Time Multiplexed LBIST for in-field testing of Automotives AI Accelerators

Umair Saeed Solangi*1,2, Muhammad Ibtesam*1 and Sungju Park1, a)

Abstract Logic BIST is a safety mechanism, which performs testing for Automotive electronics. However, pseudorandom LBIST patterns results in increased test time and test power. In this letter, a novel time multiplexed LBIST is presented to overcome test related problems of AI accelerators. First, the accelerator array is divided into smaller sub arrays, which are tested on time multiplexed clock cycles. This: 1) improves overall test time, under the given test power limit, 2) allows reduction in shift power, under given test time limits and 3) since only one sub array is clocked at a time, the peak power is reduced.

key words: Artificial Intelligence (AI) accelerators, Automotives, built-in self-test (BIST), test, time multiplexed

1. Introduction

Artificial intelligence (AI) accelerators contain array of homogenous 100-10000s Processing Elements (PEs) [1]-[9]. In an AI accelerator, the stuck-at faults degrade classification accuracy for up to 74.13% with only 0.005% faulty PEs of the Accelerator [10]. A recent study has also demonstrated that classification accuracy drops from 97.4% to 7.75% with a fault rate of 0.0003% in the accelerator [11]. Thus, testing these faults becomes critical for in-field applications such as automotive. Automotive safety integrity level (ASIL)-D addresses highest level of safety standards for automatic driver assistance system (ADAS) and necessitates a safety mechanism (SM) for periodic in-field testing of automotive electronics [12]. Generally, Logic built-in self-test (BIST) is used as a SM for ASIL-D [13,14]. To achieve an acceptable fault coverage, LBIST must run for longer times to apply sufficient pseudorandom patterns. This results in increased test time. Moreover, due to lower correlation among successive pseudorandom patterns, there is increased shift power that causes excessive heat dissipation for device under test (DUT) [15]. Also, excessive peak power during the test increases the required current flow through the DUT, which may damage DUT due to electromigration [16]. Capture cycle can also contribute to peak power/current consumption. Controlling excessive peak currents can require package modification and increases overall cost of the system [17]. The main in-field test constraints for LBIST are test time and shift power. Test time is usually dependent on shift frequency, which must be reduced to not exceed power limits of the DUT. Less test time will ensure more frequent test sessions and lower shift and peak power will allow reliable test session. In addition, International Organization for Standardization (ISO) 26262 ASIL-D requires ≥ 99% coverage for the stuck-at fault testing in automotive. Increasing coverage would require longer test times and insertion of test points in the presence of random pattern resistant faults [18]. Parallel broadcasting of test patterns to PEs/ groups of PEs is used, to shorten the test time [19]-[24]. However, shift power is still a major bottleneck and affects the overall test time. In several studies, test time improvement (with shift power constraints) is achieved by time division multiplexing based pattern application to many-core DUTs [25,26,27]. In this letter, a novel time division multiplexing (TDM) based LBIST is proposed, which alleviates the test time, test shift and peak power issues for the AI accelerators. This enables ISO 26262 complied reliable and economic testing solution. Proposed method includes: 1) a novel time multiplexed based LBIST for homogenous array testing. Where the array is partitioned in compliance with test time and test power constraints. 2) a response capture and error detection circuitry for proposed LBIST. Experimental results are presented with proposed method for Google’s tensor processing unit (TPU) array and Telsa’s neural processing unit (NPU) array.

The rest of this letter is organized as follows. Section 2 presents the array partitioning method with proposed algorithm. Section 3 presents proposed time division multiplexing based testing method. Section 4 presents the experimental work and results with proposed method. Section 5 concludes the letter.

2. Partitioning of array into smaller sub arrays

Due to large number of PEs in an accelerator array,
implementing the whole array with LBIST is not an (test power and test time) efficient scalable solution for larger arrays because of two reasons. Firstly, with increasing array size, more pseudorandom test patterns will be required to attain acceptable coverage (increases test time). Secondly, ‘switching probability = 0.5’ of Linear Feedback Shift Register (LFSR) patterns is evenly distributed among homogeneous PEs (via scan chains). So, each scan chain will exponentially (because of exponential increase in number of PEs) contribute to shift power as shown in Fig. 1. These test power related issues can degrade reliability of testing and requires costly cooling mechanism [28]. Hence, the practical approach is to subdivide the whole array into smaller and identical sub arrays for application of identical test patterns. To achieve this, the primary objective is to select a sub array size for pattern sharing to achieve test time improvement. Provided, this simultaneous parallel pattern loading in sub arrays does not exceed peak power limit.

![Fig 1. Shift and peak power with array level LBIST implementation](image)

Fig. 2. shows the counter-based clock unit with shift frequency ‘Fs’ control to restrict the shift power in each sub array. Test time improvement with parallel testing is proportional to the test time of sub array unit, which is bounded by shift power. Proposed Algorithm is used to enable the partitioning, which takes peak power as the prime bound. This algorithm is used at DFT synthesis stage. Due to parallelism, total peak power is multiple of peak power of selected sub array. It is evident that parallel testing with smallest sub array will result in shortest test time. However, this increases routing congestion and routing complexity at the PRPG and increases fanout at each stage of the PRPG.

**Algorithm—Array partitioning for division of array into \(S \times S\) sub arrays in consideration with test and power limits**

1. **Input** Array size = \(A \times A = A^2\)
2. **Input** peak power budget of array \(P_{kA}\)
3. **Input** test time limit under ASIL D ‘TT’
4. **Initialize** \(n=1\)
5. \(A \times A = 2^n \times (S \times S)\)
6. Determine peak power \(P_{kA}\) for sub array
7. Determine test time \(TT_5\)
8. if \((2^n \times P_{kA} \leq P_{kA} \text{ and } TT_5 \leq TT)\)
9. do divide \(A \times A\) array into \(2^n\) \((S \times S)\) sub arrays
10. else \(n = n + 1\)
11. repeat from step 5 to 12
12. Output sub array module size \(S \times S\)
13. Output test time based on sub array \(TT_A\), \(TT_A = TT_5\)

To avoid this, the algorithm starts by selecting a larger sub array size and checks it against peak power constraint. The array partitioning algorithm starts with 2 sub arrays initially and divides further in \(2^n\) partitions with each iteration until it satisfies the constraints. This results in computational complexity of \(O(\log n)\). Since sub arrays are clocked simultaneously, the capture power will be exponential to that of 1 sub array and may exceed peak power limit of the DUT.

![Fig 2. 4 Sub array partitions; clocked with reduced shift frequency clock unit](image)

**3. Time multiplexed array testing with LBIST**

To alleviate the test power issues of parallel testing, time multiplexed testing of subarrays is proposed. An example for the time-multiplexed sub arrays or TDM partition is shown in Fig. 3. The proposed method also uses proposed Algorithm for partitioning the array into sub arrays. TDM clock circuitry contains a counter, which feeds each sub array with time shifted clock pulses. So, each sub array clocks the input test pattern and shifts the patterns at multiplexed time intervals, as shown in Fig. 3. PRPG is clocked with reduced frequency (1/4th of operational frequency in this example). Thus, in 1 clock pulse of the PRPG, all cores receive 1 clock pulse, and 1 test pattern will be loaded into each sub array (within a TDM partition). Since each core is clocked at separate time interval, the number of shift transitions in the whole array is equal to transitions in the clocked sub array only.

When BIST_START is ‘0’, all sub arrays are clocked with operational frequency ‘Fo’. When BIST_START is ‘1’, all sub arrays are clocked with time-multiplexed clock with shift frequency ‘Fs’. Replacing BIST_START signal with TEST_MODE signal in the proposed method will allow capture cycle to be applied in time-multiplexed manner alike the shift cycle. This limits the capture power to only 1 sub array unit unlike the broadcast method, where the capture power is multiple of number of sub arrays due to simultaneous captures. The number of sub arrays in 1 TDM partition is determined by the following equation;
$K = \frac{\text{Operational Frequency (Fo)}}{\text{Test Shift Frequency (Fs)}}$  \hspace{1cm} (1)

The sub arrays grouped for time-multiplexing is limited to ‘$K$’, which is dependent on operational frequency ‘$F_o$’.

Since 1 sub array is tested per clock, the peak shift and capture power reduces to $\frac{1}{K}$. For ‘$n$’ $K$, $S=A\times A$’ (here ‘$S$’ is sub array size), patterns are simultaneously applied to ‘$n$’ groups. In this case, peak shift and capture power reduces with the factor of ‘$\frac{n}{K}$’. Fig 4. shows counter enabled TDM clock control for 4 sub arrays. Size of the ‘sub array counter’ is proportional to the number of sub arrays arranged in TDM partition. Clock tree modification needed to enable TDM based LBIST pattern application will be same as clock tree of non-TDM based LBIST (with reduced shift frequency). Performance overhead in the clock source of the TDM LBIST consists of an AND gate, which is same as for array partitioned based LBIST (as shown in Fig. 2). Existing clock buffers of the clock tree are replaced by AND gates to achieve this [29,30].

Partitioning the array into smaller homogenous modules require smaller LFSR and Multiple Input Signature Register (MISR). This may increase the aliasing error, as because this error is dependent on the size of LFSR [31]. To alleviate this aliasing issue that can degrade the coverage in the proposed method, a comparator-based response capturing register is proposed as shown in Fig. 5. This register captures the responses at the end of the shift operation. This per clock response comparison increases the probability of error detection in the proposed LBIST. Captured responses from scan chains of PEs i.e., $SC_1$, $SC_2$, …, $SC_n$ are shifted out serially from each adjacent scan chain into the response capture register. This enables per clock logic comparison among test responses, resulting in faster error detection [32]. To capture the responses, response capture register is clocked by negative edge pulse of the 4th sub array clock. This captured response is checked concurrently by the combinational logic. Mismatch among the captured response is detected by the status of the flag. Test pass/fail flag is ‘1’ in case of mismatch and ‘0’ for matched response. This flag will be checked and updated at each shift out clock.

4. Experimental results

BIST in-field testing is executed on the start-up and during Idle mode i.e., in off-line mode. This will ensure BIST sessions to have no functional impact on the accelerator’s performance. Experimental evaluation of the proposed method is performed on two accelerators; i) Google’s TPU; because of its significance in academia [10,11] ii) Tesla’s NPU; because it is specifically used in Full Self-Driving (FSD) Chip for autonomous driving applications. For power analysis, *Synopsys Primepower* is used to generate event-based power data i.e., value change dump (VCD) files of gate-level netlists. Gate-level netlists were synthesized for verilog models of NPU [1] and TPU [2]-[6] on *SAED32nm* library with *Synopsys Design Compiler*. TPU has a systolic array with 8-bit datapaths and 16-bit sum register (per PE). Whereas, NPU contains a non-systolic array with broadcasted 8-bit activation and weight inputs and 32-bit sum register (per PE). For DFT synthesis, one scan chain is inserted per PE for both array types for every size. For array partitioning of the given array module, a 4×4 sub
array module is used (in conjunction with Algorithm). A 16-bit LFSR is used as PRPG. Input of each scan chain is driven by Phase Shifter outputs. For our experiments, Operational Frequency \( F_o \) = 1 GHz and Shift Frequency \( F_s \) = 50 MHz. The area overhead of TDM BIST clock unit with the given configuration for a 8×8 array is 0.31% for TPU and is 0.14% for NPU, which decreases with increasing array size. The area overhead of response capture register for 4×4 sub array is 0.86% for TPU and is 0.4% for NPU.

LFSR patterns were applied to TPU to achieve coverage \( \geq 99% \) in compliance with ASIL D. Test time to achieve this coverage is given in Fig. 7. Due to absence of random pattern resistant faults, there were no test points inserted. However for NPU, observation test points were required to increase the coverage \( \geq 99\% \) as per ASIL D. Observation test points were inserted on a 4×4 sub array with Synopsys DFT Compiler with an area overhead of 6.96%. Test time required for LFSR patterns to achieve \( \geq 99\% \) in the presence of test points is given in Fig. 8. With test time as the primary constraint, the array is subjected to pattern application to all 4×4 sub arrays simultaneously. It is assumed that for the given array, the sub arrays tested simultaneously does not exceed power limits. For the given test time, TDM method achieves peak power reduction by limiting peak shift and capture power to only one sub array, as shown in Fig. 6. So, for the given test time, the in-field test with the proposed method is more reliable than parallel test method. This lowers the cost for package modifications required to handle large currents.

For restraining peak and shift power during in-field, the parallel method will limit the simultaneous testing to a limited cluster of sub arrays. And for parallel testing, multiple sessions will be needed to test the whole array. In proposed method, peak power in 1 TDM partition (which contains multiple sub arrays as given in Eq.1) is limited to 1 sub array. For the whole array, the peak power = sub array peak power × no. of TDM partitions. So, with the given shift frequency and peak power limits (set for the parallel test method), the proposed TDM method yields test time improvement. Which is proportional to ‘total number of TDM partitions’ over the parallel broadcast method (with multiple test sessions). For experiments, arbitrary number of test sessions i.e. ‘2’ is set, for parallel testing. Also, another scenario is assumed with fixed number (arbitrary) of 32 PEs per session for parallel test limit. The results shown in Fig. 7 and Fig. 8, corroborate the efficiency of proposed method in terms of test time in both scenarios. Reduction in test time enables more frequent in-field test sessions for the accelerators.

Parallel test method increases the shift power with increasing array size (shown in Fig. 9). So, larger arrays dissipate more heat and will require cooling mechanism, this may increase overall cost. Whereas the proposed method yields a proportional (’K’ times) reduction in shift power, as shown in Fig. 9. Consequently, with proposed method a ‘K’ times increase in test time can also be achieved (in comparison with parallel test) by increasing the shift frequency while not exceeding the given shift power limit (which are set for parallel test method).

![Fig. 7. For TPU; Test time (ns) improvement with a) 2 test sessions b) 32 PEs/session](image7.png)

![Fig. 8. For NPU; Test time (ns) improvement with a) 2 test sessions b) 32 PEs/session](image8.png)

![Fig. 9. Shift power (W) reduction with proposed LBIST](image9.png)

5. Conclusion

This letter presents a novel time multiplexed method for LBIST to enable ISO 26262 complied testing of automotive AI accelerators. Partitioning the array into identical sub arrays and clocking these sub arrays with time-multiplexed clocks allows test time improvement. This results in power efficient testing for accelerators. Experimental results with commercial accelerators validate the efficacy of proposed method.
Acknowledgments

This work was supported in part by the Higher Education Commission, Govt. of Pakistan, under the scholarship program titled "Faculty Development of UESTPs/UETs". This research was also supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

[1] P. J. Bannon: U.S. Patent 0 026 078 A1 (2017).
[2] N. P. Jouppi et al., “In-Datapath Performance Analysis of a Tensor Processing Unit,” 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), Toronto, ON, 2017, pp. 1-12, (DOI:10.1145/3079856.3080246.)
[3] J. Ross, N. Jouppi, A. Phelps, R. Young, T. Norrie, G. Thorson, and D. Liu: U.S. Patent 9747546 B2 (2015).
[4] J. Ross and A. Phelps: U.S. Patent 9697463 B2 (2015).
[5] J. Ross: U.S. Patent 9805304 B2 (2015).
[6] J. Ross and G. Thorson: U.S. Patent 9747548 B2 (2015).
[7] Y. Chen, et al., “Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices,” in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 2, pp. 292-308, June 2019, (DOI:10.1109/JETCAS.2019.2910232.)
[8] Tianshi Chen et al., “DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning”, SIGARCH Comput. Archit. News 42, 1 (2014), 269–284. (DOI:10.1145/2654822.2541967)
[9] L. Cavigelli, et al., “Origami: A convolutional network accelerator”, Proc. 25th Ed. Great Lakes Sym. VLSI, pp. 199-204, 2015. (DOI:10.1145/2742060.2743766)
[10] J. J. Zhang et al., “Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator,” in Proc. IEEE 36th VLSI Test Symp. (VTS), 2018, pp. 1-6. (DOI:10.1109/VTS.2018.8368656.)
[11] S. Kundu et al., "Toward Functional Safety of Systolic Array-Based Deep Learning Hardware Accelerators," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 29, no. 3, pp. 485-498, 2021. (DOI:10.1109/TVLSI.2020.3048829.)
[12] C. Hobbs and P. Lee, “Understanding ISO 26262 ASILs,” Electron. Des., Jul. 2013. [Online]. Available: http://electronicdesign.com
[13] S. R. Das, "Getting errors to catch themselves - self-testing of VLSI circuits with built-in hardware," IEEE Trans. on Instrum. Meas., vol. 54, no. 3, pp. 941-955, 2005. (DOI:10.1109/TIM.2005.847352.)
[14] Y. Liu et al., "Deterministic Stellar BIST for In-System Automotive Test," IEEE Int. Test Conf. (ITC), 2018, pp. 1-9. (DOI:10.1109/TEST.2018.8624872.)
[15] Weizheng Wang, Peng Liu, Shuo Cai, Lingyun Xiang, “Low power logic BIST with high test effectiveness” IEICE Electronics Express, vol. 10, no. 23, pp. 1-6, 2013. (DOI:10.1587/elex.2013.0855)
[16] D. Gizopoulos et al., "Power Aware Testing and Test Strategies for Low Power Devices," Design Autom. and Test in Europe, 2008, pp. xliv-xlvi. (DOI:10.1109/DATVE.2008.4846442)
[17] P. Girard et al., "Circuit partitioning for low power BIST design with minimized peak power consumption," in Proc. 8th Asian Test Symp. (ATS'99), 1999, pp. 89-94. (DOI:10.1109/ATS.1999.810734.)
[18] N. Mukherjee et al., "Time and Area Optimized Testing of Automotive ICs," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 29, no. 1, pp. 76-88, 2021. (DOI:10.1109/TVLSI.2020.3025138.)
[19] H. Ma et al., “A case study of testing strategy for AI SoC,” in Proc. IEEE Int. Test Conf. Asia (ITC-Asia), Tokyo, Japan, 2019, pp. 61-66. (DOI: 10.1109/ITC-Asia.2019.00024.)

IEICE Electronics Express, Vol.xx, No.xx, xx-xx