Balanced Redundancy Utilization in Embedded Memory Cores for Dependable Systems

M. Choi and N. Park
Dept. of CS
Oklahoma State University
Stillwater, OK 74078-1053
choim@a.cs.okstate.edu
npark@a.cs.okstate.edu

F. Lombardi and Y.B. Kim
Dept. of ECE
Northeastern University
Boston, MA 02115
lombardi@ece.neu.edu
ybk@ece.neu.edu

V. Piuri
Dept. of IT
University of Milan
Bramante 65, 26013
Crema (CR), Italy
piuri@dti.unimi.it

Abstract

Advances in revolutionary System-on-Chip (SoC) technology mainly depend on the high performance and 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 & 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, balanced 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 for both single dimensional (i.e., spare columns) and two dimensional (i.e., both spare columns and rows) are shown.

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* technology), solid dependability becomes an urgent requirement of such ultra density & 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 embedded memory core since memory cells are commonly considered as more prone to defects and faults than logic cells [1, 2, 3, 5, 6, 7, 8, 9]. As SoC fabrication process goes toward the era of the deep-sub-micron technology such as $0.13\mu m$, need for a high yield and ultra reliable embedded memory core becomes obvious. According to Semiconductor Industry Association and ITRS2000, embedded memory will continue to dominate SoC content in the next several years, approaching 94% of the die area by 2014 [4]. The issues surrounding high-density multi-megabit embedded memory dependability must be solved in order to facilitate this trend and to produce cost effective SoC product.

Proceedings of the 17th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems (DFT'02)
1063-6722/02 $17.00 © 2002 **IEEE**

Traditionally, reconfiguration (repair) of memory arrays using spare memory lines is the most common technique for yield enhancement of memories with faults [1, 2, 3, 5, 6, 7, 8, 9]. 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, 1, 3, 5]. [2] proposed a SRAM Embedded Memory with Low Cost, FLASH EEPROM-Switch-Controlled Redundancy. In [3], very simple built-in-self-analysis-repair scheme called CREASTA for embedded DRAM is proposed while a row-column self-repair scheme for embedded SRAM for Alpha 21264 is shown in [5]. A shared built-in self-repair analysis scheme (Shared-BISA) for multiple embedded memory cores in the SoC is proposed in [1] to realize minimum area penalty independent of the number of embedded memory cores.

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. Thus, straightforward dependability evaluation techniques for single and two dimensional 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.

The organization of the paper is as follows. In the following section (Section 2), a conceptual architectural model of the embedded memory core with both ATE and BISR repair capabilities will be shown and significance of redundancy partitioning and utilization for balanced yield and reliability will be discussed. In Section 3 and 4, detailed yield and reliability assurance techniques for single and two dimensional redundancy cases will be shown. Then, balanced redundancy partitioning and utilization techniques for both cases will be proposed as well. A set of parametric simulations further verifies effectiveness of the proposed redundancy balancing techniques in Section 5. Finally, discussion and conclusions will be given in Section 6.

2: Preliminaries

Figure 1 shows a model of 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. The given embedded memory system core consists of the following components: *(1) IEEE JTAG (Joint Test Action Group) 1149.1* : External ATE interface for factory repair, *(2) Laser Fuse* : A set of laser reconfigurable fuses to permanently program the given redundancy resources in factory, *(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.

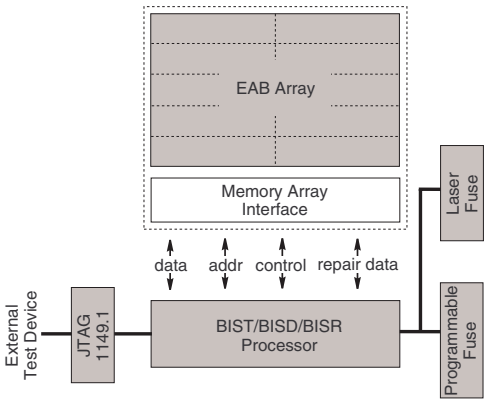

Figure 1. Embedded memory system core model under investigation

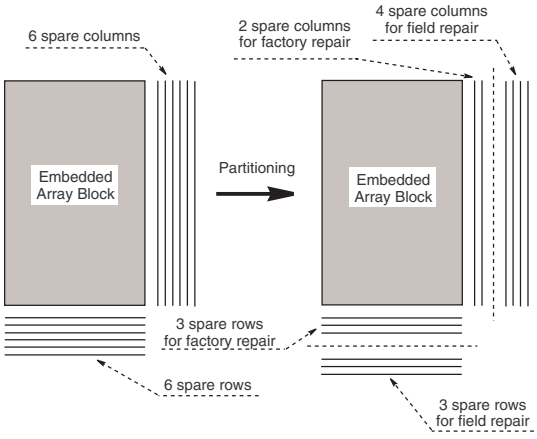

Figure 2. Example of redundancy partitioning for two dimensional redundancy

The given embedded memory system core is tested and repaired as follows: *(1) Factory Repair Process*: To circumvent the defects due to imperfect manufacturing processes, ATE communicates with the embedded memory system core via the external test equipment interface. Then, the laser fuse is permanently programmed to allocate redundancy to repair manufacturing defects in EABs, *(2) Field Repair Process*: Whenever the host SoC is reset or powered, BISR tests, diagnoses and repairs EABs. The programmable fuse is programmed to store redundancy allocation information.

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:

- n_c : Number of columns (i.e., number of bits per word).
- n_r :Number of rows (i.e., number of words).
- s_c : Number of spare columns.
- s_{cm}: Number of spare columns used for manufacturing yield enhancement (i.e., $S_c - S_{cf}$).
- s_{cf}: Number of spare columns used for field reliability enhancement (i.e., $S_c - S_{cm}$).
- λ_m : Expected number of manufacturing defects per memory cell.
- λ_f : Field failure arrival rate of memory cell per unit time interval.
- Y : Manufacturing yield.
- $R(t)$: Field reliability at time t.
- $D(t)$: Overall dependability.

The yield of a single cell can be formulated by the exponential failure law as

$$Y_{cell} = e^{-\lambda_m} \tag{1}$$

Then, the probability of having n_r non-defective cells in a column (i.e., the yield of a column) can be written as

$$Y_{column} = (Y_{cell})^{n_r} \tag{2}$$

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. Thus, yield of the given memory with column-redundancy can be formulated by the binomial distribution as follows.

$$Y = \sum_{i=0}^{s_{cm}} \binom{n_c + s_{cm}}{i} (Y_{column})^{n_c+s_{cm}-i} \cdot (1.0 - Y_{column})^i \tag{3}$$

Reliability assurance equations are similar to the yield assurance equations and can be shown as follows.

$$R_{cell}(t) = e^{-\lambda_f \cdot t} \tag{4}$$

$$R_{column}(t) = (R_{cell})^{n_r} \tag{5}$$

$$R(t) = \sum_{i=0}^{s_{cf}} \binom{n_c + s_{cf}}{i} (R_{column}(t))^{n_c+s_{cf}-i} \cdot (1.0 - R_{column}(t))^i \tag{6}$$

$$= \sum_{i=0}^{s_c-s_{cm}} \binom{n_c + s_c - s_{cm}}{i} (R_{column}(t))^{n_c+s_c-s_{cm}-i} \cdot (1.0 - R_{column}(t))^i \tag{7}$$

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 equations 3 and 7 can be used to formulate $D(t)$.

$$D(t) = Y \cdot R(t) \tag{8}$$

$$= \sum_{i=0}^{s_{cm}} \binom{n_c + s_{cm}}{i} (Y_{column})^{n_c+s_{cm}-i} \cdot (1.0 - Y_{column})^i \times \sum_{i=0}^{s_c-s_{cm}} \binom{n_c + s_c - s_{cm}}{i} (R_{column}(t))^{n_c+s_c-s_{cm}-i} \cdot (1.0 - R_{column}(t))^i \tag{9}$$

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

$$\frac{dD(t)}{ds_{cm}} = 0 \tag{10}$$

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:

- s_r : Number of spare rows.
- s_{rm}: Number of spare rows used for manufacturing yield enhancement (i.e., $S_r - S_{rf}$).
- s_{rf}: Number of spare rows used for field reliability enhancement (i.e., $S_c - S_{rm}$).
- λ_{cm} : Expected number of manufacturing defects per memory column.
- λ_{rm} : Expected number of manufacturing defects per memory row.
- λ_{cf} : Field failure arrival rate of memory column per unit time interval.
- λ_{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 $R(t)$. So, in this paper, line-based fault model is used rather than cell-based faulty model for 2-D case.

The yield of a row and a column can be approximated as $Y_{row} = e^{-\lambda_{rm}}$ and $Y_{column} = e^{-\lambda_{cm}}$. Then, the yield of rows is

$$Y_{rows} = \sum_{i=0}^{s_{rm}} \binom{n_r + s_{rm}}{i} (Y_{row})^{n_r + s_{rm} - i} \cdot (1.0 - Y_{row})^i \tag{11}$$

and the yield of columns is

$$Y_{columns} = \sum_{i=0}^{s_{cm}} \binom{n_c + s_{cm}}{i} (Y_{column})^{n_c + s_{cm} - i} \cdot (1.0 - Y_{column})^i \tag{12}$$

Thus, the overall yield is

$$Y = Y_{rows} \times Y_{columns} \tag{13}$$

Likewise,

$$R_{row}(t) = e^{-\lambda_{rm} \cdot t} \tag{14}$$

$$R_{column}(t) = e^{-\lambda_{cm} \cdot t} \tag{15}$$

$$\begin{aligned} R_{rows}(t) &= \sum_{i=0}^{s_{rf}} \binom{n_r + s_{rf}}{i} (R_{row}(t))^{n_r + s_{rf} - i} \cdot (1.0 - R_{row}(t))^i \\ &= \sum_{i=0}^{s_r - s_{rm}} \binom{n_r + s_r - s_{rm}}{i} (R_{row}(t))^{n_r + s_r - s_{rm} - i} \cdot (1.0 - R_{row}(t))^i \end{aligned} \tag{16}$$

$$\begin{aligned} R_{columns}(t) &= \sum_{i=0}^{s_{cf}} \binom{n_c + s_{cf}}{i} (R_{column}(t))^{n_c + s_{cf} - i} \cdot (1.0 - R_{column}(t))^i \\ &= \sum_{i=0}^{s_c - s_{cm}} \binom{n_c + s_c - s_{cm}}{i} (R_{column}(t))^{n_c + s_c - s_{cm} - i} \cdot (1.0 - R_{column}(t))^i \end{aligned} \tag{17}$$

$$R(t) = R_{rows}(t) \times R_{columns}(t) \tag{18}$$

The overall dependability, then, can be written as

$$D(t) = Y \times R(t) \tag{19}$$

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 follows.

$$\frac{d^2 D(t)}{ds_{cm} ds_{rm}} = 0 \tag{20}$$

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 $\lceil s_{cm} \rceil$ & $\lfloor s_{cm} \rfloor$ and $\lceil s_{rm} \rceil$ & $\lfloor s_{rm} \rfloor$ must be evaluated to determine the final partitioning positions. Figure 2 shows an example of a EAB with six spare columns and six spare rows. 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 Simulations and Results

The effect of the redundancy balancing for both single and two dimensional cases will be studied through numerical experiments. Parameters used in the simulation for the single dimensional redundancy case are summarized in Table 1 and for the two dimensional redundancy case are summarized in Table 1

Table 1. Simulation parameters for the one and two dimensional cases

Parameters	n_c & n_r	s_c & s_r	λ_m	λ_f	λ_{cm} & λ_{rm}	λ_{cf} & λ_{rf}	t
Values	128	8	10^{-4}	10^{-5}	10^{-2}	10^{-3}	10

The simulation results for the single dimensional redundancy case are shown in Figure 3 - 8, the following observations can be addressed.

- *Symmetric Case* : In Figure 3 and 4, the dependability and its derivative of the given EAB are plotted with respect to s_{cm}. In this case, parameters are selected in order that Y and $R(t)$ show exactly symmetric behaviors. Thus, partitioning result is [4,4]: 4 spare columns for factory repair and 4 spare columns for field repair. Spares are evenly partitioned and utilized in this case.
- *Reliability-Intensive Case* : In Figure 5 and 6, the dependability and its derivative of the given EAB are plotted with respect to s_{cm}. In this case, parameters are selected in order that more field faults than manufacturing defects are induced (i.e., $t = 10 \rightarrow 20$). Thus, partitioning result is [3,5]: 3 spare columns for factory repair and 5 spare columns for field repair. Spares are partitioned and utilized in favor of field reliability enhancement in this case.
- *Yield-Intensive Case* : In Figure 7 and 8, the dependability and its derivative of the given EAB are plotted with respect to s_{cm}. In this case, parameters are selected in order that more manufacturing defects than field faults are induced (i.e., $\lambda_m = 10^{-4} \rightarrow 2 \times 10^{-4}$). Thus,

partitioning result is [5,3]: 5 spare columns for factory repair and 3 spare columns for field repair. Spares are partitioned and utilized in favor of manufacturing yield enhancement in this case.

Note that, for the simulation results of the two dimensional redundancy case shown in Figure 9 - 14, similar observations can be extended and omitted in this paper.

6: Discussion and Conclusions

Among the cores for SoC integration, one of the most sensitive cores is the embedded memory core since memory cells are commonly considered as more prone to defects and faults than logic cells. Since cores cannot be physically replaceable once they are fabricated onto a SoC, combination of both ATE and BISR is commonly practiced. Proper partitioning and utilization of given shared redundancy is significantly desirable to achieve balanced manufacturing yield and field reliability of the embedded memory system core. Thus, yield and reliability assurance techniques have been initially proposed for the single dimensional redundancy case, then extended to two dimensional redundancy case. Since yield and reliability trade off each other, dependability (i.e., $Y \times R(t)$) reaches its maximum only if properly partitioned groups of the given redundancy are utilized to repair both manufacturing defects (i.e., ATE-based repair) and field faults (i.e., BISR-based repair). To effectively achieve the balanced redundancy partitioning and utilization, the dependability equations are differentiated and solved with respect to the number of spares used to enhance manufacturing yield. Parametric simulation results have been further verified that the proposed redundancy partitioning and utilization techniques for embedded memory system core achieves the theoretically optimal redundancy balancing. The proposed redundancy balancing techniques can be possibly incorporated with the existing CAD compilers for embedded memory system cores, thereby cost-effective partitioning and utilization of the shared redundancy can be realized.

References

[1] J. Ohtani, T. Ooishi, et al., "A Shared Built-In Self-Repair Analysis for Multiple Embedded Memories", *Custom Integrated Circuits, 2001, IEEE Conference on.*, pp. 187-190, May 2001.

[2] R.J. McPartland, D.J. Loeper et al, "SRAM Embedded Memory with Low Cost, FLASH EEPROM-Switch-Controlled Redundancy", *Custom Integrated Circuits Conference, 2000. CICC. Proceedings of the IEEE 2000* , Vol. 36, No. 11, pp. 287-289, May 2000.

[3] T. Kawagoe, J. Ohtani, et al., "A Built-In Self-Repair Analyzer (CRESTA) for Embedded DRAMs", *Test Conference, 2000. Proceedings. International*, Vol. 41, No. 9, pp. 567-574, Oct. 2000.

[4] International Technology Roadmap for Semiconductors. "International Technology Roadmap for Semiconductors 2000", *http://public.itrs.net/Files/2000UpdateFinal/2kUdFinal.htm*, 2000.

[5] D.K. Bhavsar, "An Algorithm for Row-Column Self-Repair of RAMs and its Implementation in the Alpha 21264", *Test Conference, 1999. Proceedings. International*, pp. 311-318, Sep. 1999.

[6] Low C.P. and Leong H.W., "A New Class of Efficient Algorithms for Reconfiguration of Memory Arrays", *IEEE Transactions on Computers*, Vol 45, No 5, pp. 614-618, 1996.

[7] D. M. Blough, "Performance Evaluation of a Reconfiguration-Algorithm for Memory Arrays containing Clustered Faults", *IEEE Transactions on Reliability*, Vol. 45, No. 2, pp. 274-284 June 1996.

[8] C.H. Stapper, H.-S. Lee, "Synergistic Fault-Tolerance for Memory Chips", *IEEE Trans. on Computers*, Vol. 41, No. 9, pp. 1078-1087, September 1992.

[9] S.Y. Kuo and W.K. Fucks, "Efficient Spare Allocation in Reconfigurable Arrays", *IEEE Design and Test*, Vol. 41, Issue. 9, pp. 24-31, Feb. 1987.

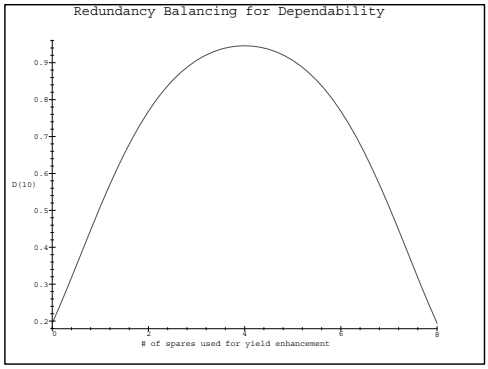

Figure 3. Symmetric 1D dependability graph

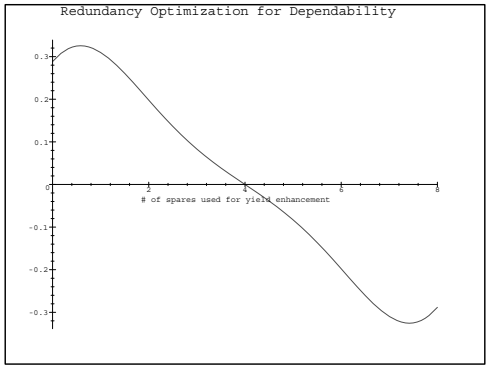

Figure 4. Balanced partitioning result [4,4]

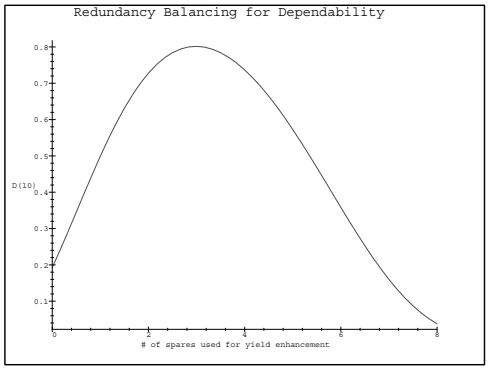

Figure 5. Reliability-intensive 1D dependability graph

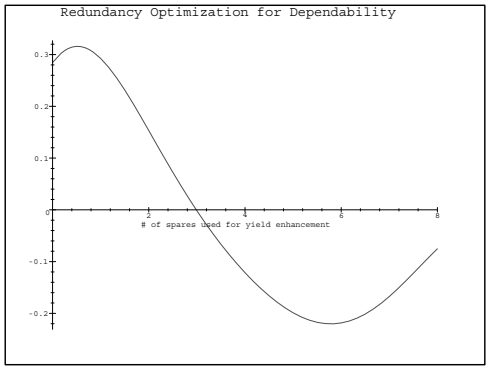

Figure 6. Balanced partitioning result [3,5]

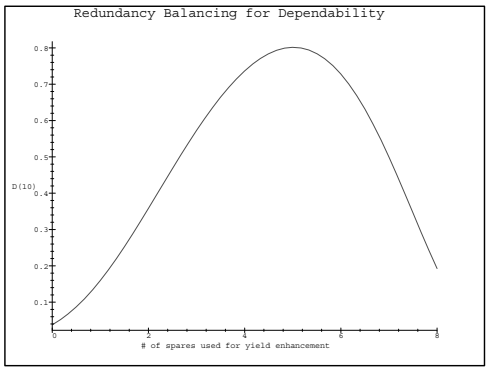

Figure 7. Yield-intensive 1D dependability graph

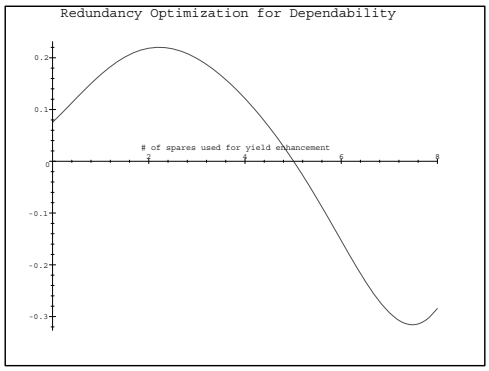

Figure 8. Balanced partitioning result [5,3]

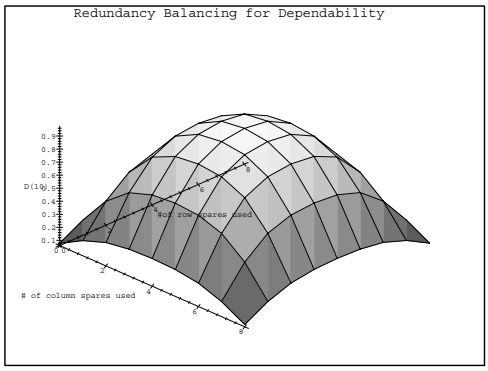

Figure 9. Symmetric 2D dependability graph

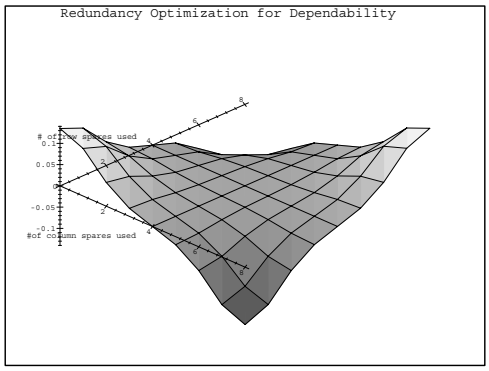

Figure 10. Balanced partitioning result $\{[4,4],[4,4]\}$

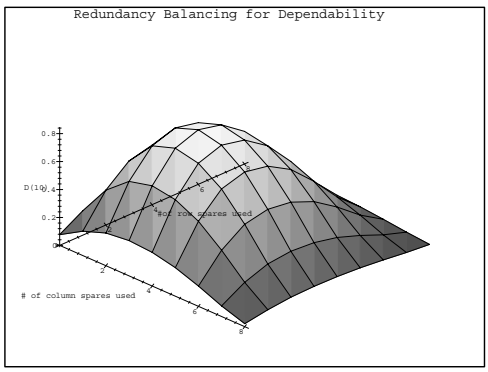

Figure 11. Reliability-intensive 2D dependability graph

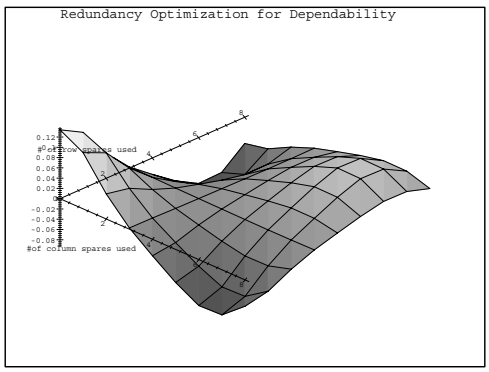

Figure 12. Balanced partitioning result $\{[3,5],[3,5]\}$

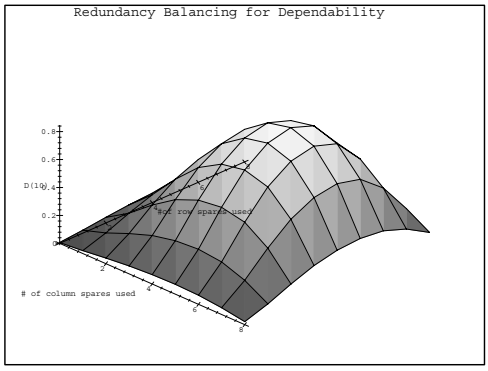

Figure 13. Yield-intensive 2D dependability graph

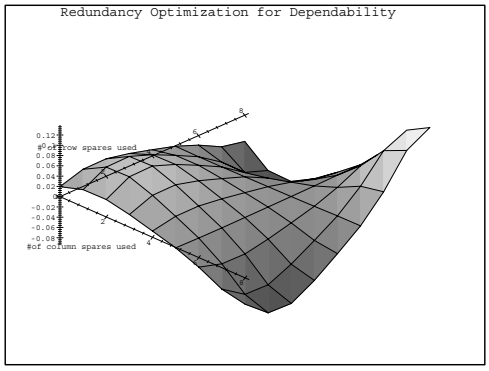

Figure 14. Balanced partitioning result $\{[5,3],[5,3]\}$