MISSOURI S&T

Missouri University of Science and Technology

Scholars' Mine

Electrical and Computer Engineering Faculty Research & Creative Works

Electrical and Computer Engineering

01 May 2005

Spare Line Borrowing Technique for Distributed Memory Cores in SoC

B. Jang

Minsu Choi
Missouri University of Science and Technology, choim@mst.edu

Nohpill Park

Vincenzo Piuri

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/ele_comeng_facwork/1884

Follow this and additional works at: https://scholarsmine.mst.edu/ele_comeng_facwork

Part of the Electrical and Computer Engineering Commons

Recommended Citation

Spare Line Borrowing Technique for Distributed Memory Cores in SoC

B. Jang[1], M. Choi[2], N. Park[3], Y.B. Kim[4], V. Piuri[5], and F. Lombardi[4]

[1]Mobile Platform Lab., Digital Media R&D Center, Samsung Electronics
416 Maetan-3Dong, Suwon-City 443-742 South Korea
[2]Dept of ECE, University of Missouri-Rolla
134 Emerson Electric Co. Hall, Rolla, MO 65409-0040
[3]Dept of CS, Oklahoma State University
219 Math Building, Stillwater, OK 74078 USA
[4]Dept of ECE, Northeastern University
{327, 309} Dana Research Center, Boston, MA 02115 USA
[5]Department of Information Technologies, University of Milan
Via Bramante 65, 26013 Crema (CR), Italy

Abstract - *In this paper, a new architecture of distributed embedded memory cores for SoC is proposed and an effective memory repair method by using the proposed Spare Line Borrowing (software-driven reconfiguration) technique is investigated. It is known that faulty cells in memory core show spatial locality, also known as fault clustering. This physical phenomenon tends to occur more often as deep submicron technology advances due to defects that span multiple circuit elements and sophisticated circuit design. The combination of new architecture & repair method proposed in this paper ensures fault tolerance enhancement in SoC, especially in case of fault clustering. This fault tolerance enhancement is obtained through optimal redundancy utilization: Spare redundancy in a fault-resistant memory core is used to fix the fault in a fault-prone memory core. The effect of Spare Line Borrowing technique on the reliability of distributed memory cores is analyzed through modeling and extensive parametric simulation.*

Keywords – *Memory Repair, System on a Chip (SoC), Reconfiguration*

I. INTRODUCTION

System on a Chip (SoC) is driving the VLSI (Very Large Scale Integration) industry today. Chip designers can implement everything from processors, memories, and other logic cores to bus interfaces onto a single chip with the deep submicron technology. Due to the fact that once System on a Chip (SoC) is fabricated it is impossible to remove, recondition, and replace a faulty core deeply embedded into the chip, test and repair process is one of the challenging issues for the success of SoC design. Therefore, more reliable and dependable SoC design with built-in self test/diagnosis/ repair is desirable and more suggested than in traditional ASIC (Application Specific Integrated Circuit) or MCM (Multichip Module) design. In the mean time, it is forecast by ITRS (International Technology Roadmap in Semiconductor) that memory cores will be dominant in SoC area and tend to be distributed into small available spaces left after placing logic cores and others into the SoC area. This opens up a new concept of memory test and repair.

Memory cores are the most space-dominant among numerous cores in SoC and ITRS forecast that more than 90% of the SoC area will be populated with embedded memory cores in a decade. Due to not only this fact but the fact that memory is a fault-prone component on SoC, the reliability of memory core plays a critical role in overall reliability of SoC. Also, it is well known that defects in VLSI circuits tend to occur in clusters due to defects that span multiple circuit elements [2].

The objective of this paper is to study the repair process of distributed embedded memory cores in SoC having new architecture for memory repair. We propose *Spare Line Borrowing* technique which utilize available redundancy of fault resistant memory cores for the repair of faulty cells in fault prone memory cores in SoC. The effect of fault tolerance in the architecture proposed will be proved through reliability analysis and extensive parametric simulation.

II. LITERATURE REVIEW

Since memory module plays a critical role in electronic devices and is fault-prone component in the system there have been significant efforts made on memory repair for

reliability. [4] has proposed the approaches for the repair of redundant RAMs in which redundant rows and columns are utilized as spares. [3] has proposed that the combined use of redundant circuits and error-correcting codes (ECC) can achieve a fault-tolerance scheme that is far more effective than an individual employment of either one of these schemes separately (so called synergistic fault tolerance).

Fault clustering has been extensively studied as well to take into account practical fault patterns. [1][2] have proposed a clustered failure model under which the problem of array reconfiguration is studied. [2] has adopted the center-satellite approach while [1] has adopted quadrant-based fault model. The quadrant-based model is preferred to the center-satellite model in practice since many parameters used in the center-satellite model makes its parameter estimation difficult [1]. [5] has proposed a reliability model of fault-tolerant onboard memory system under fault clustering.

With the emergence of SoC in the semiconductor industry, test/repair and reliability of SoC have become very important issues during the development process. [6] has addressed the challenges in testing core-based system ICs by describing and comparing the differences between traditional test method and core-based test method. [7] has proposed a repair method based on the connectivity of the chips on MCM (Multichip Module) in which yield degradation due to neighboring chips and interconnect structure was modeled and analyzed. Several MCM repair scheduling strategies based on the number of interconnections and the number of neighboring chips has been shown in [7], and it has evaluated the impact of connectivity-based repair scheduling on the overall yield of MCMs. In [8] the reliability of SoC design when bus errors affect the SoC interconnection architecture has been addressed and an approach to enhance the error detection and correction mechanism of the system bus has been proposed based on the concept of distributed bus guardians.

There have been a few works which deal with interconnections driven repair in the system. [9] presents the model and architecture for interconnection driven repair in embedded memory cores on a chip and demonstrates its repairability based on the model it suggests. Even though it gives a novel approach it doesn't show its implementation and practicality in detail. Also, it doesn't consider fault clustering case which happens often in real semiconductor chip.

III. DISTRIBUTED EMBEDDED MEMORY CORE ARCHITECTURE AND ITS REPAIR

In this section the architecture of distributed embedded memory core under investigation is introduced and its repair process is explained.

A typical distributed embedded memory core architecture in SoC suggested in this work is shown in Figure 1.
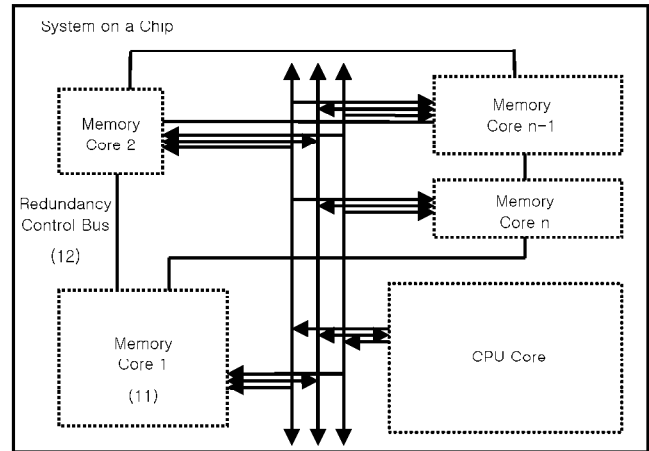

Figure 1. Typical distributed embedded memory core architecture in SoC

Bunch of memory cores are fabricated onto the available space in SoC where each memory cores are of different size. Note that, for simplicity, only embedded memory cores, CPU core, system buses, and redundancy control bus for repair are illustrated. Each memory core and CPU core are connected through system buses which generally consists of data bus, address bus, and control bus. In addition, distributed memory cores also have a interconnection between them through redundancy control bus which aims at control of adjacent memory core's redundant array. Each memory core implements redundancy for replacement of error block of either bit or word size along with ECC code inside. The architecture of redundancy control bus shown above is simply ring topology since each module is connected one another through one bus with ring-shape. This ring topology can go further into chordal ring of degree N if desired. By the way, the cost of implementation of interconnection of higher degree due to limited die area and reconfiguration algorithm complexity is unavoidable.

The interconnection topology investigated in this work is ring based. (i.e. degree of connectivity 2), and it goes to higher degree of connectivity such as chordal ring of degree 3 (i.e. degree of connectivity 3), chordal ring of degree 4 (i.e. degree of connectivity 4), etc.. Ring based interconnection topology provides the benefits such that its evenly dedicated interconnection not only offer simplicity of hardware implementation but gives fairness to all embedded memory cores in SoC in which the clustering fault phenomenon can possibly happen.
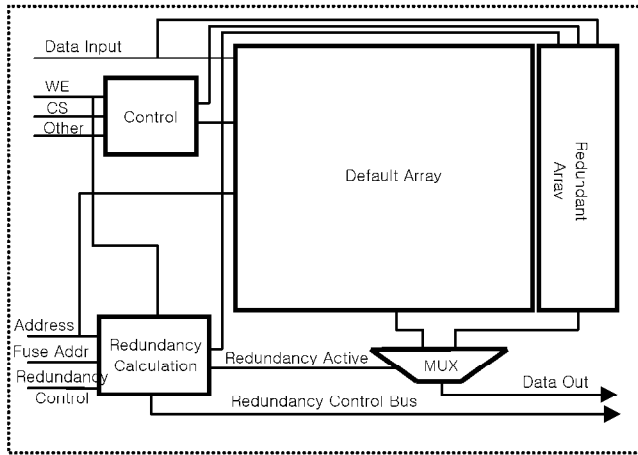
Figure 2. Memory core architecture featuring spare-array
borrowing technique

Figure 3. Spare Line Borrowing

Memory core architecture featuring *spare-array borrowing* technique under investigation in this work is shown in Figure 2. The memory array comprises default array and redundant array. The usual control signals from control circuit including write enable, chip select, and others are applied to the memory array of default array and redundant array. The addresses generated by processor for read and write operations are provided to the array of default memory array as well as redundancy calculation. Redundancy calculation then investigates first if the applied address matches the address borrowed from adjacent memory core and if so, reroutes the address signal to adjacent memory core through redundancy control bus. If the redundancy calculation does not have any log which it borrowed redundant array from adjacent memory core then it investigates if the applied address matches any defective addresses stored in the fuse banks, and if so, access the redundant array. In fault-free case, memory core doesn't have any delay time caused by redundancy calculation and thereafter since address data accesses default array directly and stop redundancy calculation processing right away, resulting same performance as normal memory cores[9]. In the last stage, MUX decides which data bus is valid between from default array and from redundant array in case that redundant array is used and final effective data is out to data line.

Figure 3 shows the result of spare line borrowing. Memory Core N have used up all its redundant lines: one for column 2, one for column 3, one for column 8. It has no more redundancy. On occurring another faulty cell(column 9) its redundancy calculation unit maps its faulty cell to the redundant column in its adjacent memory core. Consequently it is more fault tolerance than it has. Without spare line borrowing technique memory core N fails to operate resulting in overall memory system failure.
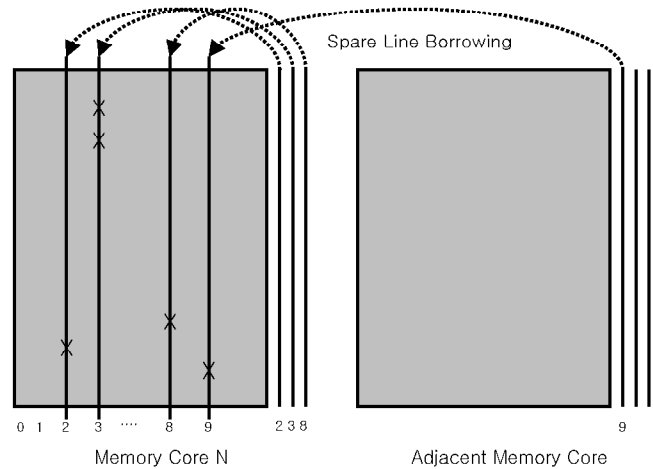
By simulating the scheme explained above each embedded memory core expects the redundant array lines as many as the total number of available redundant array lines adjacent embedded memory cores have, which results in significantly higher repairability of memory system in SoC.

The scheme proposed in this paper makes the most use of its repairability in the case where faulty cells show spatial locality, so called fault clustering. Figure 4 illustrates the memory repair in fault clustering. Figure (a) in Figure 4 shows the traditional memory repair without spare line borrowing technique. SoC has 4 memory cores in it: M1, M2, M3, M4. M1 and M3 memory cores have 4 faulty cells respectively. M2 has one faulty cell and M4 has no faulty cell. Memory cores use its redundant columns to fix the faulty columns. M2 uses one redundant column to fix faulty column. M1 and M3 use 3 redundant columns to fix its faulty columns respectively. M4 doesn't need to use its redundant column because it has no faulty column. On occurring one more faulty cell in M1, M3, M1 and M3 fails to function its operation because it can not fix its faulty cells because it has no more redundant columns. By the way, in memory repair with spare line borrowing technique proposed in this paper M1 and M3 borrows spare column from M4 which has available spare columns to fix its faulty cell respectively as can be seen in figure (b) in figure 4. Consequently M1 and M3 are still fault-tolerant and function correctly.
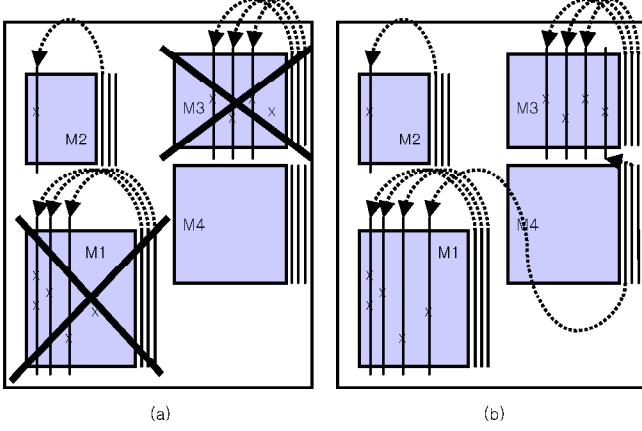
Figure 4. Memory Repair in Fault Clustering in SoC

IV. RELIABILITY ANALYSIS AND PARAMETRIC SIMULATION

In this section, the reliability of the distributed embedded memory cores featuring spare-array borrowing technique under investigation in this work is analyzed based on the proposed repair process.

The acronyms and notations to be used throughout this paper are summarized as follows.

• N_m : Number of memory modules in SoC

• N_c : Number of columns in a memory module,
     excluding spares

• N_r : Number of rows in a memory core

• S : Number of spare columns

• η : Degree of Interconnection(the number of adjacent memory cores)

• λ : Failure rate of memory cell

• P_{FP} : Probability that a memory core is fault prone

• λ_{FP} : Failure rate of memory cell in fault prone memory core

• λ_{FR} : Failure rate of memory cell in fault resistant memory core

The expected number of memory cell failures for a given time period Δt is defined by the *failure rate*. The exponential relationship between reliability and time is the *exponential failure law* which claims that for a constant failure rate, the component reliability changes exponentially as a function of time t. Therefore, with a fault arrival rate λ, the reliability is denoted as follows.

$$R(t) = e^{-\lambda t} \qquad (1)$$

Based on the reliability expressions, the reliability of M-N system where M components should operate properly in order for the whole system to be operational can be expressed as follows using the binomial distribution without loss of generality.

$$R_{M-N}(t) = \sum_{i=0}^{N-M} \binom{N}{i} \left(R(t)\right)^{N-i} \left(1.0 - R(t)\right)^{i}. \qquad (2)$$

Having the formula above the reliability of a memory module with S spare columns can be expressed as follows.

$$R_{module}(t) = \sum_{i=0}^{S} \binom{Nc+S}{i} \left(R_{column}(t)\right)^{Nc+S-i} \left(1.0 - R_{column}(t)\right)^{i}. \qquad (3)$$

Figure 5 shows the reliability of SoC based on the formula above when assuming that fault arrival rate $\lambda = 0.003$, $N_c = 32$, $N_r = 8$, and $S = 2,4,8,16$. As can be seen from the graph the reliability of a memory module grows when a module has more spare columns [3].

With the proposed spare line borrowing technique, by the way, the number of spare columns available for each module may be greater than S depending the interconnection topology. The probability that a column in a module is faulty is expressed as follows.

Pr { a column in a module is faulty }

$$= 1 - \text{Pr} \{ \text{a column in a module is fault-free} \}. \qquad (4)$$

Since each column contains N_r cells, the probability that a column in a module is fault-free is $(1 - \lambda)^{N_r}$. Thus Pr {a column in a module is faulty} is

$$1 - (1 - \lambda)^{N_r} \qquad (5)$$

The total expected number of faulty columns in a module then can be expressed as follows.

$$N_c \cdot (1 - (1 - \lambda)^{N_r}) \qquad (6)$$

Suppose there are N_m memory modules, and each memory module has the same fault rate λ. The spare columns that a module can lend to other modules would be the remaining ones after its own repair. The total number of spare columns available to a module in this case then is as follows.

$$S' = S + N_m \cdot \{S - (N_c \cdot (1 - (1 - \lambda)^{N_r}))\} \qquad (7)$$

The reliability of a module can be expressed as follows.

$$\sum_{i=0}^{S'} \binom{Nc+S'}{i} \left(R_{column}(t)\right)^{Nc+S'-i} \left(1.0 - R_{column}(t)\right)^{i} \qquad (8)$$

Figure 6 shows the reliability of a memory module in SoC based on the formula above when assuming the same configuration with Figure 5. The reliability of a memory module in SoC is much higher than the one without spare line borrowing technique. It is caused by the fact that each memory module has more spare columns than it originally has. Figure 7 and Figure 8 shows the difference of

reliability between the system with spare line borrowing technique and without it when assuming the same configuration. As can be seen from the graph the memory module with spare line borrowing technique is much more reliable which is desirable, especially, in SoC.
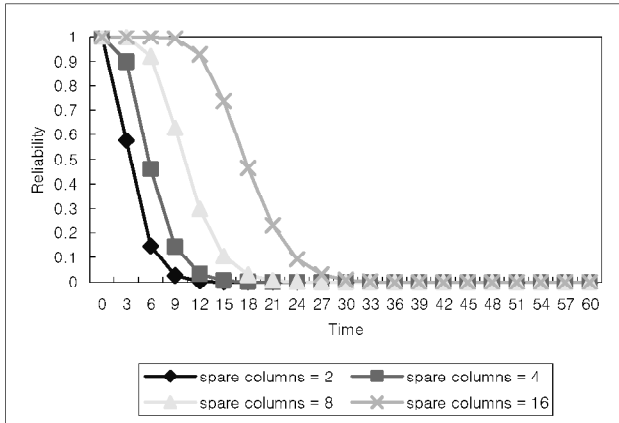

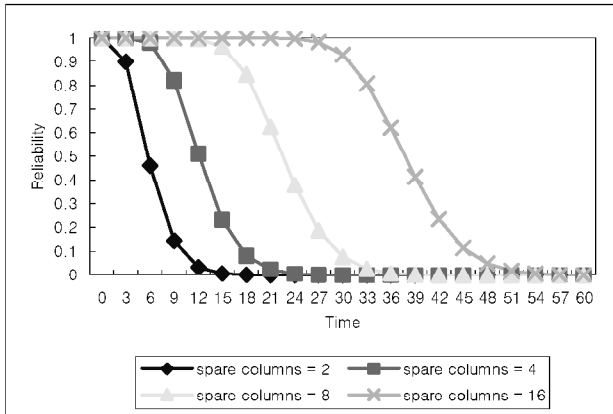Figure 5. Reliability of SoC without *Spare Line Borrowing Technique*

Figure 6. Reliability of SoC with *Spare Line Borrowing Technique*
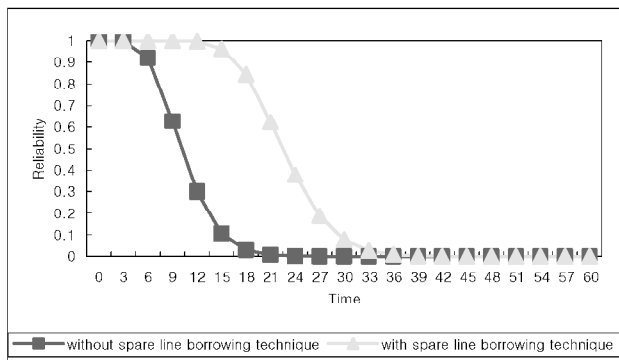

Figure 7. Reliability Comparison between the System with and without Spare Line Borrowing Technique when $S = 8$
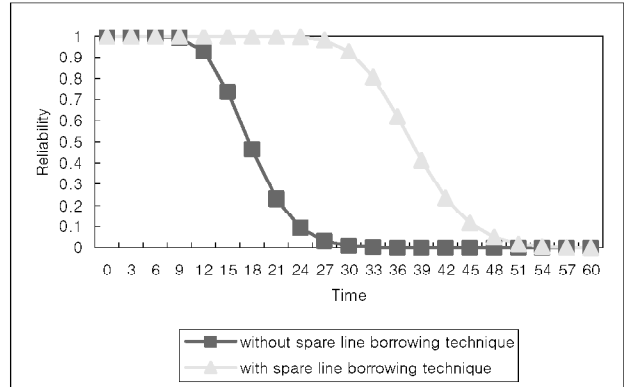
Figure 8. Reliability Comparison between the System with and without Spare Line Borrowing Technique when $S = 16$

The reliability of memory module in SoC in different interconnection topology is illustrated in Figure 9. Since higher interconnection topology means that a module has more adjacent modules it can borrow spare lines the higher interconnection topology the memory system in SoC has the higher reliability the memory system has. Note that implementing higher interconnection also carries more complex hardware design as well as higher cost to implement it.
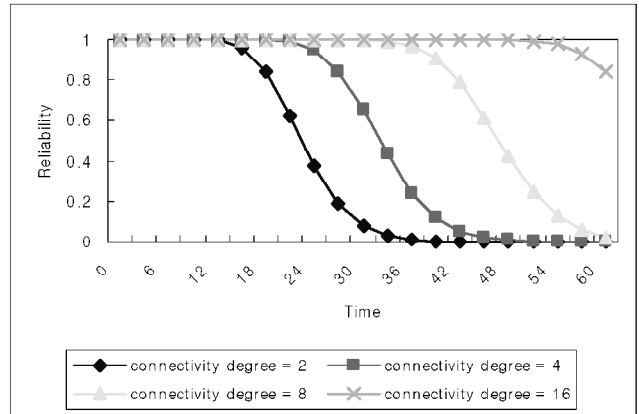

Figure 9. Reliability of SoC in different Interconnection Topology

In case of fault clustering, we have fault prone memory cores with probability λ_{FP} and fault resistant memory cores with probability $(1 - \lambda_{FP})$. For a fault prone cores, we have the probability that a column in a fault prone memory core is expressed as follows.

$$1 - (1 - \lambda_{FP})^{N_r} \qquad (9)$$

47

Then the total expected number of faulty columns in a fault prone memory core can be expressed as follows.

$$N_c \cdot (1 - (1 - \lambda_{FP})^{N_r}) \qquad (10)$$

In a similar fashion, the total expected number of faulty columns in a fault resistant memory module can be expressed as follows.

$$N_c \cdot (1 - (1 - \lambda_{FR})^{N_r}) \qquad (11)$$

Since we have the probability that a memory core is a fault prone is P_{FP}, the total expected number of faulty columns in a memory core is expressed as follows.

$$P_{FP} \cdot (N_c \cdot (1 - (1 - \lambda_{FP})^{N_r})) + (1 - P_{FP}) \cdot (N_c \cdot (1 - (1 - \lambda_{FR})^{N_r})) \qquad (12)$$

Then the total number of spare columns available to a memory core in a fault clustering case is as follows.

$$S' = N_m \cdot \{S - \{P_{FP} \cdot (N_c \cdot (1 - (1 - \lambda_{FP})^{N_r})) + (1 - P_{FP}) \cdot (N_c \cdot (1 - (1 - \lambda_{FR})^{N_r}))\}\} \qquad (13)$$

The reliability of a memory core can be expressed as follows.

$$R'_{\text{module}}(t) = \sum_{i=0}^{S'} \binom{Nc+S'}{i} (R_{\text{column}}(t))^{Nc+S'-i} (1.0 - R_{\text{column}}(t))^i \qquad (14)$$

V. CONCLUSION AND DISCUSSION

This work has proposed a architecture for memory repair process for distributed embedded memory cores in SoC. The spare line borrowing technique where one memory core borrows redundant columns from its adjacent memory cores through software reconfiguration is introduced and explained. The architecture and repair process proposed in this work has significant fault tolerance enhancement in fault clustering case. The effect of interconnection topology on the reliability of distributed embedded memory modules has been evaluated through parametric simulation with respect to the proposed process to exploit the utilization of distributed spare columns over the interconnection network. It has been observed that the system with spare line borrowing technique has higher reliability than the system without it. It also has been observed that the reliability of memory system increases as the number of spare columns and the degree of connectivity increases. Even though more spare columns in a memory module and higher interconnection topology make higher reliable system possible it also carries higher hardware implementing cost. Therefore, it is concluded based on the modeling and analysis in this paper that system requirements for reliability and a practical cost justification is two of the most important factors when designing SoC.

REFERENCES

[1] D.M. Blough, "Performance Evaluation of a Reconfiguration-Algorithm for Memory Arrays containing Clustered Faults". IEEE Transactions on Reliability, 45(2), 1996, 274-284.

[2] D.M. Blough, and A. Pelc, "A Clustered Failure Model for the Memory Array Reconfiguration Problem". IEEE Transactions on Computers, 42(5), 1993, 518-528.

[3] C.H. Stapper, and H.S. Lee, "Synergistic fault-tolerance for memory chips". Computers, IEEE Transactions on, 41(9), 1992, 1078-1087.

[4] F. Lombardi, and W.K. Huang, "Approaches for the repair of VLSI/WSI RRAMs by row / column deletion". Fault-Tolerant Computing, 1988. FTCS-18, Digest of Papers., Eighteenth International Symposium on, 1988, 342-347.

[5] M. Choi, N. Park, F.J. Meyer, F. Lombardi, and V. Piuri, "Reliability measurement of fault-tolerant onboard memory system under fault clustering". Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE, Vol. 2, 2002, 1161-1166.

[6] E.J. Marinissen, and Y. Zorian, "Challenges in testing core-based system ICs". IEEE Communication Magazine, 37(6), 1999, 104-109.

[7] M. Choi, N. Park, F. Meyer, and F. Lombardi, "Connectivity-based multichip module repair". Dependable Computing, 2001. Proceedings. 2001 Pacific Rim International Symposium on, 2001, 19-26.

[8] M. Lajolo, "Bus guardians: an effective solution for online detection and correction of faults affecting system-on-chip buses". Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 9(6), Dec. 2001, 974-982.

[9] B. Jang, N. Park, K.M. George, G.E. Hedrick, "Modeling and Evaluation of the Interconnection-Driven Repairability for Distributed Embedded Memory Cores on Chip". International Conference on Modeling, Identification and Control (MIC 2003), IASTEAD, Innsbruck, Austria, Feb. 2003.