Pre-print version

M. Choi, N. Park, V. Piuri, Y.-B. Kim, F. Lombardi, "Balanced dual-stage repair for dependable embedded memory cores", in *Journal of Systems Architecture*, pp. 281-285, April, 2004. ISSN: 1383-7621

Balanced dual-stage repair for dependable embedded memory cores

Minsu Choi [a,*], Nohpill Park [b], Vincenzo Piuri [c], Yong-Bin Kim [d], Fabrizio Lombardi [d]

[a] Department of Electrical and Computer Engineering, University of Missouri-Rolla, Rolla, MO 65409, USA
[b] Department of Computer Science, Oklahoma State University, Stillwater, OK 74078-1053, USA
[c] Department of Information Technologies, University of Milan, Bramante 65, 26013 Crema (CR), Italy
[d] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA

Abstract

Advances in revolutionary system-on-chip (SoC) technology mainly depend on the high-performance ultra-dependable system core components. Among those core components, embedded memory system core, currently acquiring 54% of SoC area share, will continue its domination of SoC area share as it is anticipated to approach about 94% of SoC area share by the year 2014. Since memory cells are considered as more prone to defects and faults than logic cells, redundancy and repair have been extensively practiced for enhancing defect and fault tolerance. Unlike in legacy PCB (printed circuit board) or MCM (multichip module) based systems, embedded core components cannot be physically replaced once they are fabricated onto a SoC. To realize enhanced manufacturing yield and field reliability, both ATE (automated test equipment) and BISR (built-in-self-repair) are commonly utilized to allocate redundancy for embedded memory system cores. Since ATE (for repairing manufacturing defects) and BISR (for repairing field faults) share the given redundancy, probabilistic redundancy partitioning and utilization techniques are proposed in this paper to achieve optimal combination of yield and reliability of the embedded memory system core. Parametric simulation results are shown extensively.

Keywords: SoC (system-on-chip) technology; Embedded memory core repair; ATE (automated test equipment); BIST/D/R (built in self-test, diagnosis and repair); Redundancy balancing

1. Introduction

As advances in ultra-large-scale-integration (ULSI) technologies make possible the seamless embedding of numerous cores on a single chip (i.e., commonly referred to as system-on-chip (SoC) technology), solid dependability becomes an urgent requirement of such ultra density and high performance system, since insignificant degradation or defect of core components could result in unacceptably low resultant SoC manufacturing yield and field reliability. Among the cores for SoC integration, one of the most sensitive cores is the

* Corresponding author. Tel.: +1-573-341-4524; fax: +1-573-341-4532.
E-mail address: choim@umr.edu (M. Choi).

1383-7621/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysarc.2003.08.004

embedded memory core since memory cells are commonly considered as more prone to defects and faults than logic cells [2–4,7–9]. As SoC fabrication process goes toward the era of the deep-sub-micron technology and beyond, need for a high yield and ultra reliable embedded memory core becomes obvious, as embedded memory will continue to dominate SoC content in the next several years, approaching 94% of the die area by 2014 [5].

Traditionally, reconfiguration (repair) of memory arrays using spare memory lines is the most common technique for yield enhancement of memories with faults [2–4,7–10]. Unfortunately, fabricated embedded memory system core cannot be physically replaced in field. Thus, built-in test, diagnosis and repair circuits are commonly practiced along with ATE-based repair to assure improved manufacturing yield and field reliability of the embedded memory core [2–4,7]. The combination of both ATE and BISR repair technologies is referred to as the dual-stage repair in this paper.

Although it is obvious that combination of ATE and BISR is able to achieve significant manufacturing yield (i.e., the probability of being manufactured and repaired as functional) and field reliability (i.e., a function of time which is defined by the conditional probability that the system performs correctly throughout the interval of time $[t_0; t]$ given the system was performing flawlessly at the initial time t_0) enhancements for embedded memory system core, one problem still remains unsolved: balanced redundancy partitioning and utilization. Since ATE (for repairing manufacturing defects) and BISR (for repairing field faults) share the given common redundancy, balanced redundancy partitioning and utilization techniques are very important to achieve ultimate combination of yield and reliability of the embedded memory system core. The main objective of the paper is to propose a balanced redundancy utilization technique for such core with dual-stage repair and shared redundancy. Thus, straightforward dependability evaluation techniques for existing redundancy architectures will be initially investigated to unveil true significance of redundancy balancing. Then, balanced redundancy partitioning and utilization techniques for both single and two dimensional redundancy architectures will be investigated. Extensive parametric simulation results will be also shown.

2. Preliminaries

Embedded memory system core under investigation, in which both ATE-based factory repair and BISR-based field repair are practiced for manufacturing yield and field reliability enhancements, consists of the following components; (1) IEEE JTAG (Joint Test Action Group) 1149.1: Boundary-scan interface for test and repair [1], (2) Laser fuse: A set of laser reconfigurable fuses to permanently program the given redundancy resources in factory [6], (3) BIST/BISD/BISR processor: This system component governs self-test, self-diagnosis, and self-repair procedures, (4) Programmable fuse: A set of programmable fuses to store additional reconfiguration signature generated by the BIST/BISD/BISR processor in field, (5) Memory array interface: This component connects the EAB (embedded array block) array and the BIST/BISD/BISR Processor together. Data, address, control and repair data flow via this component (Fig. 1).

The given embedded memory system core is tested and repaired as follows; (1) Factory repair: To circumvent defects due to imperfect manufacturing processes, the ATE communicates with the embedded memory system core via its external test

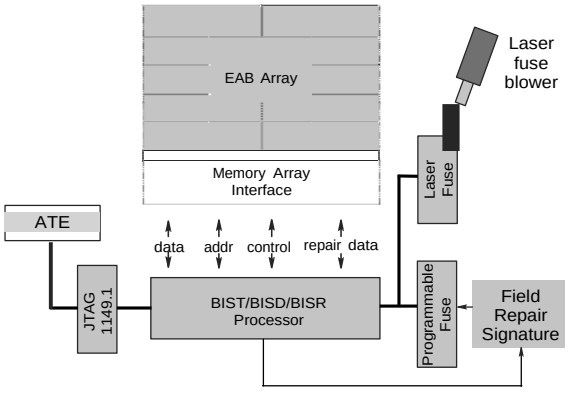

Fig. 1. Model embedded memory system core architecture.

equipment interface (as provided by pin electronics in the head architecture). Then, the laser fuse is permanently programmed to allocate redundancy and to repair manufacturing defects in the EABs, (2) Field repair: Whenever the SoC is reset or powered, the BISR circuitry tests, diagnoses and repairs the EABs. The programmable units (such as switches) are programmed to store the information for redundancy allocation.

Although it is obvious that combination of ATE and BISR is able to achieve significant manufacturing yield and field reliability enhancements for embedded memory system core, one problem still remains unsolved: How the given redundancy can be partitioned into two groups (i.e., one for ATE repair and another one for BISR repair) to balance the manufacturing yield and field reliability for the maximum dependability? Since ATE (for repairing manufacturing defects) and BISR (for repairing field faults) share the given common redundancy, balanced redundancy partitioning and utilization is very important to achieve ultimate combination of yield and reliability of the embedded memory system core. Balanced redundancy partitioning and utilization of the given embedded memory system core with single dimensional redundancy case will be investigated in the next section followed by two dimensional case in Section 4.

3. Single dimensional redundancy case

The following notations will be used throughout this research; (1) n_c: Number of columns (i.e., number of bits per word), (2) n_r: Number of rows (i.e., number of words), (3) s_c: Number of spare columns, (4) s_{cm}: Number of spare columns used for manufacturing yield enhancement (i.e., $S_c - S_{cf}$), (5) s_{cf}: Number of spare columns used for field reliability enhancement (i.e., $S_c - S_{cm}$), (6) k_m: Expected number of manufacturing defects per memory cell, (7) k_f: Field failure arrival rate of memory cell per unit time interval, (8) Y: Manufacturing yield, and (9) $R(t)$: Field reliability at time t, (10) $D(t)$: Overall dependability.

The yield of a single cell can be formulated by the exponential failure law as $Y_{cell} = e^{-k_m}$. Then, the probability of having n_r non-defective cells in a column (i.e., the yield of a column) can be written as $Y_{column} = (Y_{cell})^{n_r}$. The given memory consists of n_c memory columns and s_{cm} spare memory columns for yield enhancement. The quorum size of n_c of the total of $n_c + s_{cm}$ columns are required to be functional. Success of exactly r-out-of-n identical and independent items, where each item has a success rate of p, is given by $P_s(r : n : p) = \sum_{k=r}^{n} \binom{n}{r} p^r (1-p)^{(n-r)}$. Thus, yield of the given memory with column-redundancy can be formulated by the binomial distribution as follows.

$$Y = P_s(s_{cm} : n_c + s_{cm} : Y_{column}) \quad (1)$$

Reliability assurance equations are similar to the yield assurance equations and can be shown as follows; $R_{cell}(t) = e^{-k_f \cdot t}$, $R_{column}(t) = (R_{cell})^{n_r}$, and then,

$$R(t) = P_s(s_{cf} : n_c + s_{cf} : R_{column}(t)) \quad (2)$$

The conditional probability of having manufactured-as-good (i.e., Y) and not-failing-in-field during the time interval $[t_0; t]$ (i.e., $R(t)$) is referred to as dependability denoted by $D(t)$. Since Y and $R(t)$ are serial probabilities, product of Eqs. (1) and (2) can be used to formulate $D(t)$; $D(t) = Y \cdot R(t)$.

To find the most balanced s_{cm}, $D(t)$ can be differentiated and solved with respect to s_{cm} as $dD(t)/ds_{cm} = 0$.

Note that s_{cf} follows since $s_{cf} = s_c - s_{cm}$ holds. s_{cm} must be an integer value. So, both $\lceil s_{cm} \rceil$ and $\lfloor s_{cm} \rfloor$ must be evaluated to determine the final partitioning position. Later, they are partitioned into two groups: two spare columns for ATE repair and four spare columns for BISR repair.

4. Two dimensional redundancy case

The following notations will be used in addition to the ones given in the previous section throughout this research; (1) s_r: Number of spare rows, (2) s_{rm}: Number of spare rows used for manufacturing yield enhancement (i.e., $S_r - S_{rf}$), (4) s_{rf}: Number of spare rows used for field reliability enhancement (i.e., $S_c - S_{rm}$), (5) k_{cm}: Expected number of manufacturing defects per memory column, (6) k_{rm}: Expected

number of manufacturing defects per memory row, (7) k_{cf}: Field failure arrival rate of memory column per unit time interval, and (8) k_{rf}: Field failure arrival rate of memory row per unit time interval. Since row/column deletion is one of the NP-complete problems. There is no effective way to derive closed formulae for Y and Rðtþ. So, in this paper, line-based fault model is used rather than cell-based faulty model for 2D case to assure effectiveness of the proposed redundancy balancing method. Mathematical yield and reliability models from any repair algorithms can be inputted to find the most balanced redundancy partitioning locations.

The yield of a row and a column can be approximated as $Y_{row} = e^{-k_{rm}}$ and $Y_{column} = e^{-k_{cm}}$.

Then, the yield of rows is

$$Y_{rows} = P_s(s_{rm} : n_r - s_{rm} : Y_{row}) \quad (3)$$

and the yield of columns is

$$Y_{columns} = P_s(s_{cm} : n_c - s_{cm} : Y_{column}) \quad (4)$$

Thus, the overall yield is $Y = Y_{rows} \times Y_{columns}$. Likewise, $R_{row}(t) = e^{-k_{rm} \cdot t}$, $R_{column}(t) = e^{-k_{cm} \cdot t}$, and

$$R_{rows}(t) = P_s(s_{rf} : n_r - s_{rf} : Y_{row}) \quad (5)$$

$$R_{columns}(t) = P_s(s_{cf} : n_c - s_{cf} : Y_{column}) \quad (6)$$

Therefore, $R(t) = R_{rows}(t) \times R_{columns}(t)$ and the overall dependability, then, can be written as $D(t) = Y \times R(t)$.

To find the most balanced s_{cm} and s_{rm}, D(t) can be differentiated and solved with respect to s_{cm} and s_{rm} as $\frac{d^2 D(t)}{ds_{cm}\, ds_{rm}} = 0$. Note that s_{cf} and s_{rf} follow since $s_{cf} = s_c - s_{cm}$ and $s_{rf} = s_r - s_{rm}$ hold. s_{cm} and s_{rm} must be integer values. So, both $\lfloor s_{cm} \rfloor$ & $\lceil s_{cm} \rceil$ and $\lfloor s_{rm} \rfloor$ & $\lceil s_{rm} \rceil$ must be evaluated to determine the final partitioning positions. Later, spare columns are partitioned into two groups: two spare columns for ATE repair and four spare columns for BISR repair and spare rows are also partitioned into two groups: three spare columns for ATE repair and three spare columns for BISR repair.

5. Parametric simulation and results

The effect of redundancy balancing for the one and two dimensional cases will be studied through numerical experiments. The initial parameters used in the simulation are $n_c = 128$, $n_r = 128$, $s_c = 8$, $k_m = 10^{-4}$, $k_f = 10^{-5}$ and $t = 10$ for one-dimensional case and $n_c = 128$, $n_r = 128$, $s_c = 8$, $s_r = 8$, $k_{cm} = 10^{-2}$, $k_{rm} = 10^{-2}$, $k_c f = 10^{-3}$, $k_r f = 10^{-3}$ and $t = 10$ for two-dimensional case.

The simulation results for the single dimensional redundancy case are shown in Fig. 2, the following observations are made; (1) Symmetric case: The dependability derivative is plotted with respect to s_{cm}, using points with blank box symbol. In this case, parameters are selected in order to show symmetric behavior in both Y and Rðtþ. Thus, the partition is [4; 4]: 4 spare columns for factory repair and 4 spare columns for field repair, i.e. spares are evenly partitioned and utilized, (2) Reliability-intensive case: The dependability derivative is plotted with respect to s_{cm}, using points with cross symbol. In this case, parameters are selected to take into account the presence of more field faults than manufacturing defects (i.e., $t = 10 \blacksquare 20$). Thus, the partition is now [2,4]: 3 spare columns for factory repair and 5 spare columns for field repair, i.e. in this case spares are partitioned and utilized to achieve an enhancement in field reliability, (3) Yield-intensive case: The dependability derivative is plotted with respect to s_{cm}, using points with blank diamond symbol. In this case, parameters are selected such that more manufacturing defects than field faults are induced (i.e., $k_m = 10^{-4} \blacksquare 2 \times 10^{-4}$). Thus, the partition is [5; 3]: 5 spare columns for factory repair and 3 spare columns for field repair. Spares are partitioned and utilized to increase the manufacturing yield.

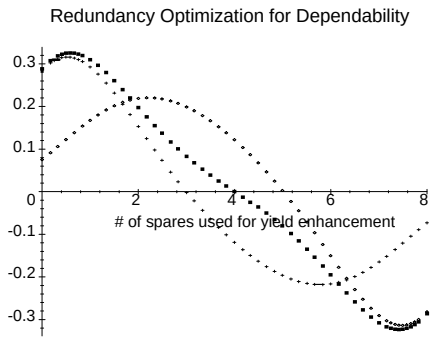

Fig. 2. Balanced partition results: [3; 5], [4; 4] and [5; 3].

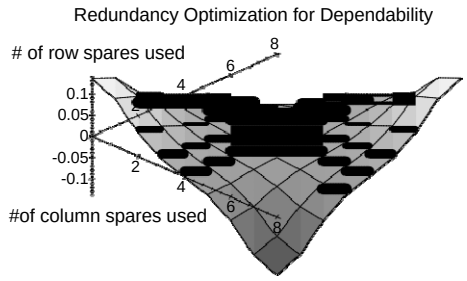

Fig. 3. Balanced partition for {[4,4],[4,4]}.

For the two-dimensional redundancy case, similar observations can be made. For instance, the redundancy balancing result of {[4,4],[4,4]} is shown in Fig. 3.

6. Discussion and conclusions

Among cores for SoC integration, the embedded memory core occupies the largest area and its performance in terms of manufacturing yield and field reliability is very important because memory cells are usually more prone to defects and faults than other components such as logic cells. Since cores cannot be physically replaced once they are fabricated onto a SoC, a combination of ATE and BISR is common practice. Proper partition and utilization of the spare redundancy is significantly desirable to achieve a balance of manufacturing yield and field reliability of the embedded memory system. Thus, yield and reliability assurance techniques have been proposed for the one dimensional redundancy case (only spare columns), then extended to two-dimensional redundancy case (both spare rows and columns are present). As a figure of merit between yield and reliability, dependability (i.e., Y × Rðtþ) is proposed. Maximum dependability is possible only if the provided redundancy is partitioned properly into two groups to repair both manufacturing defects (i.e., ATE-based repair) and field faults (i.e., BISR). To accomplish a balanced redundancy partition and utilization, the equations for the dependability are differentiated and solved with respect to the number of spares used to enhance the manufacturing yield. Parametric simulation results have verified that the proposed redundancy partition and utilization techniques for embedded memory system cores achieves the theoretically optimal value. The proposed techniques can be incorporated with existing CAD compilers for embedded memory system cores, thereby providing a cost-effective partition and utilization of the shared redundant spares.

References

[1] JTAG Technologies BV, "Boundary-scan: Unlock the power of boundary-scan—Technical Brochure, http://www.jtag.com, JTAG Technologies BV, Stevensville, MD, USA.

[2] J. Ohtani, T. Ooishi, et al., A shared built-in self-repair analysis for multiple embedded memories, in: Custom Integrated Circuits, IEEE Conference on, May 2001, pp. 187–190.

[3] R.J. McPartland, D.J. Loeper et al., SRAM embedded memory with low cost, FLASH EEPROM-switch-controlled redundancy, in: Custom Integrated Circuits Conference, CICC. Proceedings of the IEEE 2000, Vol. 36, No. 11, May 2000, pp. 287–289.

[4] T. Kawagoe, J. Ohtani, et al., A built-in self-repair analyzer (CRESTA) for embedded DRAMs, in: Test Conference. Proceedings. International, Vol. 41, No. 9, October 2000, pp. 567–574.

[5] International Technology Roadmap for Semiconductors, 2000.

[6] K. Arndt, C. Narayan et al., Reliability of laser activated metal fuses in DRAMs, in: Electronics Manufacturing Technology Symposium, Twenty-Fourth IEEE/CPMT, October 1999, pp. 389–394.

[7] D.K. Bhavsar, An algorithm for row-column self-repair of RAMs and its implementation in the Alpha 21264, in: Test Conference. Proceedings. International, September 1999, pp. 311–318.

[8] C.P. Low, H.W. Leong, A new class of efficient algorithms for reconfiguration of memory arrays, IEEE Transactions on Computers 45 (5) (1996) 614–618.

[9] D.M. Blough, Performance evaluation of a reconfiguration-alogorithm for memory arrays containing clustered faults, IEEE Transactions on Reliability 45 (2) (1996) 274–284.

[10] S.Y. Kuo, W.K. Fucks, Efficient spare allocation in reconfigurable arrays, IEEE Design and Test 41 (9) (1987) 24–31.