed scan test methodology for circuits with synchronous clocks," *Proc. ITC*, 2008, paper 9.3.

[12] J. Rearick and R. Rodgers, "Calibrating clock stretch during AC scan testing," *Proc. ITC*, 2005, paper 11.3.

[13] S. Remersaro, X. Lin, Z. Zhang, S.M. Reddy, I. Pomeranz, and J. Rajski, "Preferred fill: a scalable method to reduce capture power for scan based designs," *Proc. ITC*, 2006, paper 32.2.