NEW EMBEDDED MEMORY ARCHITECTURE FOR ENHANCED YIELD, PERFORMANCE AND POWER CONSUMPTION

Boris Polianskikh and Zeljko Zilic

Department of Electrical and Computer Engineering, McGill University.
3480 University Street, Montreal, Quebec, H3A 2A7
Email: {borisp, zeljko}@macs.ece.mcgill.ca

A new Cross-Shared Redundancy (CSR) architecture of embedded memory for yield improvement is proposed. The model of CSR takes into account cluster errors, which are common for deep-submicron technologies. The redundancy scheme is optimized in consideration of low-power and fast operation. A yield model of cross-shared redundancy for the embedded memory is presented.

1. INTRODUCTION

With advances in deep-submicron CMOS technology, it is practical to create Systems-On-Chip (SOCs), where designer can place successfully many components on the same die [3]. SOCs provide a lot of flexibility, but engineers have to account for effects such as interference between digital and analog parts, yield and reliability. Proper function of the whole system depends on function of each block.

Since software gives more flexibility, distinguishing features of SOCs are often implemented in software, rather than in hardware. It is often more profitable to create hardware as small as possible and use instead a large software component. To accept these trends, engineers need more and more fast and reliable memory, which consumes less power. Embedded memories perfectly fit to these conditions. This paper investigates means to improve the yield of embedded memory, while retaining speed and power performance.

2. CROSS-SHARED REDUNDANCY

Architecture of the CSR model was designed in consideration of yield improvement and protection of the embedded memory against most important types of failures, such as single-cell, row, column and chip-kill failures. As shown in Fig. 1, on example of 4 Mbit memory, the memory core is organized as a square array of M independent blocks (M=16 in Fig.1).

Each block consists of 512 rows and 512 columns and has its own column and row decoders. There are also redundant columns and rows for fault tolerance, and BIST (Built-In-Self-Test) circuitry and *main* MC (Memory Controller) in the middle of the core. In addition to the main controller, there are four redundant memory controllers to protect memory from chip-kill or fatal defects. As known from practice, defects tend to occur as clusters and do not spread evenly over the chip [2],[6],[7]. For this model, if a cluster error occurs and destroys the memory controller, the defected controller can be easily replaced with one of the four spare controllers. This feature significantly augments yield of the whole chip (probability of chip kill defect reduced by power of 4), since memory functioning is very important for a SOC.

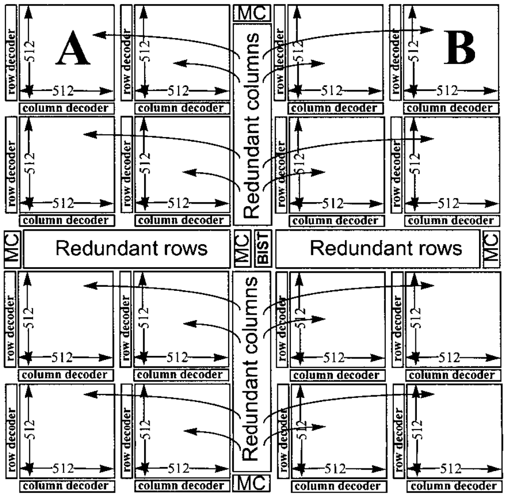

Fig. 1. Cross-Shared Redundancy memory

As seen from Fig. 1, redundant rows and columns are positioned in such a way, that redundant columns between blocks A and B may be used for both blocks depending on where the fault happens. The same is true for rows.

The same effect could be achieved just by placing the redundancy on the side and at the bottom of the core, but for speed and low-power consumption reasons, this model is more suitable. Power consumption is defined by equation $P_d = kV_{DD}^2 C_L f$, where C_L is load capacitance, k switching activity, V_{DD} is a power supply, f is a switching frequency. One way to reduce power consumption is to reduce load capacitance. For CSR model, by placing redundancy in the middle of the core, load capacitance of the access wire is reduced by a factor of 2. For example, considering the worst case, if there is a couple of destroyed columns in the left-top part of block **A** and there are redundant columns on the right side of the global core, each signal has to propagate through the whole memory and it needs to charge wire twice as much compared to our model, in order to activate neces-

0-7803-7057-0/01/$10.00 ©2001 IEEE.

sary operation.

Meanwhile, with redundancy in the middle of the memory core, the time for the signal to propagate will be reduced to half, and the wire to charge will be twice shorter.

3. FAILURE TYPES

As shown in [1], if all faults in memory were single-cell failures, the error-correcting code could bring yield up to 99.9%. However, most of these failures affect the chip support circuits, and the word and bit lines. For deep-submicron technology, this is even more true. Transistor sizes and, consequently, cell sizes become smaller. Chances that defect cluster affects only one cell, are very low. Normally, it is several cells, word and bit lines that are covered with a cluster failure. These types of defects can not be repaired by ECC (Error-Correcting Code) alone, due to 2-dimensional nature of cluster defects. It is essential to have sufficient row and column redundancy to repair these faults.

This model covers three major types of failures. Each type, in turn, consists of different subtypes of faults.

The first type is a single cell failure. This type of fault describes the situations when the defect is contained inside the cell. This can happen for several reasons.

1) Transistor damage happens more often for deep-submicron technologies.

2) Capacitor damage (very important for state-of-art 1T DRAMs), happens due to several reasons, such as absence of metal [5] or a short circuit to another metal layer. As shown in Fig. 2, cluster B damages only four cells and does not harm whole row or column. This type of fault is efficiently repaired with ECC, in order not to waste unnecessary cells of redundant rows or columns.

The second type is a row failure. If a cluster error breaks a word line close to the row decoder, the whole row does not function and can be replaced only with a redundant row. As shown in Fig. 2, cluster *A* destroys rows 5 and 6 completely and it is necessary to have two redundant rows *I* and *II* to repair such a damage. This type of faults happens for several reasons, such as bridging faults, cluster faults and many others.

The third type of faults is a column failure. It is known that columns are more susceptible to failure than rows [1]. There are several reasons why column failure happens. They can be classified in following way:

a) one or both bit lines are damaged;

b) precharge circuitry failure;

c) sense amplifier failure.

This type is especially important because the very sensitive and precisely tuned-up sense amplifiers have to sense shrinking voltage levels, as the technology shrinks.

For this type of failures, ECC is not a suitable solution, because it is impossible to repair sense amplifiers and the only acceptable solution is to replace the column, which contains a damaged sense amplifier, with the redundant one.

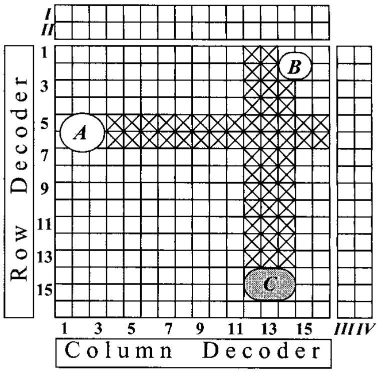

Fig. 2. Embedded memories major types of failures

4. YIELD MODELLING

Since redundancy allocation is NP-complete problem it is impossible to describe exact allocation of redundant columns and rows. The model presented below describes rather approximate outcome of spare allocation. Since all three types of failures are independent and can happen at the same time, yield is obtained if their probabilities are multiplied [9]. Our model uses combined Poisson and Binomial distributions for yield modelling [6]. The Poisson distribution is used to describe the yield of a single cell and the whole core, in case when there is no redundancy. Binomial distribution describes the yields of individual columns and rows. The yield of one cell, i. e. probability that one cell functions correctly, is given by equation (1)

$$P(cell) = \lambda_{cell} = e^{-(A_{cell} \cdot D_0)} \qquad (1)$$

where A_{cell} is the area of one cell and D_0 is the defect density, which depends on process variations and process conditions. D_0 is defined on the basis of empirical data for a specific process. There are three possible cases to consider to obtain memory yield.

4.1 Memory without redundancy

In this case, memory is not protected with redundancy and can not tolerate any damage. In this case, yield is expressed as probability of not having any faults in the memory core. Since there are N_{row} times N_{col} cells in memory block (where N_{row} is the number of rows and N_{col} is the number of columns), probability that a block functions without any faults will be expressed by equation(2)

$$Y_1 = e^{-(A_{cell} \cdot D_0 \cdot N_{col} \cdot N_{row})} \qquad (2)$$

There are *M* blocks in the core and there is equal possibility for any of them to be damaged, the yield of the whole core is going to be as in equation (3)

$$Y_{wor} = (Y_1)^M \qquad (3)$$

where Y_{wor} stands for *Y*ield *w*ith*o*ut *r*edundancy.

4.2 Memory protected only with redundant rows or columns

In this case, memory can tolerate only one type of damage: either row or column, but not both. Normally, this is a column redundancy because columns are more susceptible to failure. In this case, yield is expressed as the probability of having a certain number of failed columns or rows. Yield is described with combined Poisson and Binomial distributions. Yield of one column is expressed with Poisson distribution and the event of damage happening in several columns out of total number of columns is described with Binomial distribution. Since the number of cells per column is N_{row} , probability that one column functions is defined by equation (4)

$$\lambda_{col} = e^{-(A_{cell} \cdot D_0 \cdot N_{row})} \quad (4)$$

Probability of having c failed columns out of N_{col} is given by Binomial distribution, equation(5)

$$P(c\ \ failed\ \ columns) = \binom{N_{col}}{c}(1-\lambda_{col})^c \lambda_{col}^{(N_{col}-c)} \quad (5)$$

Now consider assumption that there are c failed columns that may be replaced with redundant columns. As is shown in Fig. 3, any cell in redundancy-protected columns c can be repaired for any damage, but if one of the rows of length N_{col}-c is damaged, the memory can not function properly already. Since no row redundancy is employed, memory can not tolerate any row damage and 1D-redundancy for one block is described by equation (6)

$$Y_2 = \sum_{c=1}^{R_{col}} \binom{N_{col}}{c}(1-\lambda_{col})^c \lambda_{col}^{(N_{col}-c)} \lambda_{row}^{N_{row}} \quad (6)$$

where λ_{row} , the probability that one row of size N_{col}-c will function is given by Equation (7) and R_{col} is the number of redundant columns

$$\lambda_{row} = e^{-(A_{cell} \cdot D_0 \cdot (N_{col}-c))} \quad (7)$$

For function memory core only several outcomes are possible: either there are no faults in the core and this event is described with equation(2), or there are several faults but not more than number of redundant columns and these events are described with equation(6). Since all these events are mutually exclusive, the yield of the whole chip protected with 1D-redundancy can be given as in equation(8), where Y_{1D} stands for *y*ield *w*ith *1D*-*r*edundancy.

$$Y_{1D} = (Y_1 + Y_2)^M \quad (8)$$

4.3 Memory protected with redundant rows and columns

Now, memory is able to tolerate 2D-cluster damage. The yield of any of M blocks is

$$Yield = P(no\ \ faults) + P(1D) + P(2D)$$
$$Y = Y_1 + Y_2 + Y_3 \quad (9)$$

where Y_1 describes yield in case when memory core is free of damage, Y_2 describes yield in case when memory is damaged and it is possible to repair the damage only with redundant columns or rows, Y_3 is the part that describes failures which may be repaired only with redundant columns and rows together given by equation (10).

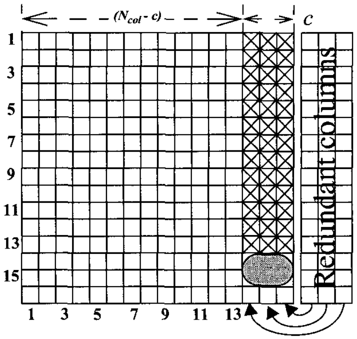

Fig. 3. Effect of redundancy on yield

The double summation in Y_3 goes through all possible combinations, i. e. 1 column+1 row, 1 column+2 rows, 2 columns+1 row, 2 columns+2 rows and so on, until all redundant columns and rows are exhausted.. For example, if there are three redundant columns and four rows (R_{col} =3, R_{col} =4), the model will sum all 12 possible situations which memory can tolerate, with P (c col, r row) denoting probability of having c damaged columns and r damaged rows in the memory core.

$$Y_3 = \sum_{c=1}^{R_{col}} \sum_{r=1}^{R_{row}} \binom{N_{col}}{c}\binom{N_{row}}{r}(1-\lambda_{col})^c \cdot \lambda_{col}^{(N_{col}-c)}(1-\lambda_{row})^r \lambda_{row}^{(N_{row}-r)} \quad (10)$$

Since the memory is divided in M blocks and yield of one block is given by equation (9) the yield of the memory will be

$$Y_{2D} = (Y_1 + Y_2 + Y_3)^M$$

where 2D stands for two dimensional redundancy.

5. RESULTS

As is shown in Fig. 5, the curve *wr* represents yield of the CSR without redundancy, curves *1c, 2c* and *15c* represent models with 1D-redundancy 1, 2, 15 redundant columns and curves *1c+1r, 2c+2r* represent models with two dimensional redundancy with 1 column and 1 row, and 2 columns and 2 rows respectively.

Comparing models with 2 redundant columns (1D-redundancy) and 1 redundant column + 1 redundant row (2D- redundancy), it is obvious that yield is better for two dimensional redundancy. The case, when there are 4 redundant columns (1D- redundancy) compared to 2 redundant columns + 2 redundant rows (2D- redundancy) gives even better results for yield. It is clearly seen from these results that the yield is much better for two dimensional redundancy than for one dimensional, not because of the quantity of the redundancy, but because of the quality of it. One might also notice that the area in space between curves *4c* and *2c+2r* is much larger than area in space between curves *2c* and *1c+1r*. This shows that even little increase in quantity of two dimensional redundancy results in dramatical improvement of yield. Two dimensional redundancy repairs much more defect types than one dimensional redundancy with the same space occupied.

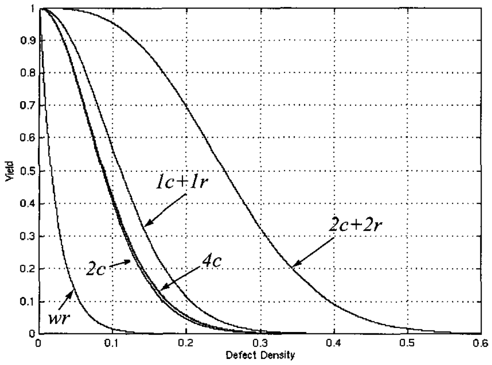

Fig. 4. One dimensional versus two dimensional redundancy.

Another interesting observation is that after certain value, the yield of the model with single redundancy converges to some point and does not improve further. As is shown in Fig. 6, comparing curves *wr, 1c, 2c, 4c, 15c*, yield improvement eventually stops when approximately 4 redundant columns are employed and curves *4c* and *15c* converged to the same line, and are seen as one single line with this scale. This happens because single redundancy alone is not sufficient to repair all types of faults in a core. Increasing the redundancy, which is not able to repair certain defects gives nothing, and memory is still going to fail.

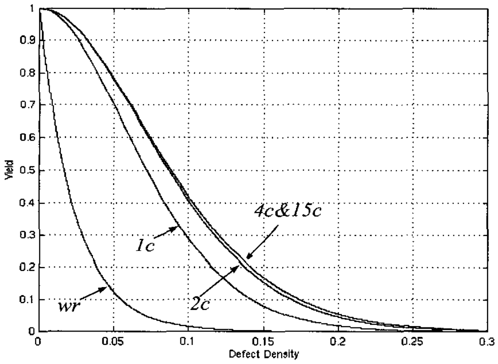

Fig. 5. Limitations of one dimensional redundancy

6. CONCLUSIONS

Selection of suitable redundancy (number of redundant columns, rows, combination of ECC and 2D-redundancy, redundant memory controllers) for specific memory model can significantly increase the yield of embedded memory. Comparing to other existing models [2], [4], [8], [10] the CSR model proves that in the future deep-submicron technologies yield might be improved without excessive performance penalty.

The CSR model achieves better yield and performance due to cross-shared redundancy positioning and additional memory controllers. Power consumption is reduced because of the placement of redundant columns and rows, which allows to charge less wires to access necessary cells.

REFERENCES

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

[2] Z.Horen and I.Horen, "A model for enhanced manufacturability of defect tolerant integrated circuits", *Proc. Defect and Fault Tolerance on VLSI systems, 1991, Pages: 81-92.*

[3] Y. Zorian, "Yield Improvement and Repair Trade-Off For Large Embedded Memories", *Proc. Conf. Design, Automation and Test in Europe, 2000, Page(s): 69-70.*

[4] S. M. Domer, S. A. Foertsch and G. D. Raskin, "Model for Yield and Manufacturing Prediction on VLSI Designs for Advanced Technologies, Mixed Circuitry and Memories", *IEEE Journal of Solid-State Circuits, Volume: 303, March 1995, Page(s): 286-294.*

[5] S. Gandemer, B. C. Tremintin and J.-J. Charlot, "Critical Area and Critical Levels Calculation in I.C. Yield Modeling", *IEEE Trans. on Electronic Devices, Vol: 35, No. 2, February 1988.*

[6] J. A. Cunningam, "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing", *IEEE Trans. on Semiconductor Manufacturing, Vol.: 32, May 1990, Page(s): 60-71.*

[7] G. Battaglini and B. Ciciani, "An Improved Analytical Yield Evaluation Method for Redundant RAM's", *Proc. Int. Workshop on Memory Technology, Design and Testing, 1998, Page(s): 117-123.*

[8] K. N. Ganapathy, A. D. Singh, D. K. Pradhan, "Yield Modeling and Optimization of Large Redundant RAM's", *Proc. [2nd] Int. Conf. on Wafer Scale Integration, 1990, Page(s): 273-287.*

[9] K.-I. Imamiya, J.-I. Miyamoto, N. Ohtsuka, N. Tomita and Y. Iyama, "Optimum Redundancy Design for New-Generation EPROM's Based on Yield Analysis of Previous Generation", *VLSI Test Symp. on Design, Test and Application: ASIC's and Systems-On-Chip', 1992. Page(s): 182-187.*

[10]M. Rudack and D. Niggemeyer, "Yield Enhancement Considerations for A Single-Chip Multiprocessor System with Embedded DRAM", *Int. Symp. on Defect and Fault Tolerance in VLSI Systems, 1999. Page(s): 31-39.*