[Article] Programmable built-in self-testing of embedded RAM clusters in system-on-chip architectures

Original Citation:

Availability:
This version is available at: http://porto.polito.it/1404469/ since: October 2006

Publisher:
IEEE

Published version:
DOI:10.1109/MCOM.2003.1232242

Terms of use:
This article is made available under terms and conditions applicable to Open Access Policy Article ("Public - All rights reserved"), as described at http://porto.polito.it/terms_and_conditions.html

Porto, the institutional repository of the Politecnico di Torino, is provided by the University Library and the IT-Services. The aim is to enable open access to all the world. Please share with us how this access benefits you. Your story matters.

(Article begins on next page)
Programmable Built-In Self-Testing of Embedded RAM Clusters in System-on-Chip Architectures

Alfredo Benso, Stefano Di Carlo, Giorgio Di Natale, and Paolo Prinetto, Politecnico di Torino
Monica Lobetti Bodoni, Siemens Mobile Communications

ABSTRACT

Multiport memories are widely used as embedded cores in all communication system-on-chip devices. Due to their high complexity and very low accessibility, built-in self-test (BIST) is the most common solution implemented to test the different memories embedded in the system. This article presents a programmable BIST architecture based on a single microprogrammable BIST processor and a set of memory wrappers designed to simplify the test of a system containing a large number of distributed multiport memories of different sizes (number of hits, number of words), access protocols (asynchronous, synchronous), and timing.

INTRODUCTION

Silicon area is now so cheap and integration technologies so advanced that industries can embed in a single chip, usually referred to as system-on-chip (SoC), all the components and functions that historically were placed on a hardware board. Each component or function is now available as a predesigned complex functional block, or embedded core.

Embedded memories are the most dense components within an SoC, accounting for up to 90 percent of its real estate. Today's technologies allow the design and manufacturing of memory cores with many I/O ports, and multiport RAM core generators are commonly available in many application-specific integrated circuit (ASIC) vendors' libraries (e.g., LSI-Logic, Texas Instruments, and ST Microelectronics). To get an idea of today's SoC complexity, it is enough to consider that typically more than 30 embedded memories are placed on a single chip; they are scattered around the device rather than concentrated in one location; they all have different types, sizes, and access protocols and timing; and they can even be doubly embedded inside embedded cores. From a testability point of view, memories are also the most sensitive to process defects, making it essential to thoroughly test them in the SoCs.

This new design philosophy, based on the use of embedded cores, leads to a radical change in the test engineering process. First of all, direct accessibility to interconnections and cores' boundaries is not possible; however, test patterns and test responses still need to be delivered to the core or the SoC boundaries.

In the case of memory cores, the test methodology of choice is built-in self-test (BIST). BIST offers a simple low-cost means to test for failures of embedded memories without significantly impacting device performance. In this scenario, the implementation of an efficient BIST strategy for SoCs including several multiport RAMs requires taking into account the different sizes (number of bits, number of words), access protocols (asynchronous, synchronous), and timing of the memories embedded in the system, to minimize the BIST area and routing overhead and fulfill power budget constraints. Moreover, while it has been used primarily for production pass/fail testing, BIST should be extended to provide the diagnostic data required for process monitoring and repair. A successful BIST for embedded memories has to guarantee core accessibility, scalability, in-system programmability (ISP), low overhead, and flexibility in the test scheduling.

This article presents the efforts and the results obtained in designing a proprietary BIST architecture, to tackle the above-mentioned set of problems.

The article is organized as follows. We summarize some of the most significant memory BIST architectures presented in literature; we give a general overview of the proposed BIST architectures and then detail the structure of the main blocks of the architecture. The scheduling
and diagnosis facilities of the proposed BIST are detailed, and a possible optimization is discussed. We present a real application of the proposed approach on an industrial case study, and finally we summarize the main contributions of the work and conclude the article.

STATE OF THE ART

Several memory BIST solutions have been proposed to test both single- and multipORT memories [1, 2], and static and dynamic memories [3, 4]. Programmable memory BIST has been proposed in [5-7] to increase flexibility in applying different combinations of test patterns targeting different types of faults. Despite their effectiveness, all these solutions are designed to address the problem of testing a single type of memory, and none focuses on the problem of concurrently testing several heterogeneous memory arrays. This problem has been addressed in [8, 9], where the authors propose a built-in self-diagnostic method to simultaneously diagnose spatially distributed memory modules with different sizes. The approach is based on the serial interfacing technique proposed in [10]. The basic idea is to synthesize the I/O port of each buffer as a scan chain from which the test patterns can be provided and memory contents can be read. The solution is very easy to implement, but it is not so efficient in terms of test speed and area overhead, and does not take into account power consumption constraints. Moreover, all the memory tested in parallel must be of the same type. A deterministic BIST state machine was designed in [11] to test multiple RAMs with different characteristics. Although all the memory modules are tested (truly) concurrently, each memory module receives its own control signals from the BIST controller. This solution has the disadvantages of large routing area overhead and a complex design of the BIST controller.

MEMORY BIST

MANAGEMENT ARCHITECTURE

The goal of this article is the design of a proprietary BIST scheme to tackle the problem of testing the memory subsystem of a complex SoC. Figure 1 gives an overview of the proposed BIST architecture.

A single BIST processor is in charge of performing the test of all (or a subset of) the memories of the system. Using a minimal set of communication signals, the BIST processor coordinates, executes, and synchronizes the test algorithm of the memories under test. The BIST processor is μ-programmable: the test algorithm is stored as a sequence of elementary test primitives in a dedicated memory (μ-program memory); these instructions include (but are not limited to) update of the address generators, application of a test pattern, and comparison of a memory cell with an expected value. This solution allows, if necessary, programming the system at runtime to execute any required test algorithm. The BIST processor functionalities and communication protocol are independent of the number and characteristics of the memories embedded in the system.

The different test primitives that constitute the test algorithm are received by a wrapper placed around each memory. In particular, each wrapper is composed of a set of port wrappers (one per each memory port) and a dispatcher. The port wrapper contains the standard blocks required to implement BIST capabilities (i.e., an
address generator, a pattern generator, and a comparator), and an interface block designed to translate the communication received from the BIST processor into the particular memory access protocol. The dispatcher is in charge of collecting the test commands from the BIST processor, and delivering them to the different port wrappers.

Finally, two scan chains are used to connect all the port wrappers to allow the scheduling of the memory under test and full diagnosis capability.

The proposed scheme guarantees the following goals:

Core accessibility: The test of each memory is controllable and observable using a minimal set of communication signals, including only two scan chains connecting all the port wrappers.

Scalability: The BIST processor implementation is independent of the number and characteristics of the memories embedded in the system.

In-system programmability (ISP): Implementing the BIST processor as a µ-programmable machine provides the test engineer a flexible and reusable block that can be used to manage the BIST of any number of memories of any size, and is independent from the test algorithm, which can be dynamically downloaded into the BIST processor itself.

Low overhead: Using a single BIST processor and a minimum set of communication signals allows minimizing, with respect to a traditional BIST solution where each memory has a dedicated BIST controller, the area overhead and connectivity around each RAM.

Flexible test scheduling: The set of memories to be tested can be freely selected by the test engineer, using either test primitives stored in the test program or a dedicated scan chain to properly set a status bit in each memory. Moreover, the proposed solution allows concurrently running the BIST of a set of memories with different number of ports, sizes, access protocols, and timing.

The following sections will further detail the blocks composing the architecture.

THE BIST PROCESSOR

The proposed memory BIST is based on a single BIST processor used to test all the memories of the target SoC. To increase flexibility, BIST execution is based on a µ-programmable approach. Due to their regular structure, the most popular and widely accepted deterministic test algorithms for memory BIST are March tests. A March test is a finite sequence of operations (March elements) applied to each memory cell in the memory array in either ascending or descending order before proceeding to the next cell [1]. March tests are popular because of their low temporal complexity, regular structure, and their ability to detect different types of faults. The proposed BIST processor has therefore been optimized to implement March tests. The chosen algorithm is stored in a dedicated µ-program memory, coded through a set of test primitives. The µ-program memory can be either a ROM or an ISP device. In the former case, the test program is fixed at design time, whereas in the latter any custom test algorithm can be downloaded into the µ-program memory at test time.

After selecting the set of memories under test, the BIST processor reads from the µ-program memory one test primitive at a time, forwards it to all the wrappers of the memories under test, and waits until its completion by all the target memories.

When the test program is completed (i.e., all the test primitives have been applied), the BIST processor reads the test results from each memory. If a fault is detected, the faulty memories can be located by resorting to a set of diagnosis capabilities.

The architecture of the BIST processor and the µ-program memory are strongly influenced by the peculiar characteristics of multiport memories. In fact, due to the possibility of concurrently accessing several cells, new fault models must be targeted [12], and ad hoc March algorithms must be adopted to cover these new types of fault. In particular, the proposed implementation is optimized to implement March algorithms for multiport memories presented in [13]. The main characteristic of these algorithms is the use of nested cycles to access the different memory ports,

\[
(p_A, p_B, p_C) = ((A \text{ to } B), (B \text{ to } C), (C \text{ to } B))
\]

where \(1 \leq B < B + 1 \leq n\) denotes a nested addressing sequence in which cell \(B\) goes from \(0\) to \(A - 1\) and, for each value of \(B\), cell \(C\) goes from \(B + 1\) to \(n\). A pseudo-C code of this nested addressing sequence would correspond to two nested for cycles:

\[
\text{for (} B = 0; B < A - 1; B++)
\]
\[
\text{for (} C = B + 1; B < n; C++)
\]

As previously explained, each step of the test program is coded in the µ-program memory as a sequence of test primitives, one for each memory port. The set of test primitives needed to implement the proposed family of March algorithms are:

- **W0**: Write pattern
- **W1**: Write not (pattern)
- **R0**: Read and verify a pattern
- **R1**: Read and verify a not (pattern)
- **INC**: Increment the address generator and define the end of a March element
- **DEC**: Decrement the address generator and define the end of a March element
- **INCCOND**: Conditionally increment the address generator
- **DECCOND**: Conditionally decrement the address generator
- **LOAD**: Load a value in the address generator
- **NME**: New March element
- **NEXTP**: Next pattern

\[ x \text{ or } x \]

\[ x \text{ or } x \]
The external interface of the BIST processor can be designed in order to match the target system requirements. Possible solutions are a P1500 compliant interface, an addressable device on the system bus, or a JTAG interface, as in the case study presented later.

**The Memory Wrapper**

The wrapper placed around each memory has to execute the test primitives broadcast by the BIST processor regardless of the particular memory access protocol. The wrapper is therefore the only element in the architecture taking care of the number of ports, the size, and the access protocol of the memory it wraps.

The wrapper generates the correct test patterns and memory addresses required to execute the received test primitives, and compares the values read during the test with the expected ones.

The wrapper architecture consists of a dispatcher and a set of port wrappers.

**Dispatcher**

Each RAM under test has a dedicated dispatcher, which receives the test primitives for all the port wrappers from the BIST processor. Since the primitives are sent sequentially but must be applied at the same time in order to execute the required operations concurrently on all the ports of the memory, each dispatcher saves all the primitives in a temporary register and delivers them to each port wrapper only after receiving a synchronization test primitive (RUN). This solution allows a dramatic reduction of the routing overhead that would be required to send all the primitives in parallel using a dedicated bus for each port.

**Port Wrapper**

Each memory port has a dedicated port wrapper that generates the test patterns (address and data) and verifies the correct behavior of the memory according to the primitive received from the dispatcher. The result of each primitive is signaled on an output line.

The internal structure of a port wrapper is drawn in Fig. 2. The address generator (AG) is in charge of generating the correct address where the test pattern, provided by the pattern generator (PG), has to be written or verified. PGs can easily be customized in order to target different fault types [13]. Its implementation is nevertheless always very simple, and never more complex than an up counter. The correctness of the content of a memory cell is evaluated using a simple comparator.

Two status bits are used to set the memory in transparent or test mode (the mode status bit) and to store the test results at the end of the BIST algorithm (the result status bit). All the memories set in test mode are tested in parallel, whereas those set in transparent mode are bypassed and not tested; this feature is required to allow flexible scheduling of the memories under test. To set and read them, the status bits of all the port wrappers are dynamically connected in a global scan chain.

Finally, each port wrapper includes an interfacing block able to receive the test primitives (command) from the dispatcher and execute them on the memory using the required protocol. Moreover, the interfacing block receives a synchronization signal (Sync_IN) from the previous port wrapper, and produces an output synchronization signal (Sync_OUT) needed by the other wrappers and the BIST processor to synchronize the scheduling of the next test primitive.

The Sync_IN signal of each port wrapper is directly connected to the Sync_OUT signal of the previous one, except for the last port wrapper whose Sync_OUT signal is connected to the BIST processor. The Sync_OUT signal is enabled only when the Sync_OUT signal of the previous port wrapper is asserted. Therefore, the BIST processor receives the logic-AND of the output signals generated by all the port wrappers.

From a functional point of view, Sync_OUT assumes different meanings depending on the received test primitive. As an example, for a read or write operation, it has the meaning of end of instruction (EOIN). It is asserted when the memory actually ends the execution of the command. This mechanism guarantees the synchronization among memories with different timing and access protocols. For a primitive to increment or decrement the value of the address generator, Sync_OUT has the meaning of end of address (EOAD). It is asserted when the addressing space has been visited by the address generator, allowing the synchronization among memories of different sizes.

Two types of port wrappers are available: one for the first port of each memory and one for the other ports. The main difference between
In order to minimize the routing overhead, the signals exchanged between the BIST processor and the memory wrappers (command signals, synchronization signal, scan chain signals) are multiplexed. In particular, these signals are multiplexed at the port wrapper level.

Figure 3. Scheduling using the Conf primitive.

The two lies in the fact that the port wrapper connected to the first port of the memory implements the main addressing loop of the March test family discussed earlier, whereas the addresses applied to the memory by port wrappers connected to the remaining ports are relative to the value of the address generated by the previous port wrapper.

In order to minimize the routing overhead, the signals exchanged between the BIST processor and the memory wrappers (command signals, synchronization signal, scan chain signals) are multiplexed. In particular, these signals are multiplexed at the port wrapper level. All the information is routed using only six signals (four command signals and two synchronization signals).

TEST SCHEDULING
An important issue to be faced when running concurrently the BIST of several modules is fulfilling power budget constraints. In fact, BIST typically results in a circuit activation rate higher than the normal one, and overdissipation of power may seriously damage the devices. Moreover, the variety of memories that can be found in a complex architecture may require different test algorithms. To address these two issues, the proposed approach implements a very flexible scheduling mechanism. In particular, it is possible to select the set of memories to be tested using either a dedicated test primitive as part of the test algorithm or setting the mode status bit flag into the memory wrapper through a scan chain. Only the wrappers of the selected memories will execute the test primitives received from the BIST processor; all the others will be set in transparent mode and therefore bypassed. In this way, several test algorithms may be stored in the μ-program memory and may be applied sequentially to different sets of memories. The definition of algorithms or guidelines for selection of the best scheduling is a task that depends on the particular target system and is therefore outside the scope of this article. Our main focus is on the design of an architecture that allows flexible definition of test scheduling. The two mechanisms implemented to allow the scheduling of the memories under test are briefly explained in the following.

SCHEDULING USING THE CONF PRIMITIVE
Using the Conf primitive, it is possible to embed scheduling information into the test program. The representation of this primitive in the μ-program memory is defined as follows:
- The Conf opcode.
- The number of 4-bit words used to code the ActivationMask.
- The ActivationMask, a mask of bits where each bit corresponds to one memory in the system. To include a memory in the set of SRAMs under test, the corresponding bit in the ActivationMask has to be set.

As an example, let's consider the system in Fig. 3. When the BIST processor reaches a Conf primitive during the test program execution, it reads the ActivationMask and configures all the memory wrappers using the scan chain defined earlier in order to activate the required scheduling plan. The first ActivationMask shown in Fig. 3 sets RAM1 and RAM4 under test, whereas the second one sets RAM2 and RAM3 under test.
In order to define different test sessions and collect test results, at the end of each algorithm the BIST processor stops the test program execution and waits for a new start primitive to continue with the next one.

**SCHEDULING USING THE SCAN CHAIN OPTION**

In order to give the designer greater flexibility, the set of memories under test can also be set loading the appropriate ActivationMask directly from the outside using a scan chain protocol. In order to jump to the appropriate test algorithm in the program memory, the starting value of the program memory Address Register can also be loaded in the BIST processor using the same protocol.

**DIAGNOSIS**

Fail map extraction is required to output the relevant data needed to determine why a failure occurred within a memory. This data is post-processed using diagnostic software to isolate the defective memory and location within the memory. Therefore, when a faulty memory is detected, the proposed approach allows collection of diagnostic information about the location of the faulty memories, the ports where the fault has been detected, the addresses of the faulty cells, and the detecting patterns. This information is stored into the result status bit, address generator, and background pattern generator of each port wrapper and can be scanned out via the Results_Scan_Chain. To allow even more detailed diagnostic capabilities, it is also possible to include in the Result_Scan_Chain the test primitive that triggered the detection of the fault. To reduce the scan chain length, depending on the result of the test (Result_Status_Bit), each port wrapper configures its portion of the Results_Scan_Chain in one of the following two ways (Fig. 4):

- **Result_Status_Bit=1**: The memory is not faulty; only the Result_Status_Bit is placed on the scan chain.
- **Result_Status_Bit=0**: The memory is faulty; the Result_Status_Bit is chained to the content of the address generator and the background pattern generator.

**FURTHER OPTIMIZATION**

To further reduce the BIST area overhead, the designer can share a single wrapper for a cluster of identical memories (same type, width, and size) to be tested in parallel.

This optimization is made at the port wrapper level. For each port wrapper only one address generator and one background pattern generator are needed. The only difference from the previously described port wrapper structure is that a shared port wrapper contains a pair of status bits and a comparator for each memory. In this way, when a fault is detected, the result status bit of the faulty memory is set, the memory is disconnected, and the wrapper keeps on testing the remaining memories of the cluster. Obviously, in this case the status of the address generator and pattern generator of the faulty memory are not preserved. To collect diagnostic information, the test must be reexecuted on the faulty memory only by properly setting its mode status bit.

Finally, since a fault in the BIST logic can be detected only if it causes an error that is detectable as a memory fault by the test algorithm, the stuck-at fault coverage cannot be precisely computed a priori and, anyway, will be quite low. Therefore, to allow high fault coverage at the end of production, the BIST logic can be synthesized and tested using full scan.

**CASE STUDY**

A case study has been used to evaluate the proposed approach and gather experimental results. The target circuit, VC12AD, is part of a telecommunications ASIC designed by Italtel SpA. Both Italtel SpA and Siemens ICN have also used the same circuit as a benchmark for the evaluation of commercial BIST insertion tools. The target circuit has been described in VHDL and synthesized using the G10 LSILogic™ library, which provides a set of RAMs of different sizes.

The VC12AD counts up to 860,000 equivalent gates (excluding RAMs), plus 36 small-sized RAMs, for a total of 14,704 bits and 380,503 equivalent gates.

The case study aims at evaluating the BIST architecture complexity when applied to a set of memories with very different characteristics, and the area overhead after the BIST insertion.

The 36 RAMs of the circuit are grouped into four distinct macro areas whose characteristics are listed in Fig. 5.

**BIST ARCHITECTURE**

In the definition of the BIST architecture, we tried to minimize the number of wrappers resorting, whenever possible, to clusters of memories (described earlier). As a consequence:
The proposed memory BIST architecture deals with memory modules only. If additional modules (e.g., random logic, legacy cores) have to be BISTed as well, more complex and sophisticated approaches will have to be adopted.

**Figure 5. VC12AD BIST architecture.**

- Within C12A, the two modules tpa21x8 and the two modules spa21x26 are treated as two clusters.
- Within C12D, the two modules spa21x34 and the two modules spa* are treated as two clusters.
- Within SYNDES, the memories are organized in four clusters of seven, seven, six, and one elements, respectively.

The memory clustering has been strongly influenced by the actual floor plan: for example, the three spa21x34 memories (two located inside C12D and one in PDH_INT) are too far apart to be included in a single cluster.

The overall VC12AD structure after BIST insertion is shown in Fig. 5.

**BIST SCHEDULING**

Due to the different characteristics of the VC12AD memories (read/write ports, read-only ports, and write-only ports are present), it is not possible to adopt a single March algorithm for all of them. We therefore organized the BIST in four sessions, each executing a different March algorithm:

- **Session 1:** All the single-port RAMs are tested concurrently.
- **Session 2:** All the dual-port RAMs are tested concurrently.
- **Session 3:** All the triple-port RAMs are tested concurrently.
- **Session 4:** All the quadruple-port RAMS are tested concurrently.

**EXPERIMENTAL RESULTS**

The total area overhead introduced by the port wrappers is 68,177 equivalent gates. This area is not proportional to the number of memory ports, but depends more on the port sizes and functionalities.

The BIST processor and μ-program memory area overhead (5431 and 4459 equivalent gates, respectively) are a fixed contribution and are not influenced by the number of memories present in the system.

The total area overhead is, in this case study, 17.02 percent. Although this result may seem quite high, it is necessary to consider that the target circuit has a lot of small memories, and therefore the overhead introduced by the wrapper is significant. With larger memories the overhead would be much lower.

The area overhead introduced by a commercial BIST insertion tool is 22.5 percent.

**CONCLUSIONS**

In this article we present a proprietary solution for a particular industrial scenario in which it is necessary to define the BIST strategy of a complex communication SoC, including several multiport memories of different sizes, access protocols, and timing. The proposed architecture consists of a single BIST processor, implemented as a μ-programmable machine and able to execute different test algorithms, a wrapper for each memory (or cluster of memories), each wrapper including one port wrapper for each memory port and a special block named dispatcher. Each port wrapper contains standard memory BIST modules and an interface block to manage the communications between the memory and the BIST processor. The dispatcher collects the
instructions from the test processor and delivers them to the port wrappers. The proposed scheme presents several advantages. It allows running concurrently the BIST of a set of memories of different number of ports, sizes, and access protocols, minimizing the BIST area overhead and connectivity around each memory. In addition, the set of memories to be tested can be freely selected by the designer, as well as the test algorithm to be executed on each set.

The proposed memory BIST architecture deals with memory modules only. If additional modules (e.g., random logic, legacy cores) have to be BISTed as well, more complex and sophisticated approaches will have to be adopted.

ACKNOWLEDGMENTS
This work is partially supported by Istituto Superiore per le ICT Mario Boella under contract Test DOC: Quality and Reliability of Complex SoC.

REFERENCES

BIOGRAPHIES
ALFREDO BENIO [M] received his M.S. degree in computer engineering (1995) and Ph.D. (1998) from the Politecnico di Torino, Italy. He is currently a researcher at the same university, where his research interests include design-for-testability techniques, BIST for complex digital systems, and software-implemented hardware fault tolerance (SWIFT). He is the chair of the IEEE Computer Society Test Technology Technical Council (TTTC) Web-Based Activities Group.

PAOLO PRINETTO [M] received an M.S. in electronic engineering in 1976 from the Politecnico di Torino, Italy. Since 1980 he has been a full professor of computer engineering at the same university, and since 1998 joint professor at the University of Illinois at Chicago. His research interests cover testing, test generation, BIST, and dependability. He is the Golden Core Member of the IEEE Computer Society and elected chair of the IEEE Computer Society TTC.

STEFANO DI CARLO [M] (stefano.dicarlo@polito.it) is a research assistant in the Department of Automation and Information Technology at Politecnico di Torino. His research interests include DFT techniques, SoC testing, BIST, and FPGA testing. He has an M.S. in computer engineering and a Ph.D. in information technologies, both from Politecnico di Torino. He chairs the IEEE Computer Society TTC Electronic Submission Committee.

GIORGIO DI NATALE [M] is a research assistant in the Department of Automation and Information Technology at Politecnico di Torino. His research interests include DFT techniques, SoC testing, BIST, and FPGA testing. He has an M.S. in computer engineering and a Ph.D. in information technologies, both from Politecnico di Torino. He is associate Webmaster of the IEEE Computer Society TTC.

MONICA LODetti [B] (lodetti@polito.it) is a DFT and test engineer at Siemens Mobile Communications. Her research interests include ASIC and PCB test, with an emphasis on defining a test methodology and providing the tools, training, and support required for the chosen strategy's application. She has an M.S. in nuclear engineering and a Master's in information technology, both from Politecnico di Milano. She is a member of the Board Test Action Group for the International Test Conference's Test Week, and General Chair of both BTW '02 and BTW '03.