J Electron Test (2017) 33:157–169
DOI 10.1007/s10836-017-5656-y

A Parallel Test Application Method towards Power Reduction

Ding Deng[1] · **Yang Guo**[1] · **Zhentao Li**[1]

Received: 14 July 2016 / Accepted: 15 March 2017 / Published online: 25 March 2017

Abstract As the serial scan design has been one of the most popular methods in VLSI circuit test, power consumption during test increases significantly because of its inherent shift mode. To solve this problem, this paper proposes a novel test scheme, which makes a few improvements in the traditional scan architecture and adopts a new two-phase approach. First, each clock chain is activated in turn and the vectors for scan cells in the activated chain are applied in parallel within a test clock period. Second, after one pattern has been applied completely, all chains are activated to capture the response altogether. In addition, a compression algorithm is proposed to augment the parallelism of our method. Experimental results on benchmark circuits and industrial modules show that, compared with the traditional serial scan scheme, the proposed approach can reduce average power by 88.98% and peak power by 59.99% at acceptable area and wire length cost.

Keywords Low power · Scan test · Parallel · Compression

Responsible Editor: A. Orailoglu

✉ Ding Deng
dengding15@nudt.edu.cn

Yang Guo
guoyang@nudt.edu.cn

Zhentao Li
lizhtao@nudt.edu.cn

[1] School of Computer, National University of Defense Technology, Changsha 410073, China

1 Introduction

Serial Scan (**SS**) design, which makes it possible to test sequential circuits with reduced complexity in practical time, is one of the most widely used Design-For-Testability (**DFT**) techniques in industry [4]. However, as the complexity of digital circuits increases, it suffers from enormous challenges in high power dissipation. Power dissipation is usually measured in two aspects: average power and peak power. Average power, which has a close relationship with heat dissipation, imposes strict requirements on the cooling mechanisms. Dissipating excessive heat too long during test may degrade the circuit performance. Even worse, it can lead to a yield loss or undesirable malfunctions which in turn increases the chip cost. On the other hand, peak power determines the thermal and electrical limits of circuits [19]. Once the Circuit Under Test (**CUT**) exceeds the limit, reliability cannot be guaranteed. Varieties of alternative approaches have been proposed to tackle with these problems.

Scan clock was slowed down in [5] resulting in test time elongation. Reordering scan cells after patterns are generated is a basic approach to suppress high activities [22]. Several modifications have been also tried in scan cells to mask undesirable toggles in [11, 31]. Instead, Bhattacharya et al. [3] developed another "double tree" structure to keep power to a lower level on the scan path. Chandra et al. [6] came up with a deferred-broadcast architecture aiming to save the scan-in power. However, these savings may correspond to only a small fraction of the overall scan power when filling techniques are used. To save scan-out power, a complementary solution named expedited-compact technique was proposed in [20] targeting the scan-out power reduction. A reconfigured scan forest architecture was proposed to reduce test power in [26], which can also save test time and compress data volume. In [28], each chain was divided into several segments with

different segments activated separately to save power. However, this approach will bring challenges for routing as scan enable and scan clock come from the same signal. Actually, the scan clock must arrive a bit later than the scan enable signal; otherwise the stimuli cannot be applied into the scan cells.

Another architecture called Random Access Scan (**RAS**) was proposed in [13] and has been applied to AMDAHL [23]. Based on this architecture, test time and data volume, consequently power consumption were all reduced greatly in cooperation with the X-identification and compression/scan co-design techniques in [1, 15]. But the associated routing complexity and extra decoder area made it impractical in the past. A toggle RAS [16, 17] to remove two global signals and a Localized Random Access Scan (**LRAS**) method [30] were proposed to address these problem, but both of them ended up with a few other inherent limits.

From the aspect of test pattern, much effort has also been expended on power reduction [24]. By exploiting the correlation of the neighboring patterns and responses, the reorder technique can be well utilized in most cases [10, 12, 21]. Grouping for launch-on-capture delay testing can also reduce the power consumption as well as the test data and responses [27]. Enokimoto et al. [9] proposed an algorithm to guarantee capture safety.

For System On Chip (**SOC**), additional emphasis was laid on test scheme [7, 8] and compression [14, 29]. A scan feed-forward scheme was proposed in [18] which improved test compression. Wohl et al. [25] proposed a multi-level scan compression architecture to deal with the high test volume problem effectively.

In this paper, a Parallel Test Application (**PTA**) test scheme is proposed to reduce power without resorting to any special Automatic Test Pattern Generation (**ATPG**) algorithms. The proposed technique has the following features:

1) Clock chains are activated separately to avoid undesirable toggles in the scan cells when stimuli are assigned, thus concurrently reducing average power and peak power.
2) Traditional scan chains are cut off and stimuli are applied in parallel through another special bus. Due to this parallel transmission, ripple propagation effects between scan cells are effectively removed.
3) Vector assignment and response collection method is non-destructive, which can reduce the effort for fault diagnosis. Additionally, chain test is not required, thus saving test time.
4) A compression algorithm is proposed to augment parallelism in PTA. It is also proved to be beneficial in reducing power consumption and test time.

The rest of this paper is organized as follows. In Section 2, the proposed PTA architecture and modified scan cell are presented, respectively. Time cost, power consumption, area and routing overhead as well as timing closure of this method are analyzed in Section 3. A compression algorithm to minimize the scan input/output (**I/O**) ports is proposed in Section 4. Experimental results on ISCAS89 benchmark circuits and industrial modules are presented in Section 5. Section 6 concludes this paper and discusses the future improvement suggestions.

2 PTA Approach

To avoid the enormous power dissipation induced by shift operations in SS test, it is necessary to break up the scan chain and evade the serial shift assignment. In this Section, we develop a parallel assignment strategy for the D flip-flops (**DFF**s). Because of its parallelism, we successfully eliminate the undesirable activities induced by propagation effect along with the scan chains in traditional SS method.

2.1 PTA Architecture Design

To carry out the PTA approach, a few modifications must be made on the traditional SS architecture. Figure 1 shows the PTA architecture.

The Top module consists of two parts: the controller and the modified CUT. The rectangles standing side by side in the CUT represent the modified scan cells. In the rest of this paper, we refer to them as the **PTA-DFF**s. Assuming that there are $m \times n$ DFFs in the original CUT, we divide all the DFFs into m rows and n columns. Here m denotes the number of separate scan clock signals (or scan enable signals) and n denotes the number of DFFs in each row. However, in practice, the total number of the DFFs often fails to equal $m \times n$ exactly. To cope with this challenge, we add dummy cells to the CUT. As shown in Fig. 1, we refer to the PTA-DFFs with the same scan clock (or scan enable) signals in a row as a **CK_chain** and the PTA-DFFs in the same column as a **CL_chain**. Determining m and n is a tradeoff between I/O overhead and test time cost. The larger m is, the longer is the test time that needs to be expended. The larger n is, the more I/O ports are needed. Each CL_chain is connected to a Scan In (**Sin**) bus and a Scan Out (**Sout**) bus. Through the Sin bus, stimuli can be applied to every PTA-DFF as long as the localized Sc_en and Sc_clk are in active state. Similarly, the captured response can transmit out to the Automatic Test Equipment (**ATE**) for comparison via the Sout bus.

The controller receives the Test Clock (**Test_clk**), Reset and Test Enable (**TE**) signals from ATE and generates m separate Scan Clock (**Sc_clk**) and Scan Enable (**Sc_en**) signals for all the CK_chains respectively in the CUT. All Sc_en pulses last one period of the Test_clk while all Sc_clk pulses only last half the period of the Test_clk. Thus, the frequency of

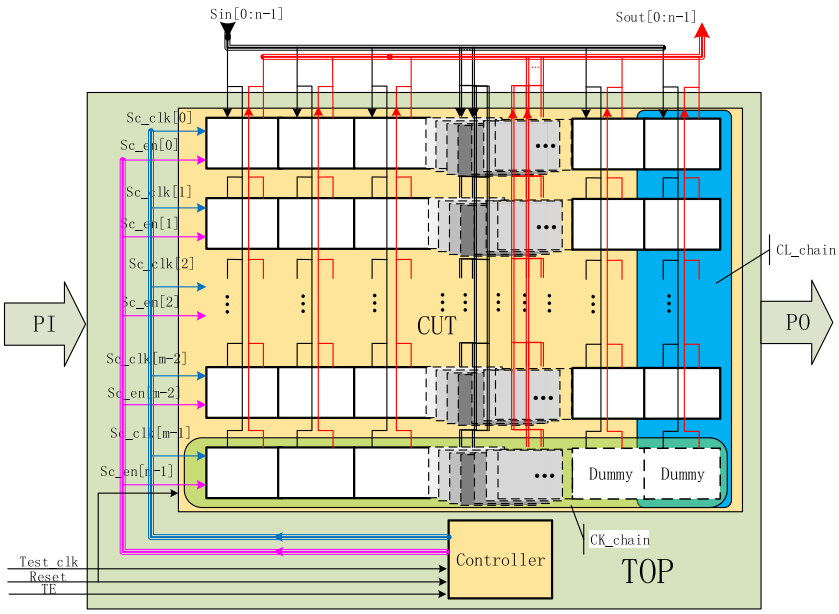


Fig. 1 Parallel test application architecture

our method can be configured by modifying the frequency of the Test_clk on the ATE.

2.2 PTA Scan Cell Structure and Controller Implementation

Figure 2 illustrates the structure of PTA-DFF in our design. As mentioned above, the key idea to reduce scan-shifting toggles is to assign test vectors directly. To satisfy this demand without additional modifications on the traditional SS scan cell, we still use the SI pin to apply stimuli. However, the stimuli come directly from another propagation path named Sin bus instead of the Q pin of the former scan cell in traditional SS method. A tri-state buffer is inserted after the Q pin of each scan cell. Assuming that the output pin Z of tri-state buffer is enabled when the Output Enable (**OE**) signal switches to high, we connect the OE pin of the tri-state buffer to the Scan Enable (**SE**) pin and all the Z pins of tri-state buffers in the same CL_chain to a Sout bus.

The function of PTA-DFF can be defined as three operations determined by two signals (Clk and SE). More details are presented in Table 1.

Whenever the SE is set to '0', the PTA-DFF samples data from the D pin at the rising edge of Clk. Then the subsequent logic (combinational or sequential) can get the stable value from the Q pin. Meanwhile, the tri-state buffer is disabled so that the Sout bus maintains high-impedance (denoted by 'Z') state. At this time, PTA-DFF acts as a normal DFF for **normal-working** mode. In the test mode, the SE switches to '1', resulting in the value on the Q pin being able to propagate through tri-state buffer to the Sout bus for observation. During the inactive period of Clk, the value on the Z pin reflects the response of the last capture. It is the **response-output** mode. Once the Clk turns to active, the PTA-DFF samples stimuli from SI pin and transmits the value to the Z pin. It works in the **vector-application** (and observation) modes at this time.

The controller can be mainly implemented by a modulo-m counter,where m is the number of CK_chains in the CUT. The count value i enables corresponding Sc_en[i] and Sc_clk[i] signals for the specific CK_chain[i]. When the count value increases up to m, one pattern has been assigned completely. Right after that, all CK_chains should be activated in the next cycle to capture the response of the last pattern. To simplify the test process, we only focus on the combinational test patterns where only one capture cycle is applied between scan load and scan unload. The sequential test patterns with more than one capture cycle are not considered in this paper.

2.3 PTA Process Explanation

The entire process of our PTA approach for a CUT containing four CK_chains is shown in Fig. 3. Assume that there are n PTA-DFFs in each CK_chain and two test patterns are applied. As soon as the Test Enable (TE) signal from the ATE switches to high, the controller bypasses the functional clock of the CUT and starts to generate Sc_clk and Sc_en for the test

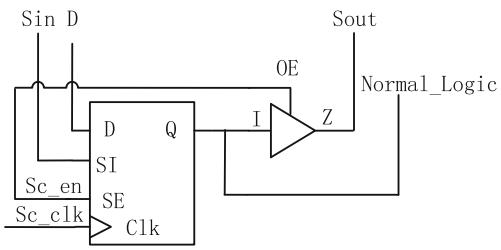


Fig. 2 PTA scan cell structure

Table 1 Operation modes of PTA-DFF

Function	Clk	SE
Normal-working	Active	0
Response-output	Inactive	1
Vector-application	Active	1

operation according to the Test_clk. The Sc_clk and Sc_en of each CK_chain are activated one by one in a certain order, meaning that only one Sc_en and corresponding Sc_clk are in active state at most during the test mode. Each pulse of Sc_en lasts a period of the test clock. And it always starts half a cycle before the rising edge of the corresponding Sc_clk. Once the Sc_en[*i*] is activated, the test vector for the whole *n* PTA-DFFs of CK_chain[*i*] is prepared on the Sin[0:n-1] bus. And as the rising edge of Sc_clk[*i*] comes, the vector is concurrently applied to the *n* PTA-DFFs of CK_chain[*i*] at the same time. In the next cycle, another CK_chain is activated and applied to stimuli while the assigned CK_chains hold previous values because their Sc_clk and Sc_en signals are inactive.

Because there are some correlated logics among CK_chains, it is necessary to capture altogether to avoid mutual impacts. After all CK_chains have been assigned, all Sc_en signals switch off to convert the CUT into normal-working mode. Meanwhile, the Primary Inputs (**PI**s) of the CUT are set to the corresponding values in the specific pattern. Then, all Sc_clk signals are activated for a cycle to capture the response of the last pattern.

It should be noted that each CK_chain, as well as its corresponding Sc_en and Sc_clk, is only activated once in a certain order for each pattern. We refer to the procedure where the CK_chains are activated in turn as a **round**. Except the first and the last round of Sc_en, the response-output operation of the last pattern and the vector-application operation of the current pattern for a certain CK_chain are conducted within the same Sc_en pulse. Specifically, in the first half of the Sc_en pulse, the response of the last pattern transmits to the Sout bus because the tri-state buffer behind the DFF is output enabled at this moment. Thus we can get an n-bit response of the *n* PTA-DFFs on a certain CK_chain within one cycle for comparison with expectation. While in the second half of the Sc_en pulse, the *n* PTA-DFFs of a certain CK_chain sample the vector for them at the rising edge of the corresponding Sc_clk. After the rising edge of the Sc_clk, the value on the Q pin of DFFs reflects the vector just applied to the DFFs, which can also be observed on the Sout bus. Therefore, we can verify whether the stimuli have been assigned to the PTA-DFFs correctly during logic test.

Attention should be paid that the response-output and vector-application operations are both non-destructive in PTA. This non-destructive property allows snap-shot of circuit states at any CK_chain, which benefits fault diagnosis a lot. In conventional serial scan, it is impossible to achieve this goal unless adding shadow latches, because the states of the circuit are serially shifted out (and in). Another difference between the PTA method and the segmentation design [28] is a phase difference between Sc_en[*i*] and Sc_clk[*i*]. In PTA, the scan enable and scan clock signals have already got a phase difference of half a period since they are generated from the controller. Hence, it almost avoids the routing problems in [28] as explained in introduction.

3 Theoretical Analysis

3.1 Test Time Cost

In the SS approach, stuck-at fault test is conducted in two steps: chain test and logic (scan) test. During the chain test, a

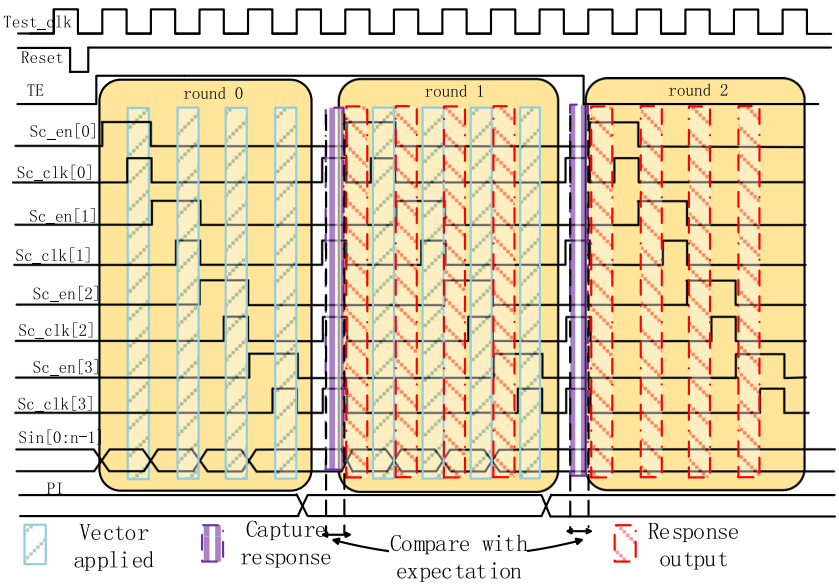


Fig. 3 PTA scheme timing example

test vector "00110011……" is applied to every scan chain. It takes $T_{Chain} = n + 4$ cycles. As mentioned in Section 2.2, sequential patterns are not considered in this paper. Therefore, the logic test takes $T_{logic} = p(n + 1) + n$ cycles. In sum, the SS scheme takes T_{SS} cycles as shown in eq. (1).

$$T_{ss} = p(n + 1) + 2n + 4 \quad (1)$$

Here n denotes the number of DFFs in each scan chain and p is the number of test patterns.

On the other hand, because PTA can verify whether the stimuli have been assigned correctly to PTA-DFFs during the vector-application operation, only logic test is required. Thus PTA approach takes T_{PTA} cycles in total:

$$T_{PTA} = p(m + 1) + m(2) \quad (2)$$

Here m is the number of CK_chains in the CUT and p is the number of test patterns.

Comparing (3.1) with (3.2), the test cycle reduction ΔT can be achieved as follows:

$$\Delta T = T_{ss} - T_{PTA} = p(n - m) + (2n - m) + 4 \quad (3)$$

To avoid extra I/O ports overhead, we let $m = n$. Hence, the test time reduction is achieved from the eq. (4),

$$\Delta T = n + 4 \quad (4)$$

In summary, because chain test is not required in PTA, our method saves at least (n + 4) cycles compared with the traditional SS method.

3.2 Power Reduction Analysis

PTA decreases the average power considerably in the following three aspects.

1) The activities between scan cells caused by the continuous "01" or "10" in stimuli or responses during traditional shift mode.
2) The extra activities of combinational logic results from the undesirable toggles of scan cells during bits shifting.
3) The clock tree power consumption. As previous studies demonstrated, for VLSI, the major portion of the energy consumption is dissipated in the clock tree [11]. In SS method, all scan chains are active simultaneously during the entire test process to reduce test time. Nevertheless, PTA scheme only offers one active clock to a certain CK_chain when stimuli are applied.

Meanwhile, it can also probably reduce the peak power in comparison to the SS approach as the CUT becomes more complex. It is because of this, for SS scheme, all scan cells and their subsequent logic tend to be triggered all the time, especially during the shift mode when the stimuli shift from '1' to '0' (or from '0' to '1'). The larger the CUT is, the higher the probability the peak power might occur in the shift period. However, for the PTA scheme, only one CK_chain and its subsequent logic are activated when the stimuli are applied. So it is almost impossible for PTA to have peak power dissipation during vector-application period. Instead, peak power nearly always occurs in the capture operation during which all CK_chains are activated. Therefore, conclusion can be drawn that if the peak power dissipation occurs during the shift mode in SS method, it can certainly be reduced in the PTA approach. The reduction ratio strongly relies on the gap between the highest shift-power consumption in SS and the highest capture-power consumption in PTA.

3.3 Area and Routing Overheads

Due to the modification of CUT and additional controller, the PTA architecture has a few area overhead compared with the traditional architecture. The overhead mainly comes from three aspects.

1) Modified DFFs in the CUT

Assume that the CUT has N DFFs in total and the area of each tri-state buffer is A_{tri}, then the area overhead of the modified DFFs in the CUT can be approximately calculated by eq. (5):

$$\text{Area overhead}_{\text{modified DFFs in the CUT}} = (N \times A_{tri}) \Big/ \text{original area} \times 100\% \quad (5)$$

2) Routing overhead of scan input/output in the CUT

Every DFF is directly connected to one of the scan inputs and outputs in the PTA architecture. Routing scan inputs/outputs to every single DFF increases wire length to some extent, which might also lead to some area overhead to avoid routing congestion. This kind of overhead highly depends on the interior structure of the CUT, the number of CL_chains denoted by n, as well as the placement and routing strategies. Therefore, there are no concrete computational formula to evaluate the routing overhead.

3) Routing and area overhead from the controller

The controller is implemented by a modulo-m counter and many clock gating cells. The area overhead of controller mainly depends on the number of CK_chains denoted by m. Besides, routing separate Sc_clk and Sc_en signals to each

CK_chain also increases the wire length. This kind of routing overhead highly depends on the parameter m, as well as the placement and routing strategies. So no concrete computational formula can be given.

In summary, the routing overhead of PTA is mainly determined by the parameter m and n. For example, as m decreases, the routing overhead of separate Sc_clk and Sc_en signals also decrease, as well as the area overhead of the controller. However, because $N = m \times n$, therefore n increases when m decreases. Hence, the routing overhead of the scan input/output will increase when m decrease. Therefore, the extreme big (or small) value of m is not appropriate for PTA from the perspective of routing overhead.

Although it's hard to determine the optimal value of m(or n), tentative experimental results show that the routing overhead changes within an acceptable range even though m varies a lot. Take our industrial module named M in Section 5.3 as an instance, the average wire length overhead only changes in a range of 11.91% ~17.83% when m varies in a range of 7 ~ 60. Therefore, if the routing overhead is not the first consideration, DFT designers can set m (or n) as what they used in traditional DFT methods.

3.4 Timing Closure

It is undeniable that the timing is a little bit worse in our method because the stimuli are directly applied by the scan input port to every DFF. However, because nowadays the test procedure is almost conducted under a slow clock frequency such as 25 MHz, timing closure of the path which starts from the scan input to the remotest DFF is still easy to meet. As for the at-speed test, the vector-application operation is also conducted under a slow clock. Even though the launch and capture operation is conducted under the high speed functional clock, data do not go along the scan I/O path in these operations. Our method adds no extra logic in the critical functional path, so it does not significantly affect the timing in the normal-working mode. Even tentative experiments have been done under a rigid constraint of a functional clock at 1GHz, the worst slack of reg-to-reg in critical functional path of the CUT only deteriorates 18 ps compared with the traditional SS method.

In addition, with careful design and tight timing constraints on the controller, skew between different Sc_clk signals and different Sc_en signals for every CK_chain can be kept within an acceptable range. The latency from the Tesk_clk to the separate Sc_clk and Sc_en can also be kept low and within acceptable range.

4 Equal-Mode Compression Algorithm

As analyzed in Section 3.1, to avoid extra I/O port overhead, the number of CK_chains should be equal to the number of DFFs in a traditional scan chain. It will lead to a large controller overhead. To increase the parallelism without adding extra I/O ports, we specially propose an equal-mode (**EM**) compression algorithm for our PTA architecture based on the theory of [26]. There are two major definitions, mostly following the terminology of [26].

Definition 1: For a group of scan cells, if any pair has no subsequent combinational logic in common,

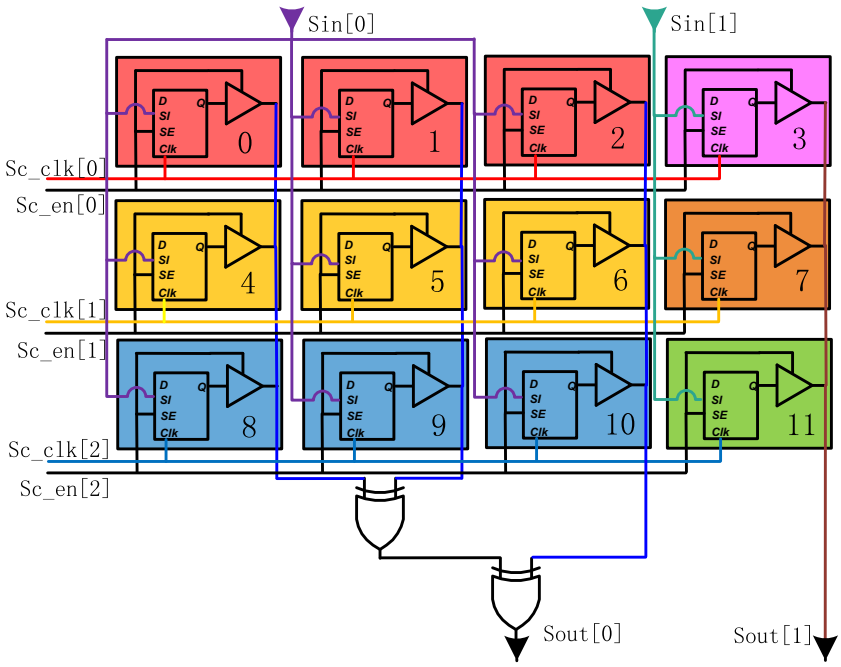


Fig. 4 PTA architecture with EM compression and exceptional DFFs

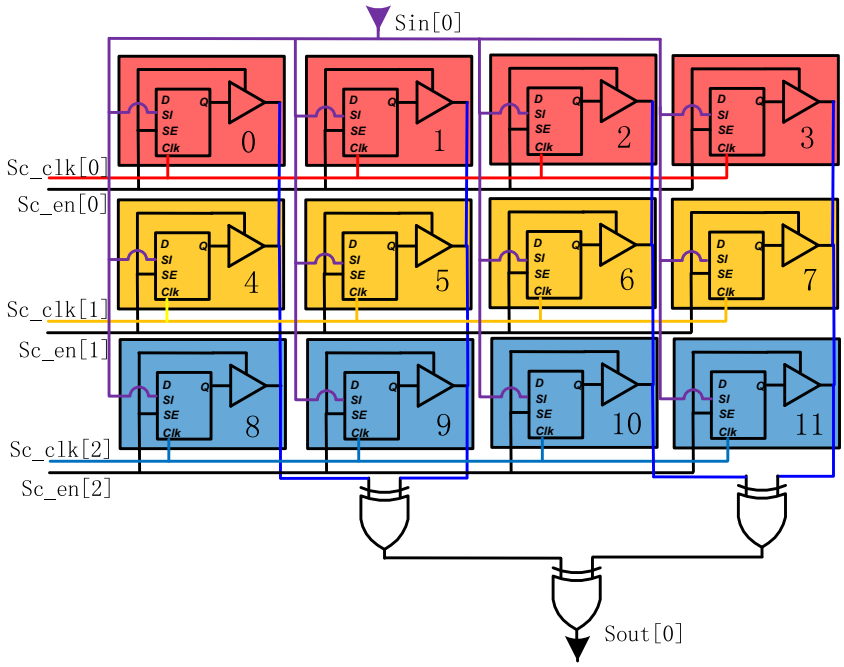


Fig. 5 PTA architecture with EM compression without exceptional DFFs

then they can share a common scan input port in the test mode.

Definition 2: For a group of scan cells, if any pair is not driven by common combinational logic, then their responses that are captured during test mode can be compressed through an XOR-tree.

Based on these two definitions, the final architecture with compression is illustrated in Fig. 4. There are 3 CK_chains each of which contains 4 DFFs. Assumes that for the first three DFFs of each CK_chain, any two of them meet both the definition 1 and 2. Then we can connect the SI pins of them to a common scan input port as Sin[0] in Fig. 4. On the other side, the Z pins of their tri-state buffers can be connected to an XOR-tree. Finally, only one scan output port driven by the XOR-tree is needed such as the Sout[0] in Fig. 4.

On the contrary, assumes that the fourth DFF of each CK_chain meets neither of the two definitions with the first three DFFs, which are referred to as **exceptional DFF**s in the remainder of this paper. It cannot share Sin[0] or Sout[0]. To guarantee the observability and controllability of these exceptional DFFs, extra scan I/O ports should be specially added for them just like Sin [4] and Sout [4] in Fig. 4.

A perfect condition occurs when all DFFs in the CUT can be divided into groups of equal members as shown in Fig. 5. In this case, all PTA-DFFs in a group are stitched in the same CK_chain and each group shares a common scan input/output port. The EM compression algorithm proposed in this paper aims at finding out these groups, i.e. the EM compression algorithm provides a novel stitching strategy to minimize the scan I/O port overhead in PTA architecture.

Table 2 Necessary data for DFF *i*

Variable	Content
CFIC[*i*]	The constraint items for scan input compression of DFF *i*
WIC[*i*]	The number (weight) of constraint items in CFIC [*i*]
CFOC[*i*]	The constraint items for scan output compression of DFF *i*
WOC[*i*]	The number (weight) of constraint items in CFOC[*i*]

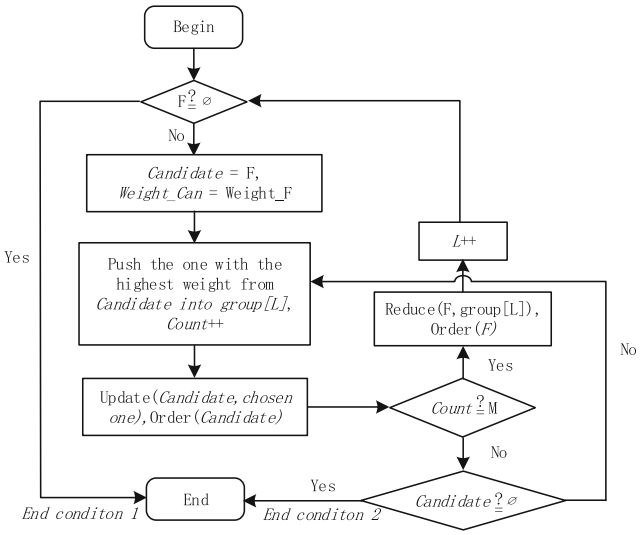


Fig. 6 Flowchart of EM compression algorithm

4.1 Data Preparation

Tracing back from the data input pin D of each DFF until to the Primary Input (**PI**) or a data output pin Q of another DFF (**PPI**) in the netlist, we can get the following information to implement EM compression as listed in Table 2.

In Table 2, the constraint items of DFF *i* denote the name of the DFFs which cannot be placed in the same group with DFF *i*.

4.2 Procedure of EM Algorithm

Assuming the expected compression ratio is set as *M*, the EM algorithm is illustrated in Fig. 6.

Algorithm Inputs:

F: the set of all DFFs in the CUT

N: the number of the total DFFs

Constraint[i]: the constraint condition set for DFF i. ($i \in (0, 1, 2\ldots, N\text{-}1)$)

Weight[i]: the number of the constraint items in Constraint[i]. ($i \in (0, 1, 2\ldots, N\text{-}1)$)

Weight_F: the set of all weights of all DFFs

Algorithm Outputs:

group[L]. ($L \in (0, 1, 2, \ldots \left\lfloor \frac{N}{M} \right\rfloor)$). Each group contains M DFFs.

Remainder: the set of remaining DFFs after the algorithm is accomplished. It is not ∅ if there are some exceptional DFFs or N%M≠0.

Function:

Update(Set1, Set2): exclude Set2 and the ones which meet the constraint conditions of Set2 from the Set1, i.e., $\text{Set1} = \text{Set1} - (\text{Set2} + \sum_{k \in Set2} \text{Constraint}[k])$

Reduce(Set1,Set2): exclude Set2 from Set1, i.e., $\text{Set1} = \text{Set1} - \text{Set2}$

Order(Set1): place the items of Set1 according to the order from the highest weight to the lowest weight.

To compress the scan input port, Constraint[i] and Weight[i] are substituted with CFIC[i] and WIC[i] respectively. After this procedure, if Remainder = ∅, i.e., the algorithm terminates because of the end condition 1. It means that for any $T \leq M$, all DFFs in the CUT can be divided into multiple groups each of which contains T DFFs. Otherwise, the algorithm ends because of the end condition 2, meaning that some DFFs are left so that extra scan I/O ports should be added or lower M can be set to see whether these DFFs are eliminated. Similarly, to compress the scan output port, we substitute the Constraint[i] and Weight[i] with CFOC[i] and WOC[i] respectively. In order to compress the scan input and output together, we substitute the Constraint[i] and Weight[i] with (CFIC[i] ∪ CFOC[i]) and the corresponding weight.

5 Experimental Results and Comparison

As reported in [11], over 90% of the total power consumption results from the switching activity of nodes in static CMOS circuits. Most previous studies utilized a model named Weighted Switching Activity (**WSA**) [2] to estimate power consumption. Some even ignored the capacitance and directly assumed that the power dissipation was proportional to the number of transitions. But it may no longer be accurate as the technology scales down below 60 nm as the leakage current and internal power consumption rise. Hence, we calculated the real power by Prime Time PX with VCD files generated after fault simulation at 250 MHz test clock, just as the power estimation flow in [6]. The clock tree power, controller power and leakage power are all included in the calculation for all experimental data.

Results of four sets of experiments are presented in this Section. The first experiment compares the power reduction of PTA with the method in [11] that is evaluated in our power estimation flow. We refer to the method in [11] as **NOR-Mask** in the rest of this paper. The second experiment explores the relationship between the chains and power reduction of PTA structure. In the third experiment, four industrial circuits are adapted for PTA structure to evaluate the practicability and effectiveness of our method. The fourth experiment provides the performance of our EM compression algorithm used with PTA structure. Because the emphasis of this paper is the power reduction ability rather than the compression ability of PTA architecture, the compression algorithm proposed in Section 4 is only applied in the fourth experiment to evaluate its validity

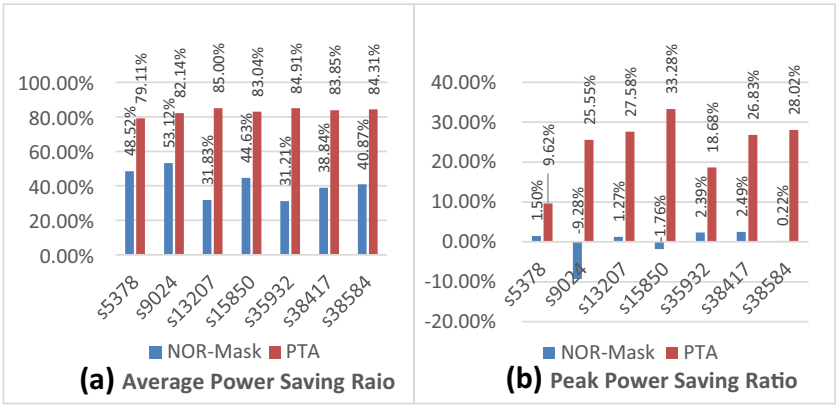


Fig. 7 Normalized power reduction compared with NOR-Mask

and efficiency. Therefore, the advantages and disadvantages of PTA architecture can be obtained from the first three experiments more directly and clearly.

5.1 PTA vs. NOR-Mask

Experiments are carried out on seven ISCAS89 benchmark circuits synthesized and inserted with full scan in 40 nm technology. ATPG was realized by TestKompress from Mentor Graphics.

Figure 7 demonstrates the capability of power saving with PTA. The power saving ratio (**SR**) is calculated by:

$$SR = 1 - Power_{new\_structure} \Big/ Power_{SS} \times 100\% \quad (6)$$

As shown in Fig. 7a, with the same number of scan I/O ports as SS and NOR-Mask, average power can be reduced by 83.19% over SS and by 41.9% over NOR-Mask on average in PTA. As for peak power consumption presented in Fig. 7b, PTA also outperforms the NOR-Mask to some extent. Because the NOR-Mask approach must assert the control signal to mask the shifting toggles of the scan cells after capture, it may even augment the peak power just as shown for s9024 and s15850. But in PTA, peak power always shows a saving more or less, up to a reduction of 33.28%. Overall, the experimental results show that the PTA scheme has a great potential to reduce average power and peak power dissipation simultaneously.

5.2 Power Reduction with Various Scan Chains

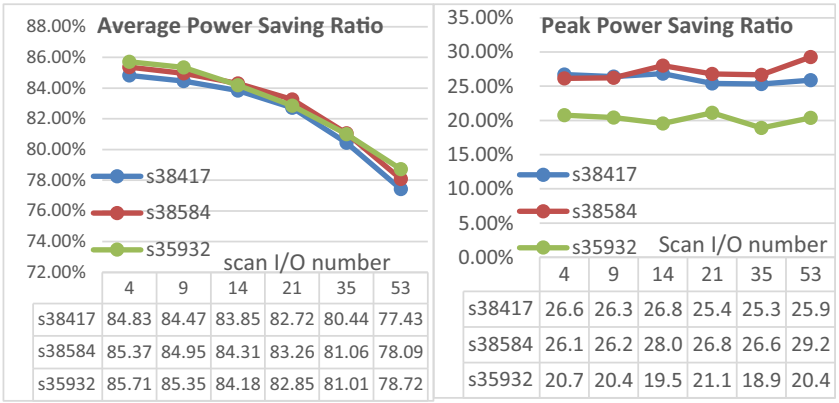
As average power has a tight relationship with test time which strongly depends on the number of scan chains (CL_chains in PTA), we conducted experiments on three benchmark circuits where the number of CL_chains (i.e. scan I/O ports) varied from 4 to 53. Figure 8 presents the trend of the power reduction as the number of CL_chains increases. As shown in Fig. 8, a clear decrease in average power reduction can be observed for all the three instances. That is because the traditional scan chains become shorter and require fewer shifts per test pattern in SS structure as more scan I/O ports are provided. Fewer shifts lead to fewer undesirable propagation power in traditional SS method. Hence, the average power saving ratio decreases as the formula (5.1) shows. Even though, the deterioration of average power reduction in PTA is actually very slight compared with [6, 20]. Our method only suffers 7.4% decrease as the number of CL_chains increases by 49. It is because that PTA makes modification on every scan cell while [6, 20] only modify a certain segment or fraction of the chains. The peak power reduction highly depends on the interior structure of the CUT as explained in Section 3.2. Hence, there is no obvious tendency as the number of scan I/O ports varies.

5.3 Industrial Cases

Table 3 provides the power reduction results on four practical industrial modules. Circuits b18 and b19 are two of the largest ITC'99 benchmarks while M and L1P are two components in the DSP processor implemented by our institute. M is a multiply unit that has 56 K cells (408 K library primitives) and 2.6 K scan cells. L1P is one of the memory components in the DSP core module which contains 19 K cells (541 K library primitives) and 3.8 K scan cells. All of them are synthesized in 40 nm high performance technology. ATPG are realized by TestKompress from Mentor Graphics. The power measurement is similar to the above experiments. The layout design is conducted by Innovus v15.20 from Cadence.

Columns 5 through 9 denote the saving ratio of average power (**SRAP**), saving ratio of peak power (**SRPP**), controller power ratio (**CPR**), area utilization overhead (**AUO**) and average wire length overhead (**AWLO**), respectively. We make two observations:

(1) From SRAP and SRPP results, we can see that PTA can reduce 88.98% average power and 59.99% peak power at most. Module M only gets 2.26% peak power



	4	9	14	21	35	53
s38417	84.83	84.47	83.85	82.72	80.44	77.43
s38584	85.37	84.95	84.31	83.26	81.06	78.09
s35932	85.71	85.35	84.18	82.85	81.01	78.72

	4	9	14	21	35	53
s38417	26.6	26.3	26.8	25.4	25.3	25.9
s38584	26.1	26.2	28.0	26.8	26.6	29.2
s35932	20.7	20.4	19.5	21.1	18.9	20.4

Fig. 8 Power reduction for PTA with various scan I/O ports

reduction, because its combinational logic proportion is much higher than its sequential proportion. In our method, the combinational logic driven by scan cells still toggles when stimuli are applied in parallel, which weakens the ability to reduce peak power.

(2) The proposed method takes a little more area and routing cost than the traditional method, especially for L1P. That is because L1P is a memory component whose sequential logic proportion is relatively higher than the other three circuits. The modifications of our PTA architecture are mainly made on the sequential logics, i.e. the DFFs, so the higher sequential logic proportion means more routing overhead. Through careful observation and comparison with the traditional method, we find the area overhead mainly results from the tri-buffers while the routing overhead almost comes from the scan I/O that is directly connected to the DFFs in the same CL_chains, which exactly coincides with our routing analysis in Section 3.3.

Although the routing cost seems a bit large, our method can still achieve smaller *Average Power-Routing Cost Product* (**PCP**). For instance, for module L1P, its PCP in PTA can be reduced to 13.14% (=(1 + 19.26%) × (1 − 88.98%)) of the original PCP of SS scheme, demonstrating that our proposal is more effective.

5.4 The Compressibility of EM Algorithm

Table 4 provides the compression results of six benchmark circuits with our EM algorithm. The second to the fourth columns represent the maximum compression ratio of scan input (under constraints of definition 1 in Section 4), scan output (under constraints of definition 2 in Section 4) as well as both scan input and output (under constraints of definitions 1 and 2), respectively. It is measured by,

$$\text{compression ratio} = \frac{\text{the number of CL\_chains}}{\text{the number of scan ports}} \tag{7}$$

Take s35932 as an example in Table 4, we only need one common scan input port for 102 CL_chains using our EM algorithm. The compression ratios of scan output for s38584 and s35932 are very small because there are always some exceptional DFFs left at the end of our EM procedure. For s35932, although more than 1000 DFFs can achieve a scan output compression of 66, there always remain almost 200 exceptional DFFs no matter how low the parameter M (the expected compression ratio in Section 4.2) is set. To avoid fault coverage loss, 13 extra scan I/O ports need to be added specially for the exceptional DFFs. Therefore, the ultimate compression is calculated as (66 + 13)/(1+13) =5.64.

Experiments to demonstrate the validity of our EM algorithm are conducted on four benchmark circuits. Two ATPG tools are used. As presented in Table 5, the second row denotes the compression ratio and compression type where CIO represents compressing both scan input and output, CI represents only compressing scan input. The row of Pat_ata represents the number of test patterns generated by ATALANTA which performs poorly in its own compression option. The row of Pat_men indicates the number of patterns generated

Table 3 Performance of industrial circuits with PTA architecture

Circuit	Scan I/O	Cells	DFFs	SRAP (%)	SRPP (%)	CPR (%)	AUO (%)	AWLO (%)
b18	8	31,518	3308	88.48	53.36	5.00	3.95	7.87
b19	10	62,343	6618	88.78	59.99	3.80	1.99	9.06
M	12	56,637	2616	82.98	2.26	1.80	2.92	11.91
L1P	8	19,478	3872	88.98	33.21	7.50	3.47	19.26

Table 4 EM compressibility on benchmark circuits

Circuit	CP_IN	CP_OUT	CP_INOUT
s5378	3	4	3
s13207	3	9	3
s15850	3	4	3
s38417	16	15	10
s38584	10	1.14	1
s35932	102	5.64	5

by TestKompress from Mentor Graphics, whose compression ability is more powerful. The "Traditional" in the table indicates the condition that the CUT is designed as a one-chain with full scan structure. The "PTA with EM algorithm" means that the CUT is modified as the PTA structure with the compressibility of scan I/O ports such as the case in Fig. 5, which also has only one scan input and output as the traditional method.

From the rows of Pat_ata and Pat_men under the same column, we can see that the number of patterns generated by TestKompress is far less than ATALANTA. From the fault coverage in the fifth and the eighth rows, we can conclude that our EM compression has no conspicuous negative effect on fault coverage. Because the XOR-tree added for output compression is also testable, the fault coverage may even become higher, for instance, as for s38417.

It is easy to see that the number of test patterns is reduced after EM compression with ATALANTA. That means the test time also reduces to some extent. Even though more patterns are generated after EM compression with TestKompress, the test time still achieves a considerable reduction if the length of per CL_chain (i.e. the number of CK_chains) is taken into account. For s38417, the number of test patterns increases about 462/172 ≈ 2.7 times, with CIO = 10 which means the number of CK_chains in PTA structure is 1/10 of the traditional method. The ultimate test time of our PTA structure is about 2.7 × (1/10) ≈ 27% of the traditional method.

Meanwhile, we can see the saving ratios for both average (**SR_aver**) and peak (**SR_peak**) power presented in the last two rows are also higher than those shown in Fig. 7. In all, our EM-compression algorithm is beneficial to PTA architecture in test time, test data volume, test power reduction and scan I/O overhead with a little bit fault coverage loss.

6 Conclusion

In this paper, we propose a new test scheme called Parallel Test Application (PTA) approach which can effectively reduce average power, peak power and test time simultaneously. Besides, shift validity can be assured during logic test which make chain test unnecessary and fault diagnosis much easier. Furthermore, the power saving ability of PTA architecture is not strongly affected by the variation of the length of clock chains (scan I/O ports). Since no extra gates are inserted in the critical path, no significant delay is introduced in comparison to the SS method. Since ATPG is conducted before the modifications of PTA architecture, the fault coverage has no loss. A controller is employed to conduct the test in two phases. In the first phase, separate scan clock signals are activated in turn and the stimuli are applied to all DFFs of a certain clock chain concurrently. In the second phase, all DFFs capture the response of the latest assigned pattern together. To meet the special requirements of this scheme, a tri-state buffer is added after each scan cell. In addition, a compression algorithm is proposed to increase the parallelism and minimize the scan I/O overhead for PTA architecture, which further reduces power consumption and test time for PTA.

However, the use of the bus and then trying to 'broadcast' through bus to scan cells on the chain may end up creating timing issues resulting in pressure to speed up scan clocks. Meanwhile, additional area and routing resources are needed for the controller and tri-state buffers. In the future, Linear

Table 5 Performance on benchmark circuits with PTA structure and EM compression

Circuit		s13207	s15850	s38417	s35854
CP_type		3(CIO)	3(CIO)	10(CIO)	10(CI)
Traditional	Pat_ata	621	516	1191	841
	Pat_men	301	180	172	198
	Cover(%)	99.12	98.07	99.40	95.64
PTA with EM algorithm	Pat_ata	520	416	1083	790
	Pat_men	318	205	462	332
	Cover(%)	99.08	98.01	99.74	95.62
SR_aver(%)		87.69	86.17	87.04	87.52
SR_peak(%)		35.51	43.53	45.59	33.60

Feedback Shift Register (**LFSR**) technique can be applied to the controller to minimize its area and power consumption. Thus the order of loading patterns and expectations should be modified accordingly. Furthermore, modifications of every DFF will be conducted at transistor level instead of at gate level as in this paper, which can improve the timing of our PTA_DFFs, as well as reduce the area and routing overheads. Besides, more effective placement and layout-aware routing strategies need to be developed to reduce the wire length overhead. In addition, our equal-mode algorithm performs poorly in cases where some exceptional DFFs exist. We will try to develop a multi-mode compression algorithm in the future study.

References

1. Baik DH, Saluja KK, Kajihara S (2004) Random Access Scan: A solution to test power, test data volume and test time. Proc. International Conference on VLSI Design, pp 883–888
2. Basker P, Arulmurugan A (2012) Survey of low power testing of VLSI circuits. Proc. International Conference on Computer Communication and Informatics, pp 1–7
3. Bhattacharya BB, Seth SC, Zhang S (2003) Double-tree scan: A novel low-power scan-path architecture. Proc. International Test Conference, pp 470–479
4. Bushnell ML, Agrawal VD (2000) Essentials of electronic testing for digital, memory and mixed-signal VLSI circuits. Springer Science+Business Media, New York
5. Chandra A, Chakrabarty K (2001) Combining Low-Power Scan Testing and Test Data Compression for System-on-a-Chip. Proc. Design Automation Conference, pp.166–169.
6. Chandra A, Ng F, Kapur R, (2008) Low power Illinois scan architecture for simultaneous power and test data volume reduction. Proc. Design, Automation and Test in Europe Conference, pp. 462–467.
7. Chou RM, Saluja KK, Agrawal VD (1994) Power constraint scheduling of tests. Proc. International Conference on VLSI Design, pp 271–274
8. Chou RM, Saluja KK, Agrawal VD (1997) Scheduling tests for VLSI systems under power constraints. IEEE Transactions on VLSI Systems 5(2):175–185
9. Enokimoto K, Wen X, Miyase K, Huang J-L, Kajihara S, Wang L-T (2013) On guaranteeing capture safety in at-speed scan testing with broadcast-scan-based test compression. Proc. 26th International Conference on VLSI Design, pp 279–284
10. Flores P, Costa J, Neto H, Monteiro J, Marques-Silva J (1999) Assignment and reordering of incompletely specified pattern sequences targetting minimum power dissipation, Proc. 12th International Conference on VLSI Design, pp 37–41
11. Gerstendörfer S, Wunderlich H-J (1999) Minimized power consumption for scan-based BIST, Proc. International Test Conference, pp 77–84
12. Girard P, Guiller L, Landrault C, Pravossoudovitch S (1999), A test vector ordering technique for switching activity reduction during test operation, Proc. 9th Great Lakes Symposium, pp 24-27
13. Ando H (1980) Testing VLSI with random access scan, Proc. COMPCON, pp 50–52
14. Jas A, Pouya B, Touba NA (2004) Test data compression technique for embedded cores using virtual scan chains. IEEE Transactions on VLSI Systems 12:775–781
15. Le KT, Baik DH, Saluja KK (2007) Test time reduction to test for path-delay faults using enhanced random-access scan. Proc. International Conference on VLSI Design, pp 769–774
16. Mudlapur AS, Agrawal VD, Singh AD (2005) A random access scans architecture to reduce hardware overhead. Proc. International Test Conference, Paper 15.1
17. Mudlapur AS, Agrawal VD, Singh AD (2005) A novel random access scan flip-flop design. Proc. VLSI Design and Test Symp., pp 226-236
18. Muthyala SS, Touba NA (2014) Improving test compression with scan feedforward techniques. Proc. International Test Conference, pp 1–10
19. Nitin P, Sun X (2004) Design of a low-power D flip-flop for test-per-scan circuits. Proc. Canadian Conference on Electrical and Computer Engineering 2:777–780
20. Saeed SM, Sinanoglu O (2011) Expedited response compaction for scan power reduction. Proc. IEEE VLSI Test Symposium, pp. 40–45.
21. Sinanoglu O, Orailoglu A (2002) Fast and Energy-Frugal Deterministic Test Through Test Vector Correlation Exploitation. Proc. International Symposium on Defect and Fault Tolerance in VLSI Systems, pp 325-333
22. Vinay D, Chakravarty S, Pomeranz I, Reddy S (1998) Techniques for minimizing power dissipation in scan and combinational circuits during test application. IEEE Trans Comput Aided Des Integr Circuits Syst 17:1325–1333
23. Wagner KD (1983) Design for testability in the AMDAHL 580, Proc. COMPCON, pp 384–388
24. Wang S, Gupta SK (1997) ATPG for heat dissipation minimization during scan testing, Proc Design Automation Conf., pp 614–619
25. Wohl P, Waicukauski JA., Colburn JE, Sonawane M (2014) Achieving extreme scan compression for SoC Designs, Proc. International Test Conference. pp.1–8
26. Xiang D, Li KW, Sun JG, Fujiwara H (2007) Reconfigured scan forest for test application cost, test data volume and test power reduction. IEEE Trans Comput 56:557–562
27. Xiang D, Chen Z, Wang LT (2012). Scan flip-flop grouping to compress test data and compact test responses for broadside delay testing, ACM Trans. on Design Automation of Electronic Systems, Vol.17, Article No. 18.
28. Yamato Y, Wen X, Kochte MA, Miyase K, Kajihara S, Wang L-T (2011) A novel scan segmentation design method for avoiding shift timing failure in scan testing, Proc. International test conference, pp 1–8
29. Yu H, Han Y-H, Li X-W, Li H-W, Wen X-Q (2005) Compression/Scan Co-Design for Reducing Test Data Volume, Scan-in Power Dissipation and Test Application Time, In Proc. 11th Pacific Rim International Symposium on Dependable Computing. 8.
30. Yu H, Fu X, Fan X, Fujiwara H (2008) Localized random access scan: Towards low area and routing overhead, Proc. ASPDAC, pp 565-570
31. Zhang X, Roy K (2000) Power reduction in test-per-scan BIST, Proc. On-Line Testing Workshop, pp 133–138

Ding Deng joined the Institute of Microelectronics and Microprocessor in 2015 and is currently working towards his Master's degree in National University of Defense Technology, Hunan, China. His research focus is on novel architectures of scan test, BIST applications for high performance microprocessor in advanced nanometric technologies.

Yang Guo received his Ph.D. degree from National University of Defense Technology, Hunan, China in 1999. Currently he is a professor at the same university, where he leads the digital signal processor group and is the director of the Integrated Circuits. He has authored or co-authored more than 50 publications in journals and conference proceedings. His primary research interests include low power VLSI circuits, microprocessor design and verification, and electronic design automation (EDA) techniques for VLSI circuits.

Zhentao Li received the Ph.D. degree in high performance microprocessor circuit design from National University of Defense Technology, Hunan, China. He has more than 10 years of industry experience and currently is doing research in electronic design automation (EDA) techniques and test architectures for VLSI.