Analysis of Flexible Redundancy Techniques for Multi-Bank Memory ICs

Israel Koren and Zahava Koren

Department of Electrical and Computer Engineering
University of Massachusetts, Amherst, MA 01003
E-mail: koren@euler.ecs.umass.edu

Incorporating defect tolerance for yield enhancement of memory ICs through redundant rows and columns has been extremely successful for more than 20 years. A new approach requiring a smaller design modification was recently implemented in [1], and is depicted in Figure 1. We present a yield analysis of this approach and demonstrate the effect of the different system parameters on the yield of the chip.

The memory IC consists of 16 banks, each of which has its own spares. Traditionally, each spare line (row or column) has its own set of fuses which are programmed so that it can replace a defective line. These fuses are designed with a much larger feature size than the memory cells and as a result, a spare line with its associated fuses consumes a much higher silicon area than a regular memory line. The design in Figure 1 separates the spare lines from the sets of fuses and makes the fuse sets globally available to all 16 banks. This allows for the use of a larger number of local spare lines with fewer fuse sets, thus achieving the yield benefits with a much reduced area penalty.

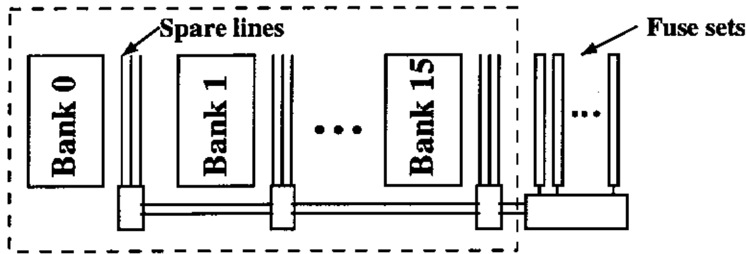

Figure 1: A block diagram of the design in [1].

Our yield analysis is based on two widely-used analytical fault models: The Poisson distribution and the negative binomial distribution [2]. We use the following notations:
λ - average number of correctable faults per row, α - clustering parameter (see [2]), r - number of rows per bank, s - number of spare rows per bank, b - number of banks, t - total number of fuse sets for all banks, Y_0 - probability of no systematic (uncorrectable) faults.

$$Chip_Yield = Y_0 \sum_{\substack{f_1+\cdots+f_b\le t \\ f_1,\ldots,f_b\le s}} \prod_{i=1}^{b} Prob\{f_i \; faulty \; rows \; in \; bank \; i\} \qquad (1)$$

Figure 2 depicts the yield as a function of t - the number of fuses - for the two distributions, with $\lambda = 0.0001$, $s = 4$, and $\alpha = 0.5$, and shows that the yield reaches its maximum at about $t = 14$,

much lower than $b \times s = 16 \times 4 = 64$, and that the optimal t is not very sensitive to α.
In Figure 3, the optimal t was calculated for two distributions with the same α but different λ's. The best t depends on λ, and it increases with λ.
Figure 4 depicts the maximum attainable yield as a function of s and shows that the optimal s depends on both λ and α. A smaller α (i.e., more clustering) and a larger λ would require a higher s to maximize the yield.

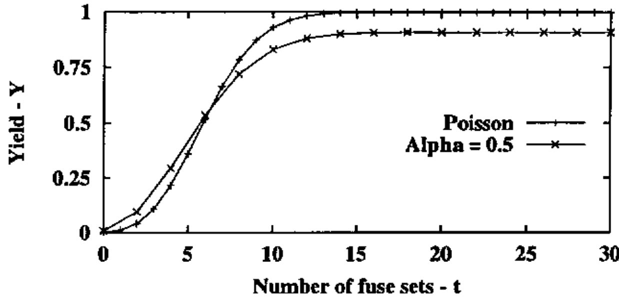

Figure 2: Yield vs. number of fuse sets for two distributions ($\lambda = 0.0001,\ s = 4$).

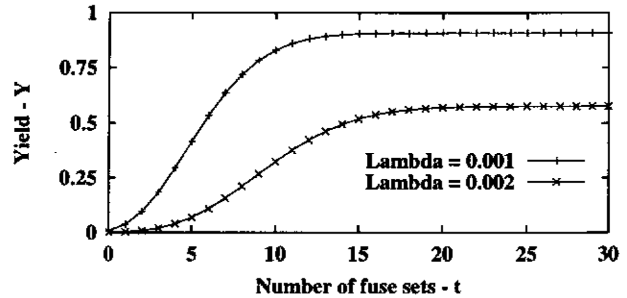

Figure 3: Yield vs. number of fuse sets for two values of λ ($\alpha = 0.5$).

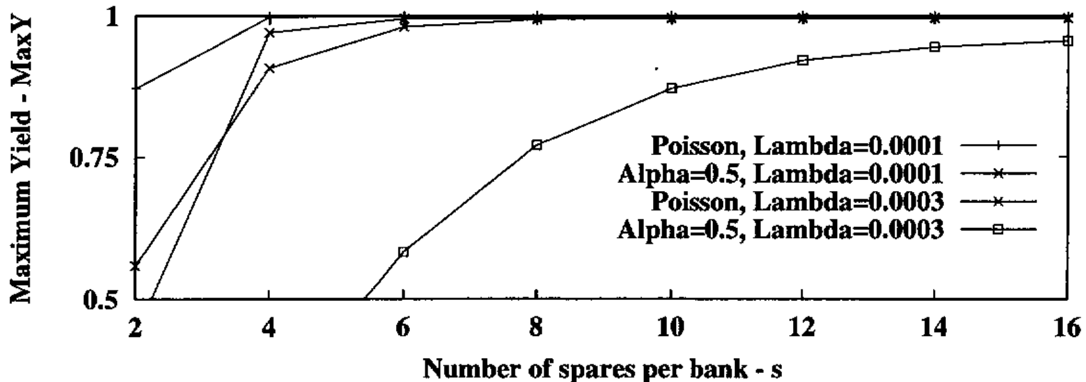

Figure 4: Max. yield vs. number of spares per bank for several distributions.

References

[1] S. Takase and N. Kushiyama, "A 1.6GByte/s DRAM with Flexible Mapping Redundancy Technique and Additional Refresh Scheme," *IEEE J. of Solid-State Circuits*, vol. 34, pp. 1600-1606, Nov. 1999.

[2] I. Koren and Z. Koren, "Defect Tolerant VLSI Circuits: Techniques and Yield Analysis," *Proceedings of the IEEE*, Vol. 86, pp. 1817-1836, Sept. 1998.