Yield Analysis of a Novel Scheme for Defect-Tolerant Memories[1]

Israel Koren and Zahava Koren

Department of Electrical and Computer Engineering
University of Massachusetts, Amherst, MA 01003
E-mail: koren@euler.ecs.umass.edu; FAX: (413)545-1993

Abstract

The recent increases in the size of memory ICs have made designers realize that there exists a need for new defect-tolerance techniques, since the traditional methods are no longer effective. One such new technique, the *Flexible Multi-Macro* (FMM) technique, has recently been suggested and implemented in a 1 Gb DRAM circuit. In this paper we present a yield analysis of the FMM design and compare its yield to that of the most common defect-tolerance technique of adding spare rows and columns to the memory array.

1. Introduction

The traditional method for incorporating defect-tolerance in memory ICs through redundant rows and columns has been extremely successful for more than 15 years. This technique has even been incorporated in the design of large cache units in microprocessors in the last five years. The advantage of employing redundant rows and columns has been especially significant in the early stages of production when the yield is still low, allowing for earlier introduction of new products into the market.

Further increases in the size of the memory array made it necessary to partition the array into several sub-arrays in order to decrease the current and reduce the access time by shortening the length of the bit and word lines [5]. Using the conventional redundancy methods implied that each sub-array should have its own redundant rows and columns, leading to situations where one sub-array had an insufficient number of spare lines to handle local defects while other sub-arrays still had several unused redundant lines.

As memory ICs become denser, the sub-micron process technology becomes more complex and the manufacturing yield is expected to decrease [5]. As a result, defect-tolerance techniques are important not only in the early stages of the production but also in the mass production stages. It becomes apparent, therefore, that new and more efficient redundancy techniques must be developed. One obvious approach is to turn some (or even all) of the local redundant lines into global redundant lines, allowing for a more efficient use of the redundant lines at the cost of higher silicon area overhead due to the larger number of required programmable fuses. This approach has been followed in [5], where the design of an experimental 4 Mb SRAM was presented. A 3% increase in the area overhead and up to 61% increase in effective yield have been reported there.

A different approach was presented in [4]. Here, fewer (compared to the traditional technique) redundant lines were used and they remained local. For added defect-tolerance, the individual sub-array (called Macro in [4]) was fabricated in such a way that it could become part of up to four different memory ICs. The proposed technique was

[1]This work was supported in part by NSF, under contract MIPS-9305912.

0-7803-3639-9/96 $4.00

named the Flexible Multi-Macro (FMM) technique and was applied to a 1 Gb DRAM in 0.25 μm CMOS technology.

In what follows we present a yield analysis of the FMM design, and compare its yield to that of the conventional defect-tolerant wafer design using spare rows and columns.

2. Yield Analysis

The 1 Gb DRAM described in [4] was partitioned into four 256 Mb modules (macros). Had the researchers adopted the traditional technique of adding redundant rows and columns to each module, the 8" wafer would contain 24 ICs, as shown in Figure 1. To implement the proposed FMM technique allowing each module to be included in any one of up to four ICs, the area of the basic module had to be increased by 2%. In order to keep the overall area of the module identical to that in the conventional design, row redundancy was eliminated, thus saving about 2% of the total area. The column redundancy was still implemented, consuming approximately a 2.5% overhead. The resulting floorplan of the wafer is depicted in Figure 2. The chip boundaries are not shown, since these are determined only after testing all modules and partitioning into subsets, each consisting of the four modules needed for a 1 Gb DRAM IC. Note that since the chip boundaries are not predetermined, four additional modules were fabricated at each corner (for a total of 16 modules) allowing further flexibility in combining modules to form ICs.

The yield of the designed memory chip was estimated in [4] using a Monte-Carlo simulation. Estimating the yield through simulation is always time-consuming (compared to analytical alternatives), making a complete analysis of any proposed scheme very costly.

We present a yield comparison of the conventional wafer design and the FMM design, based on three widely-used analytical fault models: The Poisson distribution, and the large-area and medium-area clustering negative binomial distributions [2]. We distinguish in our analysis between defects and faults - a fault being a defect which actually affects the proper operation of the chip, thus reducing the yield. We ignore defects which do not turn into faults.

The two yield measures which can be used for comparing different chip designs are G, the expected number of operational (good) chips out of a wafer, and Y, the expected proportion of operational chips out of a wafer. The two measures are related through N - the maximum number of chips that can be extracted out of a wafer:

$$Y = \frac{G}{N} \tag{1}$$

We denote by $G^{(c)}, Y^{(c)}, N^{(c)}$ and by $G^{(f)}, Y^{(f)}, N^{(f)}$ the values of the above-defined measures for the conventional chip and the FMM chip, respectively. We deal with the conventional chip first. Since the boundaries of the chips on the wafer are fixed, $Y^{(c)}$ can be calculated as the probability that a given chip is operational, or that a fixed area of four adjacent modules is operational. Thus,

$$N^{(c)} = 24$$

and

$$G^{(c)} = 24 \cdot Y^{(c)} \tag{2}$$

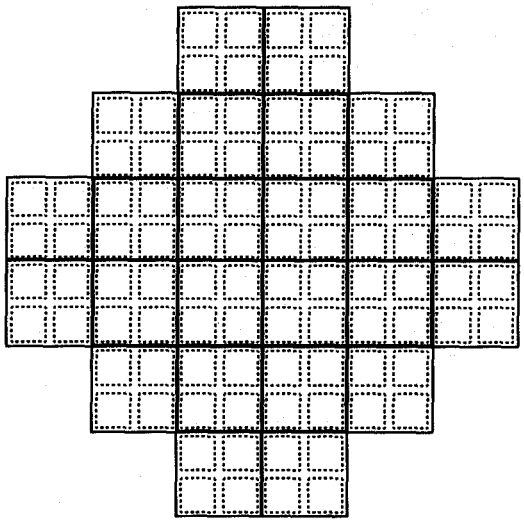


Figure 1: An 8" wafer containing 24 1Gb DRAMs with conventional redundancy.

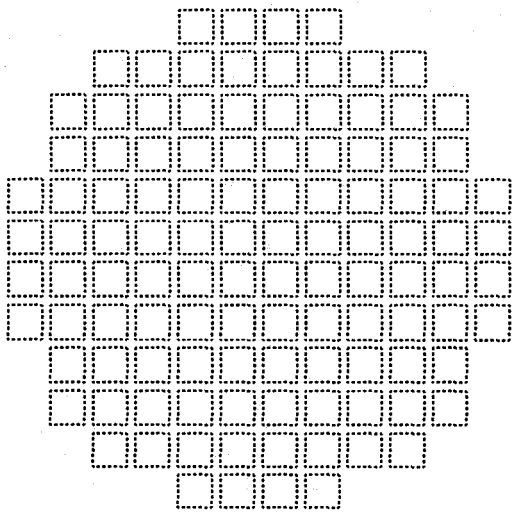


Figure 2: An 8" wafer containing 112 256-Mb macros.

To calculate $Y^{(c)}$, we first calculate $Y_M^{(c)}$ - the probability that a selected module, out of the four in the chip, is operational. Denote by n the number of columns (or rows) in a module and by d the number of spare columns (which is equal to the number of spare rows). Since d is very small compared to n, we can assume that all $2 \cdot d$ spares are columns. In addition, the module has a chip-kill area in which a fault cannot be recovered from by redundancy and is, therefore, fatal. Denoting by $L^{(c)}$ the fault density (average number of faults) per column, the fault density for the chip-kill area is $K \cdot L^{(c)}$, where K is the ratio between the chip-kill area and the area of a column. The probability of an operational module is the probability that no more than $2d$ of its columns are faulty and that the chip-kill area is completely fault-free. Assuming the Poisson distribution for the faults on the chip, we obtain

$$Y_M^{(c)} = Prob(a\ given\ module\ is\ operational) \tag{3}$$

$$= \sum_{j=0}^{2d} \binom{n+2d}{j} \left(1 - e^{-L^{(c)}}\right)^j \left(e^{-L^{(c)}}\right)^{n+2d-j} e^{-KL^{(c)}}$$

$$= \sum_{j=0}^{2d} \binom{n+2d}{j} \left(1 - e^{-L^{(c)}}\right)^j e^{-(n+2d-j+K)L^{(c)}}$$

The probability of an operational chip is

$$Y^{(c)} = (Y_M^{(c)})^4 \tag{4}$$

As is well-known by now [3], the negative binomial distribution can be obtained by compounding the Poisson distribution, which means averaging over the parameter $L^{(c)}$.

Assuming that $L^{(c)}$ is a random variable with a $Gamma(\alpha, \frac{\alpha}{\lambda})$ density function, i.e.,

$$f(L) = \frac{\alpha^\alpha}{\lambda^\alpha \Gamma(\alpha)} L^{\alpha-1} e^{-\frac{\alpha}{\lambda}L} \tag{5}$$

and integrating the Poisson probability function with respect to this density, yields the negative binomial distribution with a fault density of λ and a clustering parameter of α. The equations resulting from integrating (3) and (4) with respect to (5) are very complicated and will not be presented here. However, we need to emphasize that both the large area clustering and the medium area clustering distributions can be obtained using the procedure of compounding the Poisson distribution. The difference is in the order of integrating and raising to the 4th power in (4). If $Y_M^{(c)}$ is integrated and the result raised to the 4th power, we get the medium area negative binomial distribution, with clusters similar in size to the module. If, on the other hand, $(Y_M^{(c)})^4$ is calculated first and the result is integrated, we get the large area negative binomial distribution.

The yield analysis of the FMM wafer is more complicated. Since the boundaries of the chips are flexible and each operational module can be included in more than one chip, the yield $Y^{(f)}$ is not equal to the probability that a given area of size $2 \cdot 2$ modules is operational. Instead, the inclusion and exclusion formula needs to be used. Let $g^{(f)}$ be the number of (disjoint) operational chips that can be extracted out of an FMM wafer,

and let $I_{i,j}$ denote the number on the wafer of (not necessarily disjoint) operational rectangles consisting of $i \times j$ modules. Note that $g^{(f)}$ is a random variable, and that $G^{(f)}$ defined above is its expected value. Then,

$$g^{(f)} = \sum_{i=2}^{12}\sum_{j=2}^{12}(-1)^{i+j} I_{i,j} \tag{6}$$

or

$$g^{(f)} = I_{2,2} - I_{2,3} - I_{3,2} + I_{3,3} + \cdots$$

and

$$G^{(f)} = E(g^{(f)}) = \sum_{i=2}^{12}\sum_{j=2}^{12}(-1)^{i+j} E(I^{(i,j)}) \tag{7}$$

where $E(X)$ denotes the expected value of the random variable X.

Since the $I_{i,j}$ rectangles are not necessarily disjoint, it is easy to calculate $E(I_{i,j})$. Denoting by $N_{i,j}$ the number of (not necessarily disjoint) $i \times j$ rectangles in a wafer,

$$E(I_{i,j}) \;=\; N_{i,j} \cdot Prob(a\ given\ i \times j\ rectangle\ is\ operational) \tag{8}$$

We calculate this last probability similarly to the way $Y^{(c)}$ has been calculated earlier. Assuming the Poisson distribution for the faults on the wafer, we first calculate the probability of an operational module. Due to the overhead required for the added flexibility in the new design, the redundancy has been reduced to d columns and the chip-kill area has been increased by a factor of R ($R > 1$). Therefore,

$$Y_M^{(f)} = Prob(a\ module\ is\ operational) \tag{9}$$

$$= \sum_{j=0}^{d} \binom{n+d}{j} \left(1 - e^{-L^{(f)}}\right)^j \left(e^{-L^{(f)}}\right)^{n+d-j} e^{-RKL^{(f)}}$$

$$= \sum_{j=0}^{d} \binom{n+d}{j} \left(1 - e^{-L^{(f)}}\right)^j e^{-(n+d-j+RK)L^{(f)}}$$

where n is the number of columns, d is the number of redundant columns and $L^{(f)}$ is the fault density of a column in the new chip. Now,

$$Prob(an\ i \times j\ rectangle\ is\ operational) = \left(Y_M^{(f)}\right)^{ij} \tag{10}$$

and substituting (10) into (8) and then into (7), results in

$$G^{(f)} = \sum_{i=2}^{12}\sum_{j=2}^{12}(-1)^{i+j} N_{i,j} \left(Y_M^{(f)}\right)^{ij} \tag{11}$$

and since $N^{(f)} = 26$,

$$Y^{(f)} = \frac{G^{(f)}}{26} \tag{12}$$

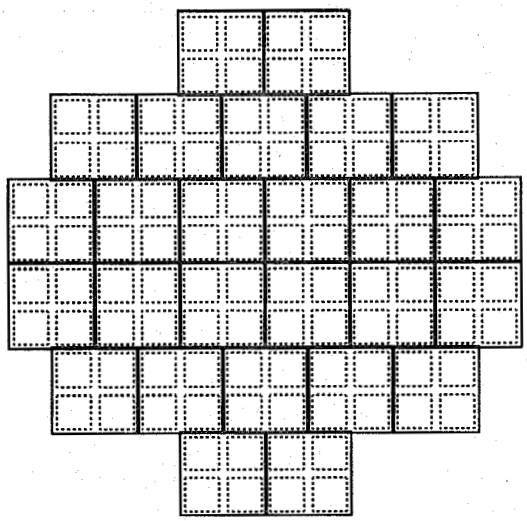


Figure 3: An 8" wafer containing 26 1Gb DRAMs.

Again, to obtain the yield for the negative binomial distribution, we average over the parameter $L^{(f)}$ with respect to the *Gamma* distribution (5) with the parameters λ_f and α_f, the the average number of faults per column and the clustering parameter, respectively, for the FMM chip. We will get results for the large area model or for the medium area model, according to whether the raising to the ijth power in (10) precedes the integration or vice-versa.

In comparing the yield of the two chip designs, we need to take into account several factors: The FMM chip has less redundancy per individual module and a larger chip-kill area, but on the other hand has a much higher flexibility in combining operational modules into complete chips. Another point that needs to be considered is that the FMM design has more modules on the wafer, and there is, therefore, a potential of 26 chips that can be extracted out of one wafer compared to the maximum number of 24 chips in the conventional design.

A question that needs to be asked in comparing the two designs is which of the two measures, G or Y, should be used. The two comparisons are not identical, since $Y^{(c)}$ is calculated by dividing $G^{(c)}$ by 24, while $Y^{(f)}$ is calculated by dividing $G^{(f)}$ by 26. Comparing the Gs reflects the full advantage of the new design, while comparing the Ys isolates the effect of the flexibility of the FMM chip, and ignores the fact that the wafer has a potential of 26 good chips (or assumes that even in the conventional design we can obtain 26 chips per wafer by adding 8 extra modules in the corners as depicted in Figure 3).

Since $\frac{Y^{(f)}}{Y^{(c)}} = \frac{G^{(f)}}{G^{(c)}} \cdot \frac{24}{26}$, it is possible to find a case in which $G^{(f)} > G^{(c)}$ but $Y^{(f)} < Y^{(c)}$, indicating that the flexibility of the FMM design alone is not enough to offset the decrease

in redundancy and increase in the chip-kill area. In the numerical results which follow, we show one example of comparing both the Gs and the Ys, and then proceed to compare the Gs, thus capturing the full impact of the FMM design.

3. Numerical Results

The main system parameters which need to be considered in the numerical analysis are n - the number of columns (and rows), d - the number of redundant columns (and rows, in the conventional design), λ_c and λ_f - the fault densities per column in the conventional and the FMM designs, respectively, α_c and α_f - the corresponding clustering parameters, K - the ratio between the chip-kill area and the column area in the conventional design, and R - the increase in the chip-kill area for the FMM design, compared to the conventional design.

Based on the information provided in [4], $n = 2^{14}$ and d is about 2% of n. We can assume that $\lambda_c = \lambda_f = \lambda$ and $\alpha_c = \alpha_f = \alpha$, and denote $\lambda_k = K\lambda$. We calculated the values of G and Y for the two chip designs, for several values of the parameters λ, α, K and R, and for the three fault distributions - Poisson, medium area negative binomial and large area negative binomial. The results are depicted in Figures 4 - 6.

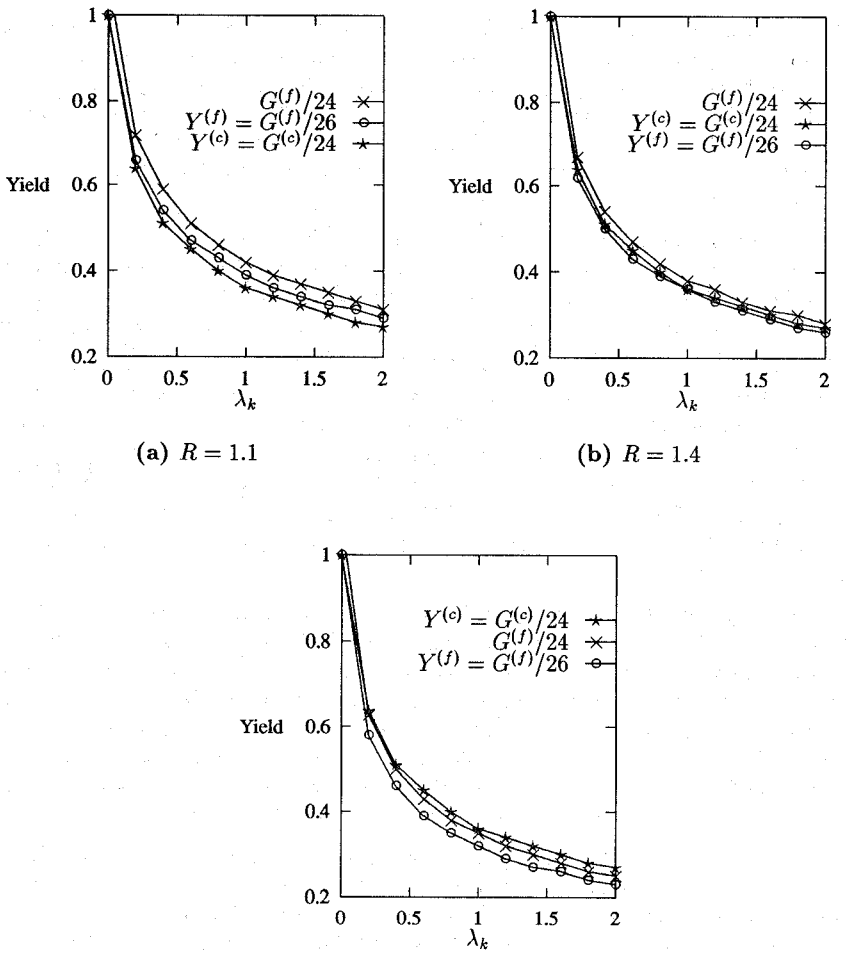
Figure 4 compares $Y^{(c)} = G^{(c)}/24$, $Y^{(f)} = G^{(f)}/26$, and $G^{(f)}/24$ as functions of λ_k, for the large area negative binomial distribution with $\lambda = 1$, $\alpha = 0.25$, and three values of R - 1.1, 1.4 and 1.8. As seen in Figure 4(a), if the chip-kill area is increased by only 10%, both $G^{(f)} > G^{(c)}$ and $Y^{(f)} > Y^{(c)}$. Figure 4(b) shows that if the chip-kill area is increased by 40%, $G^{(f)} > G^{(c)}$ but $Y^{(f)} < Y^{(c)}$, while if it is increased by 80%, we see in Figure 4(c) that both $G^{(f)} < G^{(c)}$ and $Y^{(f)} < Y^{(c)}$. Similar results are obtained for the other two distributions.

Figure 5 compares $G^{(c)}/24$ and $G^{(f)}/24$ as a function of λ_k, for $\alpha = 0.25$, $R = 1.3$ and the three fault distributions. We see in Figures 5(a), 5(b) and 5(c) that for the Poisson distribution there is practically no advantage in using the new design, there is a small improvement when the large area negative binomial distribution is assumed, and the largest increase in yield occurs for the model using the medium area negative binomial distribution. A possible explanation to this phenomenon is that in this last model we are more likely to find operational areas of size 2×2 modules (rather than smaller or larger fault-free areas), enabling us to use the flexibility provided by the FMM design.

Figure 6 depicts $G^{(c)}/24$ and $G^{(f)}/24$ as a function of R, for $\lambda = 1$, $K = 0.2$, the medium area negative binomial distribution, and two values of α - 0.25 and 5. Figures 6(a) and 6(b) show that the larger improvement in yield when using the FMM design occurs for the smaller value of α (the range of R values for which the FMM design is better is larger for $\alpha = 0.25$). The lower α indicates higher clustering, and again, we are more likely to encounter medium sized fault-free areas on the wafer which can be used by the FMM design.

4. Conclusion

Designers of memory ICs have recently realized that the traditional method for incorporating defect-tolerance (through redundant rows and columns) is no longer effective in sub-micron process technologies. Consequently, new techniques for defect-tolerance have been recently proposed and implemented. In this paper we analyzed one such design

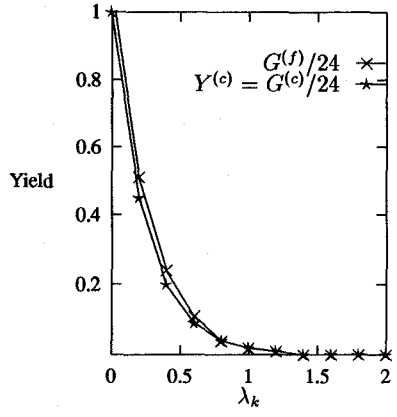


(a) $R = 1.1$

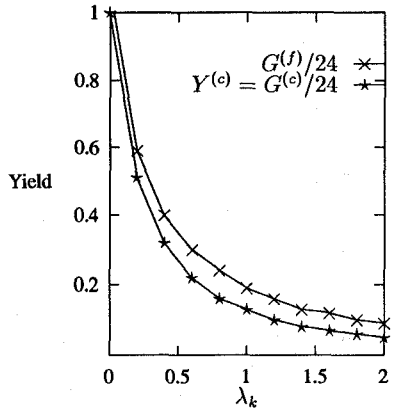
(b) $R = 1.4$

(c) $R = 1.8$

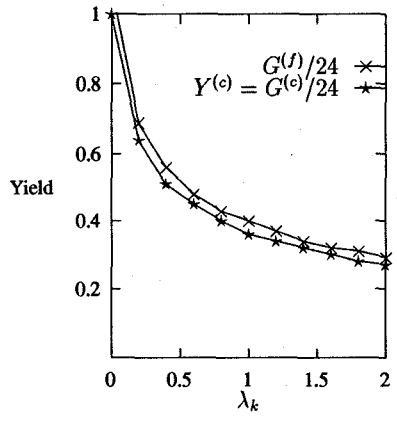
Figure 4: Yield as a function of λ_k, for the large area negative binomial distribution, $\alpha = 0.25$, and three values of R.



(a) The Poisson distribution.

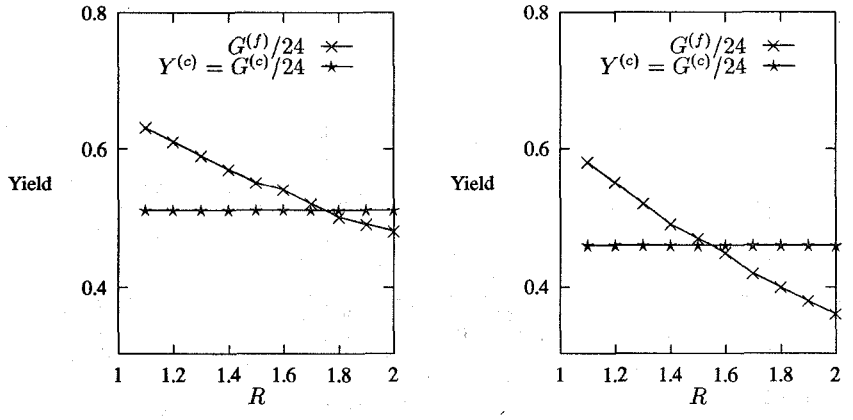


(b) The medium area negative binomial distribution.



(c) The large area negative binomial distribution.

Figure 5: Yield as a function of λ_k, for $\alpha = 0.25$, $R = 1.3$ and three fault distributions.



(a) $\alpha = 0.25$. **(b)** $\alpha = 5$.

Figure 6: Yield as a function of R, for $\lambda_k = 0.2$, the medium area negative binomial distribution, and two values of α.

(the FMM chip [4]) and our most important conclusion is that the advantage of the new technique over the traditional one cannot be guaranteed. A very careful yield analysis must be performed since, depending on the system parameters, the new design can have a higher or a lower yield than the conventional design.

References

[1] G. Kitsukawa *et al.*, "256-Mb DRAM Circuit Technologies for File Applications," *IEEE J. of Solid-State Circuits*, vol. 28, pp. 1105-11101 Nov. 1993.

[2] I. Koren, Z. Koren and C.H. Stapper, "A Unified Negative Binomial Distribution for Yield Analysis of Defect Tolerant Circuits," *IEEE Trans. on Computers*, vol. 42, pp. 724-437, June 1993.

[3] I. Koren and C.H. Stapper, "Yield Models for Defect Tolerant VLSI Circuit: A Review," *Defect and Fault Tolerance in VLSI Systems*, I. Koren (ed.), pp. 1-21, Plenum, 1989.

[4] T. Sugibayashi *et al.*, "A 1-Gb DRAM for File Applications," *IEEE J. of Solid-State Circuits*, vol. 30, pp. 1277-1280, Nov. 1995.

[5] T. Yamagata *et al.*, "A Distributed Globally Replaceable Redundancy Scheme for Sub-Half-micron ULSI Memories and Beyond," *IEEE J. of Solid-State Circuits*, vol. 31, pp. 195-201, Feb. 1996.

[6] J-H. Yoo *et al.*, "A 32-Bank 1Gb DRAM with 1GB/s Bandwidth," in ISSSC Dig. Tech. Papers, pp. 378-379, Feb. 1996.