C. H. Stapper
A. N. McLaren
M. Dreckmann

Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product

A model with mixed Poisson statistics has been developed for calculating the yield for memory chips with redundant lines and for partially good product. The mixing process requires two parameters which are readily obtained from product data. The product is described in the model by critical areas which depend on the circuit's sensitivity to defects, and they can be determined in a systematic way. The process is represented in the model by defect densities and gross yield losses. These are measured with defect monitors independently of product type. This paper shows how the yield for any product can be calculated given the critical areas, defect density, and mixing parameter. Future yields are forecast by using expected improvements in defect densities. Examples show good agreement between actual and calculated yields.

Introduction

Computer memory chips containing 65 536 memory bits are now available, and a trend towards larger chips with even greater bit densities is becoming apparent. As the number of bits goes up, the probability of having memory cell or word and bit line failures increases. Several manufacturers have therefore begun to use redundant memory bits in their product [1-3].

Using redundant memory bits to replace defective ones has been proposed by numerous inventors and authors. Sakalay, Fletcher, and Kril [4-7] devised several schemes for redundancy in core memories. Tammaru and Angell [8] described and calculated the yield of memory arrays and logic circuits with redundancy. A yield calculation with multiple word and bit line redundancy in memory arrays was published by Chen [9], and Arzubi [10] devised a method for implementing such a scheme on integrated circuit memory chips. Recently, Schuster [11] made a set of calculations to show the yield improvement possible as a function of the average number of faults on a memory chip and the total number of redundant lines.

Another scheme for enhancing yield is that of using partially good product. Elmer *et al.* [12] described this scheme for use in a CCD memory chip. Many partially good combinations are possible. We have worked with half good, two-thirds good, seven-eighths good, eight-ninths good, and nine-tenths good schemes for memory chips. The optimum scheme usually depends on the memory organization in which the chips are used.

The effectiveness of redundancy and partially good product depends strongly on the fault distribution, *i.e.*, the probabilities for having zero, one, two, three, or more faults per chip. Except for Cenker *et al.* [3], the calculations made in the papers on redundancy and partially good product referred to previously all assumed binomial or Poisson distributions. Yet, data by Moore [13] showed that this is not necessarily the case. Both Warner [14] and Stapper [15] showed that Moore's data could be represented by mixed Poisson statistics. More data showing this to be true is given in this paper, thus making the extension of mixed Poisson statistics necessary for both re-

dundancy and partially good product yield calculations. Formulas for this theory, as well as the comparison between theoretical calculations and actual process data, are the major subjects of this paper.

Statistics

Figure 1 shows the fault distribution obtained from a sample of 141 memory chips having 4K- × 9-bit organization. The chips were selected from 21 wafers which were randomly picked during three months of production. Each wafer was tested, and the sample was selected from chips for which the locations of failing cells, word lines, and bit lines were known. These chips were then visually inspected and systematically delayered to uncover all defects which caused the faults. Defects which did not cause faults or failures were not included in the distribution.

For a Poisson distribution, the mean must equal the variance. For the data in Fig. 1, the mean was 2.333 and the variance was 4.619, clearly not Poisson. These results can best be modeled with mixed Poisson statistics using a gamma distribution as the mixing function. This results in a Polya-Eggenberger distribution of the form [15]

$$P(X = x) = \frac{\Gamma(x + \alpha)\left(\frac{\bar{\lambda}}{\alpha}\right)^x}{x!\Gamma(\alpha)\left(1 + \frac{\bar{\lambda}}{\alpha}\right)^{\alpha+x}}, \quad (1)$$

where x is the number of faults per chip, $\bar{\lambda}$ the average number of faults per chip, and α a parameter that depends on the fault density variation. The quantities $\bar{\lambda}$ and α make (1) a two-parameter distribution. Poisson statistics use only a single constant parameter λ, which is equal to the mean and the variance of the number of faults per chip distribution.

The mean and variance for (1) are

$$E(X) = \bar{\lambda} \text{ and} \quad (2)$$

$$Var(X) = \bar{\lambda}\left(\frac{\bar{\lambda}}{\alpha} + 1\right). \quad (3)$$

Instead of α, it is often useful to express the above results in terms of the coefficient of variation of the fault density λ. This fault density has its own probability distribution function (pdf) with a mean, variance, and coefficient of variation σ/μ. When this pdf or mixing function is a gamma distribution with parameters $\bar{\lambda}$ and α [15], it is found that

$$\sigma/\mu = 1/\sqrt{\alpha}. \quad (4)$$

In the limits, when α approaches infinity or σ/μ approaches zero, expressions (2) and (3) approach those of Poisson statistics. In that case, only a single value for λ exists, indicating that the pdf of λ is a delta function.

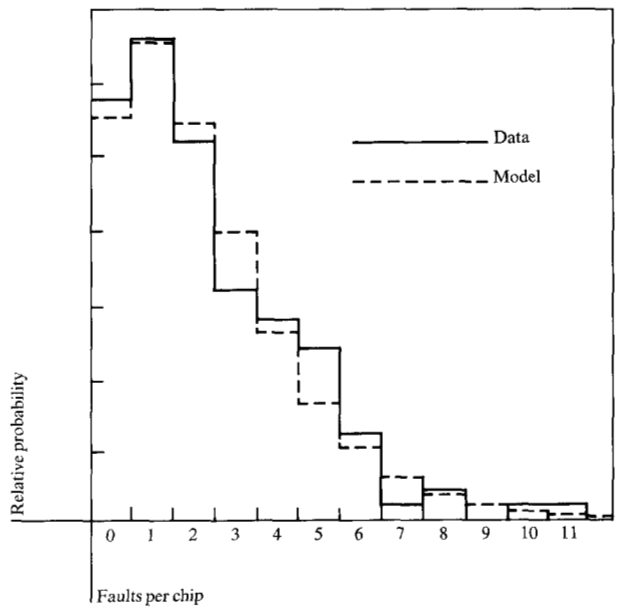

Figure 1 Fault distribution from physical analysis.

The parameters α and $\bar{\lambda}$ can readily be estimated from the mean and the variance of the data by solving (2) and (3). For the data in Fig. 1, we obtain $\bar{\lambda} = 2.333$ and $\alpha = 2.382$ ($\sigma/\mu = 0.6480$). The theoretical distribution with these parameters is shown by dashed lines in Fig. 1. Testing the calculations with a Smirnov-Kolmogorov test gives a significance level greater than 0.2, indicating that this model cannot be rejected as a fit for the data.

Another sample of wafers manufactured during the same period was subjected to a more severe retention time test. On those chips that were not completely inoperable, the number of isolated failing single cells was counted. The resulting distribution shown in Fig. 2 has a far longer tail than was observed in the visually inspected sample. These long tails are typical of the distributions for junction leakage failures, which can also be represented by the Polya-Eggenberger distribution shown by the dashed lines in Fig. 2. In this case we obtained $\bar{\lambda} = 1.572$, $\alpha = 0.3927$, and $\sigma/\mu = 1.596$.

Values of σ/μ ranging from 0.5 to 2 have been encountered in samples from different products. The long tails in these distributions have a profound effect on redundancy. Poisson distributions are short tailed, and calculations based on such statistics, therefore, require less redundancy than may be needed in actuality. To prevent this

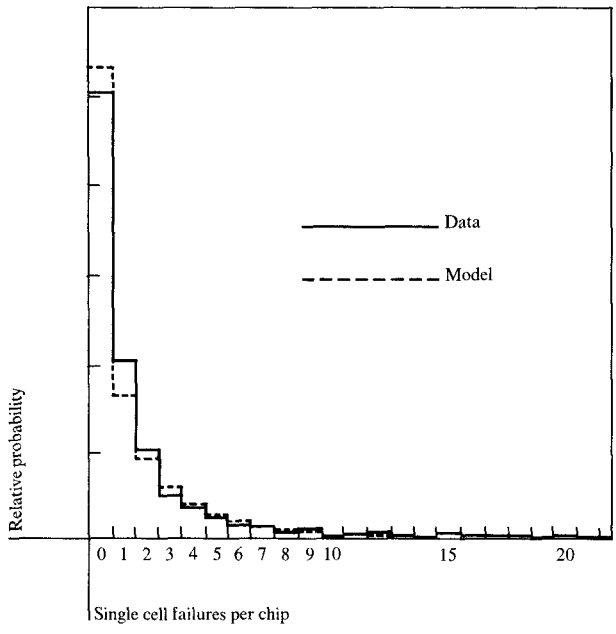

Figure 2 Single cell failure distribution for a 4K- × 9-bit memory chip.

Table 1 Defect and fault types used in the yield model.

Defect types (D)	*Fault types (λ)*
1. Node leakage	1. Single cell
2. Bit line leakage	2. Double cell
3. Peripheral circuit leakage	3. Single word line
4. Missing diffusions	4. Double word line
5. Extra diffusions	5. Single bit line
6. Missing polysilicon	6. Double bit line
7. Extra polysilicon	7. Redundant circuits
8. Missing aluminum	8. Section kill
9. Extra aluminum	9. Chip kill
10. Holes in the SNOS oxide	
11. Holes in the MNOS oxide	
12. Holes in the polysilicon oxide	

problem, we have based our model completely on mixed Poisson statistics. The mixing function in all cases is the gamma distribution of faults per chip or, if needed, per circuit.

Cenker *et al.* [3] claimed to have made their redundancy yield calculation with a model based on a theory by Price [16]. That model essentially has α and σ/μ equal to one. Although better than Poisson statistics, such a model still cannot span the actual range of data.

Organization of the model

The yield model consists of one set of parameters characterizing the process and another set describing the product. We have found it useful to define the defect densities as a set of parameters that causes faults. Mathematically, we can write this in matrix form as

$$\boldsymbol{\lambda} = \mathbf{A}_c\mathbf{D}, \tag{5}$$

where $\boldsymbol{\lambda}$ is a vector representing the nine faults per chip or circuit, $\mathbf{A}_c$, the critical area matrix of the product, and $\mathbf{D}$, a vector consisting of the defect densities of the process.

The twelve random defect types making up the defect density vector are shown in Table 1. Defect densities for missing contact holes have not been included. We have never been able to collect data that substantiated a model like that in (5) for this type of defect. Furthermore, less than 1% of the contact holes could be associated with fixable faults; the rest all caused a section or chip failure when missing. This led us to model missing contact holes conveniently as gross yield multipliers.

Other process parameters are modeled as gross yield detractors and simple yield multipliers. These include second level metal and contact holes, as well as over- or under-exposed patterns, over- or under-etched patterns, and misalignment for any photo step. Gross area-related yield losses, such as threshold voltage, transconductance, and contact resistances that are out of specification, fit this model implicitly. So do the gross components of junction leakage, which destroy complete areas of wafers as well as entire wafers.

Yield losses that are not due to the process, such as those resulting from handling, misprobing, mistesting, and errors in redundancy implementation, are also accounted for by gross yield multipliers. Our models also include gross yield estimates for circuits that fail to function due to a certain combination of device parameters that are nonetheless within process specifications.

To model the random defect losses requires first a list of the circuit areas such as the list for a 4K- × 9-bit chip shown in Table 2. Each circuit in this list is analyzed for its sensitivity to the 12 random defect types. This sensitivity has to be calculated for each type of fault that may be caused by any one of these types of defects. The list in Table 1 shows nine fault types, but we have used more in other applications. These faults are represented by the vector $\boldsymbol{\lambda}$ in formula (5).

The critical area matrix $\mathbf{A}_c$ in (5) is a 12 × 9 array of numbers representing the sensitivity of a chip to random defects. This sensitivity to defects is obtained by calculating the critical area, *i.e.*, the area of the circuit in which a defect must fall to cause a fault. For leakage defects this is assumed to be the total area of the metallurgical junc-

tion. Similarly, the critical area for dielectric pinholes is the area of overlap of the conductors. In all these cases defect size has been neglected, but this cannot be done for defects in the diffusion, polysilicon, and first level metal photo patterns. The photo defects are known to be distributed by size [17, 18]. Dennard [19] had shown the defect size distribution to be falling off as $1/x^3$ for defects with diameter x. Other workers at IBM have since verified that this is a good approximation. We have, therefore, calculated our photo-critical areas with this distribution. This requires the generation of so-called probability of failure curves. This is done by computer generation of a set of random coordinates on 2000× plots of the circuits. Disks of different diameters are then placed on these locations. The probability of failure for a given disk size is the fraction of defects of that size which we determined to have caused a fault. The probabilities of failure vary with defect size as shown, for example, in Fig. 3 for missing diffusions in the array of the 4K- × 9-bit chip. There are probability of failure curves for single cell, double cell, single bit line, and other failure combinations. These failures or faults are mutually exclusive since any given defect size at a random location can only result in one of these fault types.

Integrating these probability of failure curves over the size distribution gives the fraction of the circuit area that is sensitive to the average photo defect. Multiplying the results by the circuit area gives the critical area for each defect and fault type. In this way critical area matrices can be obtained for each circuit. The sum of the critical areas of all circuits gives the critical area matrix for the entire chip. Such a matrix for the 4K- × 9-bit chip is shown in Table 3. By establishing libraries of critical areas for different circuits, we have been able to reconfigure chips to optimize expected productivities by means of redundancy and partially good schemes.

Redundancy

The concept of redundancy is straightforward; spare word lines and bit lines are made available to replace failing lines and to bypass failing cells. The effectiveness of redundancy may be assessed by a probabilistic model to be described below. Qualitatively, however, the power of the concept can be illustrated as follows:

Suppose we have a large number of arrays for which there are, on the average, two failing word lines, one failing bit line, and two failing cells—a total of five failures on the average. It is clear that it will be rare for such an array to have no failures. In a sense, having no failures would require that none of the five typical failures occur. Thus, the yield without redundancy is low. Suppose, however, we have four redundant word lines and four redundant bit lines. Then on the average we have $4 - 2 = 2$ more word lines than failing lines and $4 - 1 = 3$ more bit lines than failing lines, giving a total of $2 + 3 = 5$ spare lines to

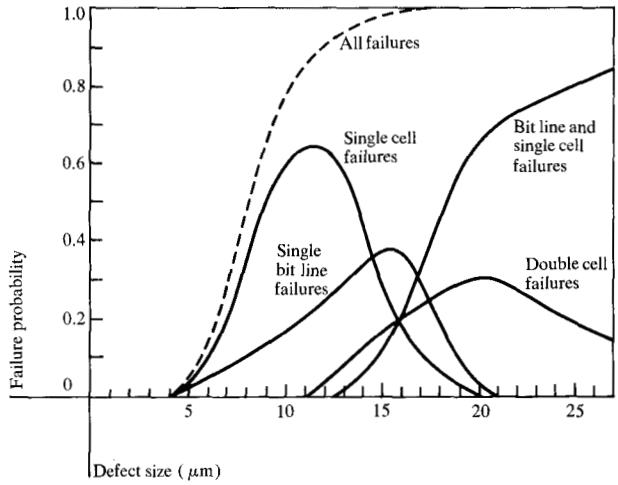

Figure 3 Probability of failure as a function of defect size for different failure modes in a memory array. These curves must be combined with the defect size distribution, defect density, and array area to obtain the average number of faults associated with each of these failure modes.

Table 2 Circuit types and corresponding areas.

Circuit type	*Circuit area* (mm^2)
Array cells	10.64
Sense amplifiers	1.53
Sense amplifier drivers	0.79
Sense amplifier underpasses	0.15
Word decoders and drivers	1.46
Word line terminators	0.56
Bit decoders and drivers	0.91
Word redundancy	
Word lines	0.17
Decoders and drivers	0.02
Terminators	0.01
Redundancy compare	0.29
Bit redundancy	
Bit lines	0.30
Sense amplifiers	0.04
Decoders and drivers	0.03
Steering circuits	0.59
Redundancy compare	0.19
True/complement generator	1.27
Timing chain	0.52
Phase drivers	1.34
Restore drivers	0.72
Internal voltage supply	0.09
Wiring	1.58
Pads	2.19
Total circuit area	25.39

Table 3 Critical area matrix for a 4K- × 9-bit chip. All areas in mm^2.

	Single cells	*Double cells*	*Single word lines*	*Double word lines*	*Single bit lines*	*Double bit lines*	*Redundant circuits*	*Chip kill*
Junction leakage								
Storage node	8.67	—	—	—	—	—	0.38	—
Bit line	—	—	—	—	3.03	—	—	—
Peripheral circuit	—	—	—	—	—	—	—	—
Diffusion								
Missing pattern	0.65	0.05	0.24	0.21	0.54	0.05	0.17	0.44
Extra pattern	3.05	3.79	0.30	0.29	0.65	0.11	0.41	0.29
Polysilicon								
Missing pattern	2.02	0.10	0.08	0.16	0.39	0.04	0.12	0.05
Extra pattern	0.44	0.01	0.08	0.03	0.18	—	0.06	0.18
Missing oxide	—	—	4.99	0.03	0.35	—	0.40	3.62
Metal								
Missing pattern	0.25	—	0.35	—	0.07	0.01	0.04	0.35
Extra pattern	—	—	0.02	0.66	0.04	0.03	0.06	0.32
Pinholes								
SNOS thin oxide	6.49	—	0.16	0.12	0.22	0.03	0.44	0.52
MNOS thin oxide	0.35	—	0.13	0.09	0.12	0.01	0.13	0.67

Table 4 Combinations constituting a fixable pattern for a 1 × 1 redundancy scheme.

1 cell	*2 cells*	*1 word line*	*2 word lines*	*1 bit line*	*2 bit lines*
0	0	0	0	0	0
1	0	0	0	0	0
2	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	1	0	1	0
0	0	0	0	1	0
1	0	1	0	0	0
1	0	0	0	1	0
1	1	0	0	0	0

correct the two failing cells. Intuitively, it will be rare for there to be enough failures to exceed redundancy capability, and the yield will be high with redundancy.

The above argument must be modified for fatal defects which are not in principle correctable by redundancy. These clearly put a limit on redundancy effectiveness; in practical cases, however, redundancy has been shown to enhance yield significantly.

Redundancy computation

For an array with no redundancy, the yield is simply the probability that no faults occur. With redundancy, the yield is the probability that a fixable pattern of failures occurs. Given critical areas and defect densities, mean fault frequencies ($\bar{\lambda}$) can be calculated, and from these yield with redundancy can be derived.

Early simulation work by McLaren indicated that one redundant line for 64 lines is appropriate for present-day manufacturing conditions. This level of redundancy has been implemented in several chips, producing significant yield increases. Since that time, an analytical approach has been developed, which we are currently using in our yield modeling work. This analytical approach is now described.

Since the yield with redundancy is the probability that a fixable pattern of failures occurs, it can be calculated as the sum of the probabilities of the (mutually exclusive) events which constitute a fixable pattern. This is illustrated by supposing that, for simplicity, we have one redundant word line and one redundant bit line. The events constituting a fixable pattern may then be represented in matrix form as in Table 4.

Each row of the matrix is an event; each entry in a row is the number of failures of each type (single cell, etc.) that is fixable with the available redundancy; thus, each row is a fixable pattern of failures. All the rows constitute the set of all fixable events.

The yield is then calculated by first deriving the probability of each event and then summing these probabilities over the rows.

If we use λ_{SC}, λ_{DC}, λ_{SWL}, λ_{DWL}, λ_{SBL}, λ_{DBL} to denote the mean frequencies per array of fixable failures, such as single cell, double cell, etc., a Poisson model for the first event (first row) gives the probability

$$P\ (000000) = e^{-\lambda_{SC}} e^{-\lambda_{DC}} e^{-\lambda_{SWL}} e^{-\lambda_{DWL}} e^{-\lambda_{SBL}} e^{-\lambda_{DBL}}. \quad (6)$$

For the second row we get

$$P\ (100000) = \frac{e^{-\lambda_{SC}} \lambda_{SC}}{1!}\, e^{-\lambda_{DC}} e^{-\lambda_{SWL}} e^{-\lambda_{DWL}} e^{-\lambda_{SBL}} e^{-\lambda_{DBL}}. \quad (7)$$

In general, we can express these probabilities as

$$P\ (i\ j\ k\ l\ m\ n) = e^{-(\lambda_{SC}+\lambda_{DC}+\lambda_{SWL}+\lambda_{DWL}+\lambda_{SBL}+\lambda_{DBL})} \times \frac{\lambda_{SC}^{i} \lambda_{DC}^{j} \lambda_{SWL}^{k} \lambda_{DWL}^{l} \lambda_{SBL}^{m} \lambda_{DBL}^{n}}{i!\ j!\ k!\ l!\ m!\ n!} \quad (8)$$

for an event with i single cell failures, j double cell failures, k single word line failures, etc.

When all the probabilities for fixable events are added together, we have the probability of a fixable array. That probability must, of course, be multiplied by the probability of no fatal (or chip-kill) defect to get the net chip yield.

The above model uses Poisson statistics. It can be readily modified to mixed Poisson statistics when λ_{SC}, λ_{DC}, λ_{SWL}, etc., are assumed to be proportional to the total mean number of faults λ for the chip. This total mean number of faults per chip is given by

$$\lambda = \lambda_{CK} + \lambda_{SC} + \lambda_{DC} + \lambda_{SWL} + \lambda_{DWL} + \lambda_{SBL} + \lambda_{DBL}, \quad (9)$$

where λ_{CK} represents the average number of fatal or chip-kill faults. These are faults caused by defects in peripheral circuits, such as timing and power supply circuits. Such faults make the chip completely nonfunctional.

Using the same gamma distribution for the mixing function of λ as before changes (8) into

$$P\ (i\ j\ k\ l\ m\ n) = \frac{\Gamma(i+j+k+l+m+n+\alpha)}{\Gamma(\alpha)} \times \frac{\left(\frac{1}{\alpha}\right)^{i+j+k+l+m+n}}{\left(1+\frac{\bar{\lambda}}{\alpha}\right)^{i+j+k+l+m+n+\alpha}} \times \frac{\bar{\lambda}_{SC}^{i} \bar{\lambda}_{DC}^{j} \bar{\lambda}_{SWL}^{k} \bar{\lambda}_{DWL}^{l} \bar{\lambda}_{SBL}^{m} \bar{\lambda}_{DBL}^{n}}{i!\ j!\ k!\ l!\ m!\ n!}. \quad (10)$$

This expression includes the fatal faults in the average number of faults per chip, $\bar{\lambda}$. Summing (10) over all fixable patterns therefore leads to the correct random defect yield of the chip.

Matrices for fixable events, such as the one in Table 4, have been generated for numerous redundancy schemes.

Table 5 Number of fixable combinations considered for various redundancy schemes.

Redundancy		*Fixable combinations*
Word lines	*Bit lines*	
0	0	1
1	0	3
2	0	8
3	0	16
4	0	30
5	0	50
1	1	10
2	1	22
3	1	43
4	1	77
5	1	126
2	2	53
3	2	100
4	2	177
5	2	284
3	3	185
4	3	320
5	3	506
4	4	548
5	4	854
5	5	1316

The number of patterns that we have used for such calculations are given in Table 5. With more redundant lines the probability of fixing the chip is increased. But fixability also depends on the spread of the fault distributions. A lower value of α (higher σ/μ ratio) causes longer tails in that distribution. These long tails can seriously lower the probability of fixing the chip.

Several reasons exist, therefore, for not always having sufficent redundancy to fix all fixable faults on the average. It is useful to define a quantity $\bar{\lambda}_{NF}$ for the average number of fixable faults that cannot be fixed. This definition is such that the random defect yield for the product is given by

$$Y = \left(1 + \frac{\bar{\lambda}_{CK} + \bar{\lambda}_{NF}}{\alpha}\right)^{-\alpha}. \quad (11)$$

But this yield should also be equal to the yield calculated by the summation of (10). As such, $\bar{\lambda}_{NF}$ can be determined by

$$\bar{\lambda}_{NF} = \left\{\alpha\left(\sqrt{\frac{1}{\Sigma\ P\ (i\ j\ k\ l\ m\ n)}} - 1\right)\right\} - \bar{\lambda}_{CK}, \quad (12)$$

where the summation is over all the fixable combinations. Good fixability schemes will result in low values of $\bar{\lambda}_{NF}$.

The model in (11) is essentially the same as the yield models used earlier by Sredni [20] and Stapper [15]. In

Table 6 Failure modes found by visual inspection.

	Single cell faults	*Double cell faults*	*Single word line faults*	*Double word line faults*	*Single bit line faults*	*Double bit line faults*
Junction leakage						
Storage node	0.248	0	0	0	0	0
Bit line	0	0	0	0	0	0
Diffusion						
Missing pattern	0.050	0	0	0	0.057	0.028
Extra pattern	0.057	0.035	0.007	0	0.021	0
Polysilicon						
Missing pattern	0	0	0	0	0	0
Extra pattern	0.106	0.007	0.021	0	0	0
Missing oxide	0.028	0	0.106	0.028	0.007	0
Metal						
Missing pattern	0.043	0.007	0.213	0.014	0.021	0.007
Extra pattern	0	0	0.028	0.099	0	0.007
Pinholes						
SNOS thin oxide	0	0	0	0	0	0
MNOS thin oxide	0	0	0.099	0.007	0	0
No visual defects	0.113	0.028	0.106	0.028	0.298	0.056
Repeating defects	0.064	0.029	0.029	0.029	0.036	0.007
Total	0.709	0.106	0.609	0.205	0.440	0.105

both of these approaches the mixing was done on the total number of random faults caused by random defects, such as photo, dielectric, and leakage defects. This differs from earlier work by Stapper [21, 22] where mixed Poisson statistics were applied to individual defect types. This approach is not readily extendable to redundancy calculations and has therefore been superseded by the methods described in this paper.

Expression (11) does not represent the complete yield model. As discussed in the section on organization of the model, the gross yields still have to be included. This is done by multiplying (11) by the total gross yield. The result is a yield formula similar to the model described by Paz and Lawson [23] but extended to include the faults per chip caused by all the different random defect types.

Redundancy results

The yield model was first used on a 32K-bit chip with four redundant word lines and four redundant bit lines. Defective bit or word lines could be replaced by redundant lines. This was achieved by blowing a pattern of fuses to steer the decoders around the defective addresses [24]. The perfect plus fixable chip yield for 13 lots is shown as a function of the perfect chip yield in Fig. 4. Each data point is the yield of a lot consisting on the average of ten wafers.

The product had a kerf structure that allowed measurement of the defect densities. This was done with serpentine and fingered defect monitors. The monitor data were converted to defect densities by the method described by Stapper [21, 22]. The average defect densities for all the lots combined were used to calculate the yield indicated by the triangle. The defect densities were then changed covariantly in order to obtain the solid line. As can be seen, the agreement between the model and the data is quite good.

Of special interest are the two lots represented by the two high yield points in Fig. 4. The defect monitor data for these lots showed that all defect densities except for leakage defects were low. The leakage defect densities were in fact higher than in the other lots. These defects caused a large number of single cell failures but very few chip-kill failures. The low defect densities for the other defects also resulted in a lower number of chip-kill failures as well as fewer fixable failures. The single cell failures can be fixed by using the redundant lines. This product is therefore highly fixable. A two-dimensional yield profile based on the actual redundant word and bit line implementation is shown in Fig. 5. The solid lines connect points obtained with these data. A set of dashed lines connect the points calculated with the model based on defect densities measured with the kerf. The product data along

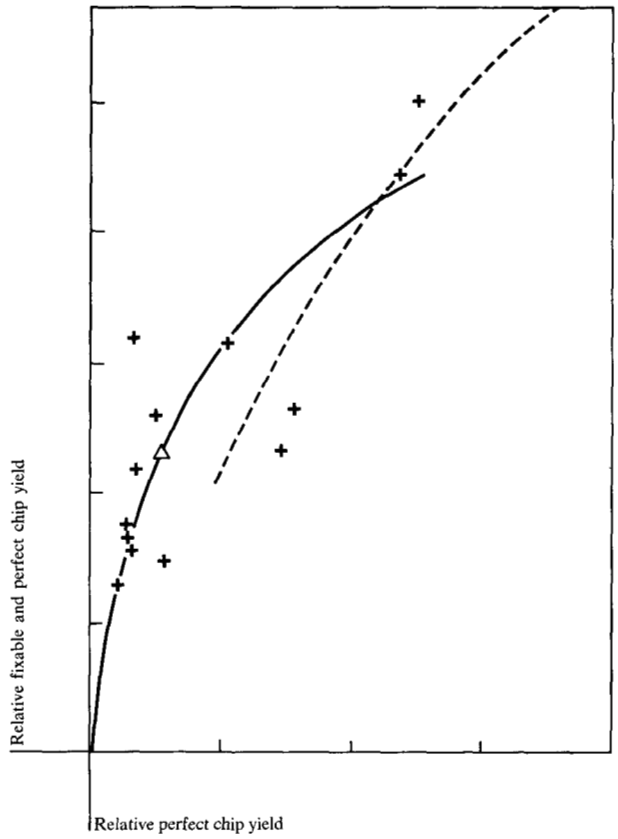

Figure 4 Redundancy profile for a 32K-bit chip. The solid line is calculated with the model.

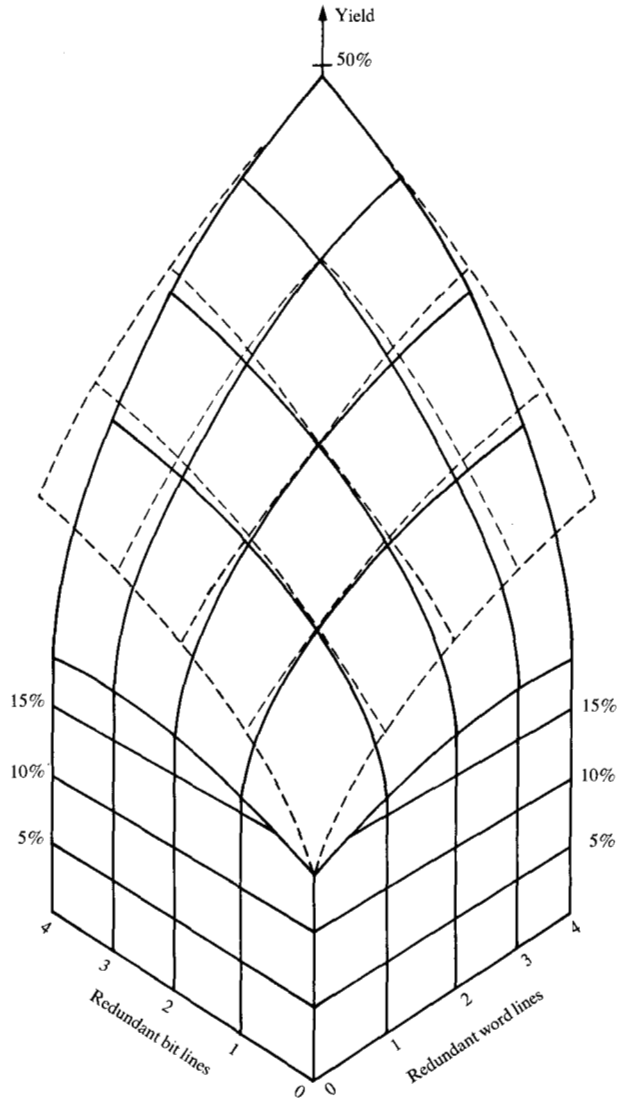

Figure 5 Yield as a function of redundancy. Data from a 32K-bit chip are in solid lines, calculations made with the model in dashed lines.

the zero redundant bit line and zero redundant word line axes are significantly lower than the calculated probabilities. This is an artifact of the actual method used for implementing the redundancy. In practice, we fixed all faulty bit lines first, then all faulty word lines, and finally other failures, using redundant bit and word lines alternately. This resulted in very few chips actually being fixed with word or bit lines only, even though this might well have been possible. The actual results of using 0×0, 1×1, 2×2, 3×3, and 4×4 redundancy are in very good agreement with theory.

We made a more comprehensive analysis of our model with the 141-chip sample of the 4K- × 9-bit chip mentioned earlier. For each of these chips, it was known which cells, word lines, and/or bit lines failed. With visual inspection and some chemical analysis, it was possible to relate defects to specific failures. Table 6 shows the results of this analysis in terms of the average number of faults per chip. Some caution must be taken when interpreting these data. Except for repeating defects, the sample for each type of fault consists of 141 chips. A single failure in the sample, therefore, leads to an average of 0.007 faults per chip. The 99% confidence interval for this result ranges between 0 and 0.052 faults per chip assuming a Poisson distribution. For higher failure counts, the 99% confidence interval gets even larger. For instance, 15 faults in this sample give an average of 0.106 faults per chip and a 99% confidence interval between 0.054 and 0.165 faults per chip when reckoned with Poisson statistics. These intervals would become even wider if mixed Poisson statistics were used.

Our objective was to compare these data to the fault densities calculated with our model. This was done by determining the defect densities with a sample of 1600

Table 7 Defect densities obtained from defect monitors. An * indicates defects greater than 2.5 μm. All densities are in defects per cm^2.

Junction leakage	
Storage node	1.93
Bit line	2.0
Diffusion	
Missing pattern	8.0*
Extra pattern	3.7*
Polysilicon	
Missing pattern	0.5*
Extra pattern	20*
Missing oxide	7.1
Metal	
Missing pattern	41.1*
Extra pattern	38.7*
Pinholes	
SNOS thin oxide	0.5
MNOS thin oxide	4.6

kerf monitors produced on the same wafers as the 141 chips. These defect densities are listed in Table 7. By means of the critical area matrix of Table 3 and matrix equation (5), the faults per chip shown in Table 8 can be determined.

The theoretical faults for cells and word lines agree with the data within the 99% confidence limits. It is clear, however, that something is missing in the case of the bit lines. This is due mostly to the nonvisual defects shown in the data. These defects occurred only under certain test conditions and, therefore, have been designated as "voltage sensitive failures." They often manifested themselves as either complete bit line failures or as varying numbers of cells failing along the bit lines. This appeared to be caused by marginal circuit operation under certain parametric combinations. In subsequent yield models, we have included these failures. This was done with an extra defect density and critical area so that the results conformed to the empirical data. As seen in Table 8, this addition to the model affects only the bit lines. Improved bias conditions on leakage and diffusion short monitors allow quantification of these defect densities, but product sensitivity depends on the design.

We did not include triple adjacent word and bit line failures in our model since the probabilities of failure and of critical areas were very small. In the data we did not find any triple bit line failures, but there were two chips with triple word line failures. These are not shown in Table 6. They represent an average of 0.014 faults per chip with 99% confidence bounds of zero and 0.066 faults per chip. This is small compared to the 0.609 and 0.205 faults per chip for single and double word line failures, respectively. These triple word line failures are therefore negligible.

Besides the failure modes, we also compared calculated yield with actual. This sample of 141 chips was part of a larger sample of wafers. The ratio of fixable to perfect chips on these wafers was 3.51. The model using kerf defect densities gave a ratio of 3.60. The absolute perfect chip yield was 1.06 times the model-predicted yield. For the fixable chip yield, the actual to model ratio was 1.03.

Partially good product

Failures often occur only in a given section of a chip, or when redundancy is present the remaining uncorrected failures may occur only in a given section of a chip. The section concerned may be one-half, one-quarter, or one-eighth of the chip, with the remainder of the chip (one-half, three-quarters, or seven-eighths) being fault-free. This suggests that the circuitry of a chip be partitioned so that the fault-free sections can function as independent units. The bit capacity of these sections is then available for packaging in modules along with other partially good or perfect product to give a total capacity of some marketable combination of memory bits.

Yield calculations for partially good product are somewhat different from those when only all good product exists. We define the equivalent yield as the fraction of usable capacity. If we denote the equivalent yield as Y_{EQ}, the all good yield by Y_{AG}, and the partially good yield as Y_{PG}, then

$$Y_{EQ} = Y_{AG} + (k/n)Y_{PG}, \tag{13}$$

where (k/n) is the fraction (1/2, 3/4, 7/8) of usable capacity for partially good chips. The concept of equivalent yield is applicable to both empirical situations (determining actual yield) and modeling situations (projecting yield). In the latter case, the partially good yield, Y_{PG}, has to be computed, and this may be done as follows:

Suppose the chip is partitioned into n independent sections and that we want the probability that k of the sections are fault-free (after correction with redundancy, if applicable). Then, if we know the yield of each section separately as $Y_{1/ns}$, the yield of defects fatal to the chip (chip-kill defects) as Y_{CK}, and the gross yield as Y_0, then the required probability (yield) is

$$\begin{aligned} Y_{k/n} &= \binom{n}{k} Y_0 Y_{CK} Y_{1/ns}^k (1 - Y_{1/ns})^{n-k} \\ &= \binom{n}{k} Y_0 Y_{CK} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j Y_{1/ns}^{j+k}. \end{aligned} \tag{14}$$

For example, in the case of two sections, the yield of half good product is

Table 8 Failure modes by type according to the model.

	Single cell faults	*Double cell faults*	*Single word line faults*	*Double word line faults*	*Single bit line faults*	*Double bit line faults*
Junction leakage						
Storage node	0.17	0	0	0	0	0
Bit line	0	0	0	0	0.06	0
Diffusion						
Missing pattern	0.05	0	0.02	0.02	0.04	0
Extra pattern	0.11	0.14	0.01	0.01	0.02	0
Polysilicon						
Missing pattern	0.01	0	0	0	0	0
Extra pattern	0.09	0	0.02	0.01	0.02	0
Missing oxide	0	0	0.35	0	0.02	0
Metal						
Missing pattern	0.10	0	0.14	0	0.03	0
Extra pattern	0	0	0.01	0.26	0.01	0.01
Pinholes						
SNOS thin oxide	0.03	0	0	0	0	0
MNOS thin oxide	0.02	0	0.01	0	0.01	0
Total	0.58	0.14	0.56	0.21	0.21	0.01
Voltage sensitive failures					0.23	0.09
Subtotal					0.44	0.10

$$Y_{1/2} = 2Y_0 Y_{CK} Y_{1/2s}(1 - Y_{1/2s}). \tag{15}$$

Note that $Y_{1/2}$ is not $Y_{2/4}$. The latter yield assumes that any two out of four sections on a chip are to be good. Six such combinations are possible.

If the mean number of faults in a partially good section is given by $\lambda_{1/ns}$ and the mean number of random fatal defects by λ_{CK}, then Eq. (14) can be rewritten as

$$Y_{k/n} = Y_0 \binom{n}{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j \, e^{-\lambda_{CK} + (j+k)\lambda_{1/ns}}, \tag{16}$$

thus expressing the yield completely in terms of the gross yield and random faults. Under the assumption that $\lambda_{CK} + (j + k)\lambda_{1/ns}$ is proportional to the total fault

$$\lambda = \lambda_{CK} + n\lambda_{1/ns}, \tag{17}$$

expression (16) can be transformed to a mixed Poisson model

$$Y_{k/n} = Y_0 \binom{n}{k} \sum_{j=0}^{n-k} \binom{n-k}{j} (-1)^j \times \frac{1}{\left\{1 + \frac{\bar{\lambda}_{CK} + (j+k)\bar{\lambda}_{1/ns}}{\alpha}\right\}^{-\alpha}}. \tag{18}$$

It is possible to combine a partially good product scheme with redundancy. This is precisely what has been done in IBM's 64K-bit low cost chip and 32K-bit high performance chip. In this case, the $\bar{\lambda}_{1/ns}$ is given by the nonfixable faults in one of the n sections. These quantities are then calculated by means of (10) and (12). This is the form used for our yield model.

Partially good product results

Data for a chip with half good product and redundancy are shown in Fig. 6. Each data point represents the average all good and half good yield for groups of five wafers which were consecutively tested at the output of an automatic VLSI wafer fabricator. The all good yield includes the yield of perfect chips as well as the yield for chips which have been fixed with redundancy. The half good yield consists similarly of perfect and fixed half good chips.

The model used for this chip calculates the yield for both redundancy and half good product. The dashed line in Fig. 6 was calculated more than two years before the product was actually manufactured. The solid line came from a yield projection made nine months before the product was made. By that time enough was known about

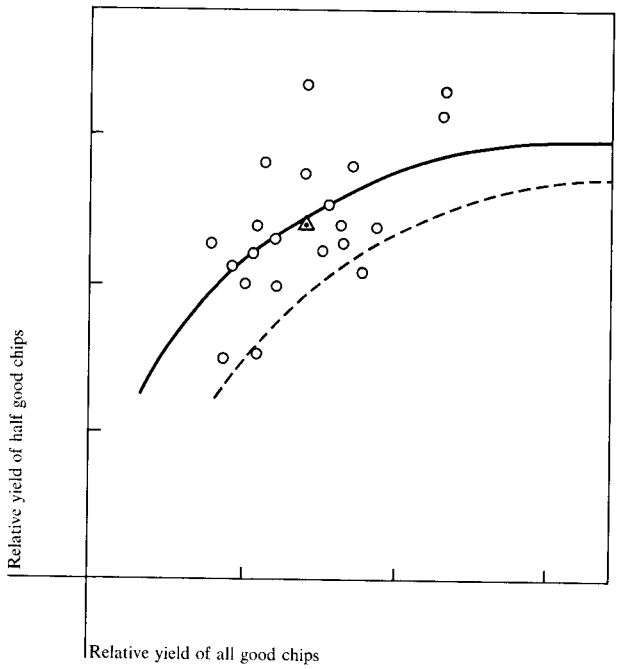

Figure 6 Yield of all good and half good chips. These chips have been fixed with redundancy. The lines are projections made with the model.

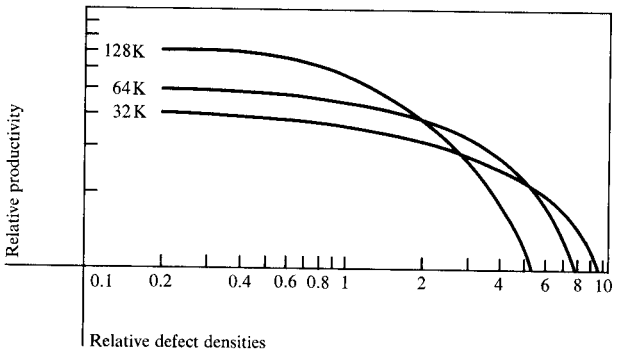

Figure 7 Bit productivity as a function of a defect density multiplier for three different chips.

the product and manufacturing line to make a more accurate forecast. The triangle shows the grand mean for the 105 chips in this sample.

Applications of the model

The model described in this paper has been used for yield projection and productivity optimization. Short term yield projections for a half year are made monthly based on existing in-line monitor and inspection data. The results of these projections are used to assess line loading and inventory of different parts needed for all good or partially good product. The long term yield projections are made using the best available technical judgment on how much can be learned for the individual parameters in the model. These parameters include the random defect densities as well as the gross yield detractors. The results are used to determine long term product cost.

Data are continually collected to see how actual results differ from the projected yield assumptions. The reasons for not making or exceeding targets are therefore continually visible and allow for quick management decisions and corrective action.

The long term yield projections have also been used to optimize productivity. We have stored the defect sensitivity factors for all circuits in computer libraries. By rearranging these circuits, we have been able to configure chips with redundancy and partially good schemes that optimize the expected number of good bits per wafer. The productivity in usable bits per wafer can be plotted as a function of defect densities, as in Fig. 7. This plot shows which chip size gives optimum productivity at a given multiple of a reference set of random defect densities. The manufacture of a 128K-bit chip at the time that a line has defect densities of five times the reference set would be disastrous, whereas productivity for either a 64K-bit or 32K-bit chip would be optimum.

Summary

We have found the yield model described in this paper to be a powerful tool for analyzing, projecting, and optimizing yield in VLSI memory chips. Although this model has been specifically developed for memory chips, we believe that the techniques can be applied to a large number of other integrated circuits. The use of this tool facilitates a systematic approach to yield analysis and has established a method for yield improvement.

Acknowledgments

In determining the probability of failure curves, we were helped by R. DeSimone, F. Girard, J. Hennessey, S. Kess, K. Kroell, D. Linke, W. Morton, E. Thoma, and D. Werling. We gratefully acknowledge the assistance of J. Hennessey in data collection and analysis, as well as programming support. It is impossible to acknowledge individually everyone involved over the years with the design, manufacture, testing, data collection, and characterization of the products made available to us for using and testing our model. We would like to express our gratitude for having been part of this team.

References

1. R. R. DeSimone, N. M. Donofrio, B. L. Flur, R. H. Kruggel, H. H. Leung, and R. Schnadt, "FET RAMS," *1979 IEEE ISSCC Digest of Technical Papers* **22**, 154–155 (1979).

2. R. P. Cenker, D. G. Clemons, W. R. Huber, J. B. Petrizzi, F. J. Procyk, and G. M. Trout, "A Fault-Tolerant 64K Dynamic RAM," *1979 IEEE ISSCC Digest of Technical Papers* **22,** 150–151 (1979).
3. R. P. Cenker, D. G. Clemons, W. R. Huber, J. B. Petrizzi, F. J. Procyk, and G. M. Trout, "A Fault-Tolerant 64K Dynamic Random-Access Memory," *IEEE Trans. Electron Devices* **ED-26,** 853–860 (1979).
4. F. E. Sakalay, "Correction of Bad Bits in a Memory Matrix," *IBM Tech. Disclosure Bull.* **6,** 1–2 (1964).
5. F. E. Sakalay, "Memory System for Using Storage Devices Containing Defective Bits," U.S. Patent 3,422,402, U.S. Cl. 340/172.5, Jan. 1969.
6. R. P. Fletcher, "Storage System Using a Storage Device Having Defective Storage Locations," U.S. Patent 3,444,526, U.S. Cl. 340/172.5, May 1969.
7. R. S. Kril, "Memory System," U.S. Patent 3,689,891, U.S. Cl. 340/172.5, Sept. 1972.
8. E. Tammaru and J. B. Angell, "Redundancy for LSI Yield Enhancement," *IEEE J. Solid-State Circuits* **SC-2,** 172–182 (1967).
9. A. Chen, "Redundancy in LSI Memory Array," *IEEE J. Solid-State Circuits* **SC-4,** 291–293 (1969).
10. L. M. Arzubi, "Memory System with Temporary or Permanent Substitution of Cells for Defective Cells," U.S. Patent 3,755,791, U.S. Cl. 340/173R, Aug. 1973.
11. S. E. Schuster, "Multiple Word/Bit Line Redundancy for Semiconductor Memories," *IEEE J. Solid-State Circuits* **SC-13,** 698–703 (1978).
12. B. R. Elmer, W. E. Tchon, A. J. Denboer, and R. Frommer, "Fault Tolerant 92160 Bit Multiphase CCD Memory," *1977 IEEE ISSCC Digest of Technical Papers* **20,** 116–117 (1977).
13. G. E. Moore, "What Level of LSI is Best for You?" *Electronics* **43,** 126–130 (1970).
14. R. M. Warner, Jr., "Applying a Composite Model to the IC Yield Problem," *IEEE J. Solid-State Circuits* **SC-9,** 86–95 (1974).
15. C. H. Stapper, "On a Composite Model to the IC Yield Problem," *IEEE J. Solid-State Circuits* **SC-10,** 537–539 (1975).
16. J. E. Price, "A New Look at Yield of Integrated Circuits," *Proc. IEEE (Lett.)* **58,** 1290–1291 (1970).
17. T. R. Lawson, Jr., "A Prediction of the Photoresist Influence on Integrated Circuit Yield," *Solid State Technol.* **7,** 22–25 (1966).
18. R. A. Maeder, F. W. Oster, and R. J. Soderman, "Semiconductor and Integrated Circuit Device Yield Modeling," U.S. Patent 3,751,647, U.S. Cl. 235/151.11, Aug. 1973.
19. R. H. Dennard, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, unpublished results; it is unfortunate that this work done by Dennard has never been published.
20. J. Sredni, "Use of Power Transformations to Model the Yield of IC's as a Function of Active Circuit Area," *1975 International Electron Device Conference Digest,* 123–125 (1975).
21. C. H. Stapper, "Defect Density Distribution for LSI Yield Calculations," *IEEE Trans. Electron Devices* **ED-20,** 655–657 (1973).
22. C. H. Stapper, "LSI Yield Modeling and Process Monitoring," *IBM J. Res. Develop.* **20,** 228–234 (1976).
23. O. Paz and T. R. Lawson, Jr., "Modification of Poisson Statistics: Modeling Defects Induced by Diffusion," *IEEE J. Solid-State Circuits* **SC-12,** 540–546 (1977).
24. B. F. Fitzgerald and E. P. Thoma, "Circuit Implementation of Fusible Redundant Addresses on RAMs for Productivity Enhancement," *IBM J. Res. Develop.* **24,** 291–298 (1980, this issue).

Received May 3, 1979; revised November 19, 1979

The authors are located at the IBM General Technology Division laboratory, Essex Junction, Vermont 05452.