On the Assumptions Contained in Semiconductor Yield Models

Albert V. Ferris-Prabhu, *Senior Member, IEEE*

Abstract—**The form of semiconductor yield models, and their predictions, are to a large extent affected by the size distribution of all defects and the spatial distribution of fatal defects. As the effects of these and other assumptions on yield models are rarely described in the literature, this paper examines them and develops scaling rules for the average number of fatal defects per chip. To trace differences in the predicted yield to the various assumptions, the treatment compares a simple Poisson model with a compound Poisson model and shows that, when appropriate scaling rules are used, yield predictions of the simple Poisson model are accurate for new product with chips of area up to an order of magnitude larger than chips of existing product.**

I. Introduction

SEMICONDUCTOR device yield modeling plays a central role in the semiconductor industry. Cost varies inversely as yield, and accurate cost estimates require accurate yield predictions. Whether used for strategic purposes to predict the yield in the year 2000 of a 2 in by 2 in analog chip with both memory and logic features, with four levels of metallization and 0.1 μm minimum design feature in a GaAs technology, or for tactical purposes to estimate the number of wafer starts needed next quarter to meet the committed volume of a 5 mm by 5 mm memory chip with two levels of metallization and 2 μm ground rules in a standard silicon technology, accurate yield predictions allow informed decisions to be made in a timely manner. Improved productivity with attendant cost reduction is central to a successful business.

There are many semiconductor yield models [1]–[6], but it is not always clear which model to use, as the underlying assumptions are not always stated fully. In principle it should be possible to compare the predictions of a model with actual results. In practice this is not so. A recent paper [7] attempts to compare the predictions of certain yield models with manufacturing data. But many manufacturing details that would affect the values of terms in the models are not given. This is understandable, as proprietary sensitivity inhibits the release of information (such as process details, tool specifics, product pattern geometry including minimum pattern width and spacing, as well as predicted and actual yields) of the kind and in the detail that would permit independent verification. Consequently the applicability, if not the validity, of some models has been questioned [8]. A more substantive comparative study [9] discusses analytical and simulation models in detail. But as neither these nor other analytical or simulational treatments [10]–[12] of yield modeling address the effect on predicted yield of the assumptions common to all yield models, it is the purpose of this paper to do so. Section II relates random-defect-limited yield to the observed yield and indicates two assumptions common to all yield models, Section III discusses how assumptions pertaining to the spatial dispersion of fatal defects affect the form of the yield equation, and Section IV examines how assumptions about the size distribution of all defects affect the expected average number of fatal defects per chip. Section V introduces scaling rules for estimating the average number of fatal defects per chip of new product from the observed yield of existing product and discusses the area scale factor, Section VI applies these rules to the yield equation and examines a compound Poisson distribution, Section VII compares data with predicted yields, and Section VIII discusses the results. Section IX concludes that, with proper scaling, the predictions of the simple Poisson model are accurate for chips of area up to at least an order of magnitude larger than the chips from which scaled.

Manuscript received August 29, 1990. This paper was recommended by Associate Editor A. E. Ruehli.

The author is with the General Technology Division, IBM, Essex Junction, VT 05452.

IEEE Log Number 9107750.

II. Random-Defect-Limited Yield

Central to all yield models is an expression that attempts to predict the probability of being able to manufacture a chip with no nonsystematic defect that will prevent the device from functioning as designed. Such a defect is referred to as a fatal random defect. If the design includes fault tolerance, nonfatal defects need to be considered as well. But if the design is not fault-tolerant, as in logic or application-specific integrated circuit devices, the yield model attempts to predict the probability of being able to manufacture a chip with no fatal random defect. This prediction is expected to agree with the actual fraction of chips on a wafer, averaged over some, many, or all wafers in a given time period, that is free of fatal random defects. It is expected that

$$p(0; \lambda) = Y_{RD} \tag{1}$$

where Y_{RD} is the random-defect-limited yield, λ is the

average number of fatal random defects per chip, and p is the prescription that predicts the probability of finding a chip free of such defects.

The fraction of chips per wafer that is functional, i.e., the *observed* yield, Y_{obs}, is the product of Y_{RD}, the fraction that is free of fatal random defects, and Y_o, the fraction that is also free of all other factors that would inhibit functionality, i.e., gross defects, field size errors, parametric irregularities, and so on. Thus,

$$Y_{\text{obs}} = Y_{RD} \times Y_o. \tag{2}$$

Of the nonfunctional chips, it is not always practical to determine which have failed test because of fatal random defects and which have failed test because of other reasons. One way of inferring the value of Y_o is by the window method, in which successively smaller grids are overlaid on a composite wafer map. The fraction of grids that contains no indication of nonfunctionality is taken to be the yield. The intercept with the ordinate, of this yield versus area curve extrapolated to zero area, is interpreted as Y_o. Thus

$$Y_{RD} = Y_{\text{obs}}/Y_o = \tilde{Y}_{\text{obs}} \tag{3}$$

and verification of the (random-defect-limited) yield model requires that the following equality hold:

$$p(0; \lambda) = Y_{\text{obs}}/Y_o. \tag{4}$$

The inability to determine the exact value of the Y_o term introduces an inherent uncertainty in the later verification of the predicted value with the actual value of the random-defect-limited-yield.

Also embedded in all models predicting random-defect-limited yield are two major assumptions that affect the predicted results. One pertains to the spatial dispersion over the wafer of **fatal** random defects, as reflected in the form of the prescription for p, i.e., of the yield equation. The other pertains to the size distribution of random defects with spatial extent, assumptions about which affect the value of λ, the average number of **fatal** defects per chip.

III. Spatial Dispersion of Fatal Defects

The form of the prescription, $p(0, \lambda)$, that predicts the probability of finding a chip with no fatal defect, where λ is the average number of fatal defects per chip, depends upon how the fatal defects are spatially dispersed across the wafer. For over 30 years, equations have been suggested in the yield literature [1]–[6] for predicting the yield (loss) of semiconductor chips caused by the dispersion of fatal defects over the wafer. However, the first treatment of spatial dispersion in general is due to Rogers [13], who has developed a rigorous theory showing that the form of the expression predicting the occurrence of an event in a given region depends on whether the occurrence probability of an additional event is related to or independent of the prior occurrence of other events in that region. Rogers shows that if the probability of occurrence of an event in a region decreases with the number of events that have already occurred in that region, the binomial model results; if the occurrence of an event is independent of the occurrence of prior events in that region, the Poisson model results; and if the probability of occurrence of an event increases linearly with the number of events that have already occurred in that region, the negative-binominal model results. All three cases are succinctly contained in the equation

$$p(0, \lambda; \beta) = (1 + \beta\lambda)^{-1/\beta} \tag{5}$$

which, when $\beta = 0$, reduces to the well-known Poisson equation:

$$Y = e^{-\lambda}. \tag{6}$$

The term β is related to a statistical measure called the coefficient of variation. Some workers use the term α, which is the reciprocal of β. However, it is more meaningful physically to interpret this term as a coefficient coupling the spatial occurrence of fatal defects [14] with $\beta = 0$ implying no coupling and $\beta = 1$ implying maximum coupling. The magnitude as well as the sign of β plays an important role. For a fixed value of λ, Fig. 1 shows that the random-defect-limited yield predicted by (5) increases as β increases. Equivalently, for a fixed value of the random-defect-limited yield, Fig. 2 shows that the average number of fatal defects per chip increases as β increases. Both support the interpretation of β as a coefficient coupling the location of fatal defects. A negative value of β implies that fatal defects are mutually inhibitory, tend to be more widely dispersed, and result in fewer chips free of fatal defects. A value of β equal to 0 implies that the occurrence of fatal defects is equiprobable anywhere. A positive value of β implies that fatal defects are mutually attractive, tend to occur in clusters, and result in more chips free of fatal defects. The question of which model is applicable for yield prediction thus depends critically upon how the fatal defects are spatially dispersed over the wafer.

Up to now no data have been reported in support of the binominal model. Moreover, when β is negative, (5) becomes negative for $\lambda > 1/|\beta|$, a physically unacceptable result for yield.

The Poisson model has been criticized for consistently predicting lower yields for larger chips than has later been observed. But that appears to be due more to the manner in which the average number of fatal defects per chip on new product of larger area is estimated than to any major weakness in the model [15].

The case where β is positive, i.e., the negative binomial distribution, was first formulated in 1714 by Montmort [16] in connection with waiting time problems in gambling. This well-known distribution has been applied to problems in eugenics [17], ecology [18], psychology [19], and astrophysics [20]. It was first proposed for yield prediction by Okabe *et al.* [21] and then by Stapper *et al.* [22], who translated Rogers's treatment to yield prediction. In all these applications, the common feature is that of dispersion: of plant species, of human characteristics,

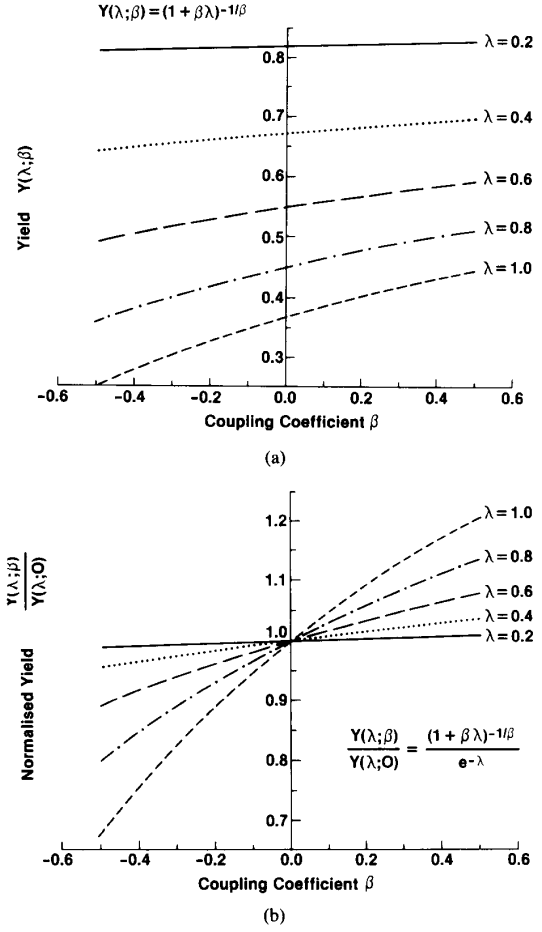

Fig. 1. Yield as a function of coupling coefficient for specified values of the average number of fatal defects per chip. (a) Absolute values. (b) Values relative to that with coupling coefficient zero.

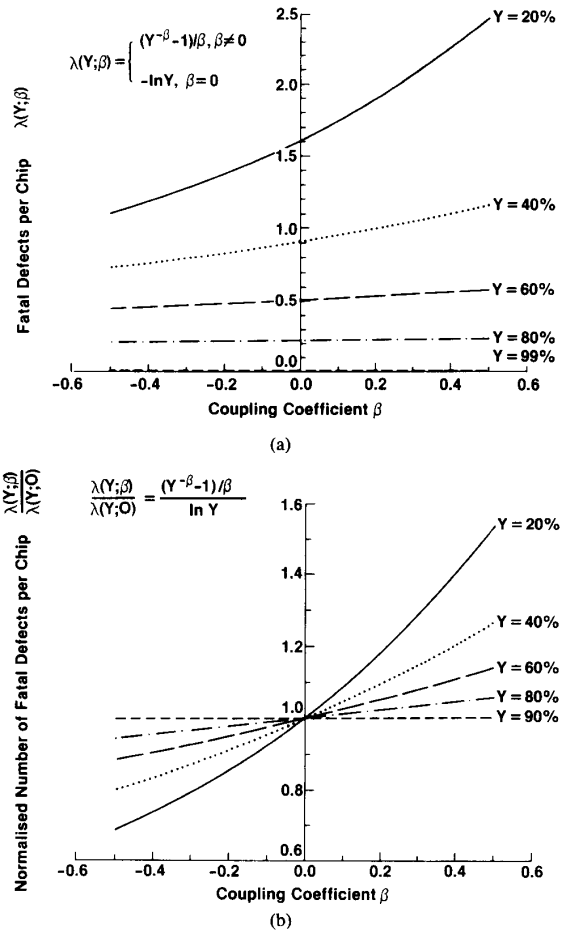

Fig. 2. Average number of fatal defects per chip as a function of coupling coefficient for specified values of the yield. (a) Absolute values. (b) Values relative to that with coupling coefficient zero.

of galaxies, or of defects in semiconductor devices. The negative binominal distribution is actually a compound Poisson distribution, and is shown by Feller [23] to be a limiting form of both the Polya distribution and of the Bose-Einstein distribution, a form of the latter being structurally similar to Seeds's yield model [24].

The Bose–Einstein distribution is usually encountered in the statistics of indistinguishable particles which have no constraint on the number that can occupy a given state. As such it is sometimes considered to be inapplicable to describing defects in semiconductor devices. However metallic inclusions leading to pipes, i.e., shorts between emitter and collector, or pinholes in the gate oxide or other insulators, are examples of defects that are indistinguishable from each other, and their areal density is extremely low. Thus negligible error will arise if their distribution is considered to be of the Bose–Einstein type. Nevertheless, for other defects that are distinguishable or for the defects as a whole, this distribution is not a suitable one.

The negative binomial model has been reported [22] to accurately predict semiconductor device yield, but if applied to each of the independent sequential processes involved in device fabrication, difficulties arise in the usual interpretation of λ as the average number of fatal defects per chip [14]. For example, if Λ is interpreted as the average number of fatal defects caused by all process steps, and λ_i as the average number of fatal defects introduced during the ith step, then the requirement that the net yield be the product of the yield of each of the intermediate processes, i.e., $Y_{\text{net}} = \Pi Y_i$, leads to the anomalous result that $\Lambda > \Sigma \lambda_i$, i.e., that the average of the sum is greater than the sum of the averages. This anomaly is due to the presence of the coefficient β, which serves to couple defects [14]. Though a statistical model of nearest-neighbor interaction that reproduces the predictions of the negative-binominal model has been developed [25], it is noteworthy that to date no **quantitative physical** model has been proposed to explain the cause of such interactions or the nature of the mechanism that would serve to couple defects introduced during presumably independent process steps, or even to suggest, *a priori*, when such a spatial dispersion of fatal defects is likely to occur.

Fatal Defects

Fatal defects are essentially of two kinds: point defects and defects with spatial extent. Examples of point defects are pinholes in the insulating medium between two conducting planes or between gate and channel in FET's and crystallographic defects and metallic inclusions that form conductive paths, or "pipes," between emitter and collector in bipolar transistors. As point defects **will** be fatal if they occur where they **can** be fatal, their distribution depends upon the device pattern geometry. For a defect with spatial extent to be fatal, however, not only must it occur where it **can** be fatal; it must also be large enough to be fatal. For example, to break a pattern meant to be conducting, the defect must be at least as large as the pattern width. Similarly, to connect two patterns meant to be disjoint, the defect must be at least as large as the space between them [26]. The spatial dispersion of **fatal** defects of this type therefore depends upon the spatial dispersion and size distribution of **all** defects, as well as on the pattern geometry of the device. As a result of this interaction between defect size and pattern geometry, even two identical chips fabricated on wafers that are simultaneously processed in the same fabricator may have a different spatial dispersion of fatal defects as well as a different number of fatal defects.

One might expect to be able to confirm the assumed choice of spatial dispersion, i.e., of yield model, by comparing predicted with observed yield. But such verification is obscured by the fact that the average number of fatal defects per chip, the value of which affects the predicted yield, is in turn affected by assumptions about the size distribution of random defects that have spatial extent [27].

IV. Size Distribution of Defects

In principle, $\lambda(A, w)$, the average number of all fatal defects introduced during fabrication on a chip of area A with minimum design feature w, is the product of the number $N_D(A)$ of **all** defects on the chip and the fault probability, $\Phi(w)$, i.e., the average probability that a defect will indeed be fatal. Thus

$$\lambda(A, w) = \Phi(w) N_D(A). \tag{7}$$

It has been shown elsewhere [28] that

$$N_D(A) = A \int dx N(A, x) \tag{8}$$

and

$$\Phi(w) = \int_w^W dx S(A, w) K(x - w) \tag{9}$$

where $S(A, x)$ is the area-dependent defect size-density function, $N(A, x)$, normalized to unity.

The fault probability kernel, $K(x - w)$, i.e., the probability that a defect of arbitrary size x will be fatal to a pattern of width w, depends only upon the pattern geometry. Once the design layout is known, it can be obtained analytically [29], [30], or by simulation [31]–[33].

In addition to an explicit size dependence, the defect size-density function has an implicit area dependence that arises because the average number of fatal defects per chip increases with the distance of the chip from the center of the wafer [34]–[37]. Consequently the wafer-averaged number of fatal defects per chip varies with chip area, and needs to be accounted for.

The actual form of the size dependence is not known and is the subject of continuing investigation [38]–[43]. It is reasonable to expect fewer large defects than smaller ones, so it is sometimes *assumed* that the size-density function follows an inverse power law, and an inverse cube relation is often used.

The difficulty in determining the defect size distribution is exacerbated by the fact that it is not possible in all cases to determine whether a defect is potentially fatal or even to define unambiguously what constitutes a defect. There are various inspection techniques that attempt to look for certain defined irregularities such as etch pits, foreign material, extra or missing metal, and spots of unidentifiable nature above a certain size. But these are done on a sample of chips on some wafers, and are subjective. The only unambiguous test is the final electrical one, where devices pass or fail a set of objective and reproducible criteria. Such tests determine only whether a chip is functional. To determine the cause requires physical failure analysis of the chip. The nature of the random fatal defect(s) and the layer at which they are introduced can be determined by various delayering methods. But these are time-consuming and in a manufacturing environment can reasonably be applied to only a small fraction of the chips that have failed. They thus provide size information only about fatal defects, and that too only on the chips that have been examined. They provide no information about fatal defects on chips that have not been examined and no information about "potentially" fatal defects on any of the chips that pass final electrical test. Size information therefore has to be obtained indirectly.

The use of hypothetical yield values to determine and optimize defect size distribution has been reported [38], but various inconsistencies have been noted [39]. A more fundamental difficulty with this approach is that it seeks to extract information about both the spatial distribution and the size distribution from basically one equation [27]. New measurement techniques have been reported recently [44] that do not share this limitation, but they determine the size of only fatal defects. Methods of analysis that assume an inverse power law and seek the power that fits the data with least error can determine the best choice of member of a particular class of functions, but are unable to determine which class is the better one.

As yet there is no physical model applicable to semiconductor device fabrication that predicts the distribution of defect sizes. So the determination of the size distribution of all defects, fatal and otherwise, is still an open question, inviting more sophisticated methods for obtain-

ing and analyzing pertinent data. Consequently, in practice, the average number of fatal defects per chip is not evaluated from (7) but rather is scaled from the average number of fatal defects per chip of existing product inferred from its observed yield [15].

V. Scaling Rules for Fatal Defects

In practice, the average number of fatal defects, $\lambda(A, w)$, expected on a new product chip of area A and minimum design feature w is obtained through the relation

$$\lambda(A, w) = \sigma \times \lambda(A_e, w_e). \tag{10}$$

The scale factor, σ, is defined in (13a) below. The average number of fatal defects, $\lambda(A_e, w_e)$, on an existing product chip of area A_e and minimum design feature w_e is inferred from the observed yield, $Y_{\text{obs},e}$, by inverting the yield equation, i.e.,

$$\lambda(A_e, w_e) = p_e^{-1}\left(\frac{Y_{\text{obs},e}}{Y_{o,e}}\right). \tag{11}$$

From (5) this can be written as

$$\lambda(A_e, w_e) = \frac{\left(\frac{Y_{\text{obs},e}}{Y_{o,e}}\right)^{-\beta_e} - 1}{\beta_e}, \tag{12a}$$

which, when $\beta_e = 0$, reduces to the Poisson expression

$$\lambda(A_e, w_e) = -\ln \frac{Y_{\text{obs},e}}{Y_{o,e}}. \tag{12b}$$

Assumptions about the value of the Y_o term and of the coupling coefficient, β, affect the inferred value of the average number of fatal defects per chip. For a given value of the observed yield, λ_e increases as Y_o increases or as β increases. As neither Y_o nor β is known exactly, this introduces flexibility into the scaled value of $\lambda(A, w)$.

The scale factor, σ, is defined by the expression [45]

$$\sigma = \xi \times \psi \times \alpha. \tag{13a}$$

The sensitivity factor, ψ, and the area factor, α, are defined in (13b) and (13c) below. The process complexity scale factor, ξ, attempts to account for the difference in the expected number of fatal defects per chip because of differences in the number of layers, process steps, technology, and tools. The value assigned to it, in the neighborhood of unity, is based upon experience and engineering judgment and is clearly an adjustable parameter.

The sensitivity scale factor, ψ, is the ratio of the fault probabilities of the new and existing product. Given by [43],

$$\psi = \frac{\Phi(w)}{\Phi(w_e)} \simeq \left(\frac{w}{w_e}\right)^{p-1}, \tag{13b}$$

and its magnitude depends on assumptions made about the defect size distribution. The magnitude of the sensitivity factor increases as p increases.

It has been shown [38] that a choice of $p = 3$ keeps constant the random-defect-limited yield of chips where the area is reduced in proportion to the minimum ground rules, keeping the number of circuits constant. In other words, the inverse cube law is effectively yield-neutral. So it is convenient to use, permitting small adjustments to the scaled number of fatal defects per chip to be made through other factors.

The magnitude of the area scale factor, α, defined by the expression [45]

$$\alpha = \left(\frac{A}{A_e}\right)^{1-b}, \quad 0 \le b \le 1, \tag{13c}$$

decreases as b increases. The term b in the exponent of the chip area ratio is an empirical measure of the extent by which the average number of defects per chip deviates from linearity with respect to the chip area.

If the density of fatal defects is invariant with respect to the chip area from which computed, and if the minimum design feature, w, and thus the fault probability or defect sensitivity, $\Phi(w)$, is the same, the average number of fatal defects per chip should increase linearly with chip area, and b should be 0. In fact this is not always so.

To demonstrate this, yield data from 118 wafers of a bipolar SRAM were analyzed [37]. The actual product chip consists of ten identical array segments with common decode and peripheral circuitry. Through appropriate diagnostic procedures, it is possible to determine whether nonfunctionality is due to failure of the array or the non-array portion and to determine which array segments have failed. This permits each wafer of single ten-array segment memory chips to be treated as though it consists of ensembles of "pseudochips," each of 1, 2, 3, $\cdots$, 10 segments. All these pseudochips have the same design rules and fault probability and experience the same defect environment. The only difference between them is their area. Analysis of their respective yields should therefore provide reasonable information about how the average number of fatal defects per chip scales with area.

Fig. 3 shows the yield of each pseudochip. Curvilinear extrapolation to zero area suggests that the Y_o term for these data is reasonably close to unity. Fig. 4 shows the average number of fatal defects per pseudochip relative to that of the one-array segment pseudochip, as inferred from the yield prescription of (5) with four intermediate choices of the defect coupling coefficient, β. For most values of the coupling coefficient, this ratio is close to the square root of the area ratio of the pseudochips, though numerical methods show that a value of $b = 0.4$ in the exponent of the area scale factor provides better agreement. As these results were obtained from actual yield data of a commercial product that had been in volume production, it is not likely that one pseudochip would differ from another in any way other than area, suggesting that it is reasonable to introduce an area factor when scaling the number of

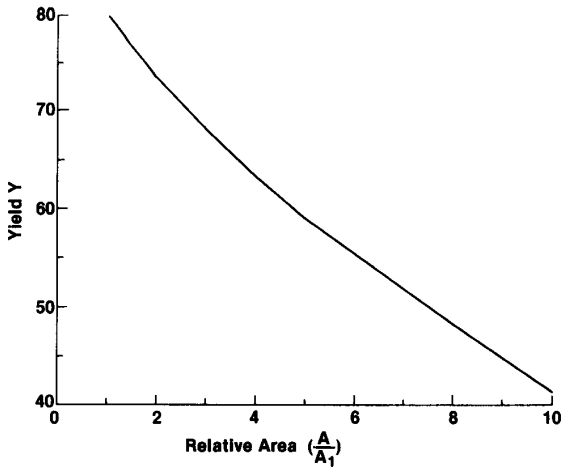

Fig. 3. Yield of multiple array-segment pseudochips versus ratio of their areas.

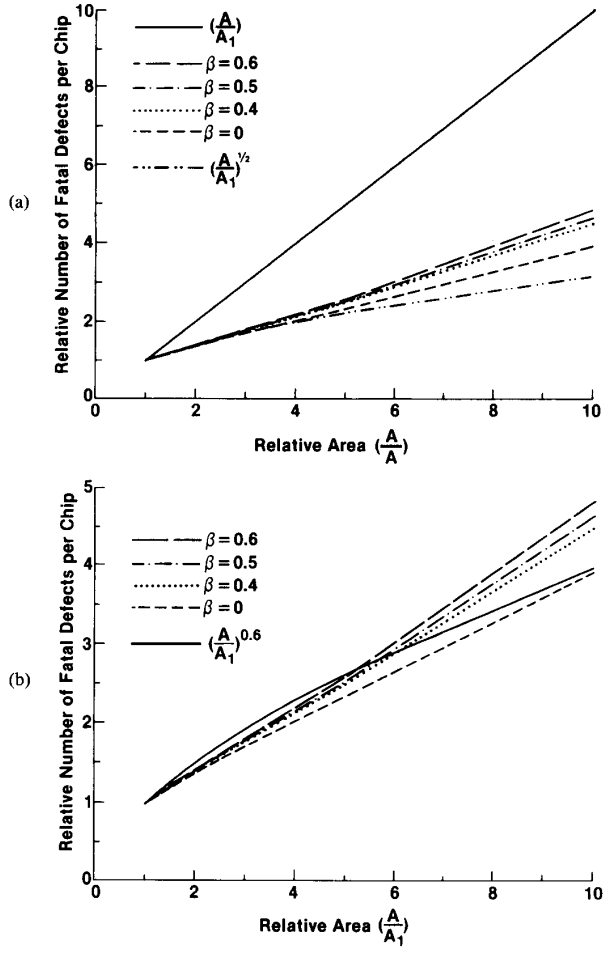

Fig. 4. Ratio of the average number of fatal defects per pseudochip versus ratio of pseudochip areas, with different values of the coupling coefficient β. (a) Shown with area scale factor with $b = 0$ and $b = 0.5$. (b) Shown with area scale factor with $b = 0.4$.

fatal defects per chip. (This is examined again in Section VII.)

Thus, the scaled average number of fatal defects per chip of the new product,

$$\lambda(A, w) = \xi \frac{\Phi(w)}{\Phi(w_e)} \left(\frac{A}{A_e}\right)^{1-b} p_e^{-1}\left(\frac{Y_{\text{obs},e}}{Y_{o,e}}\right), \quad (14)$$

where, as shown in (13b),

$$\frac{\Phi(w)}{\Phi(w_e)} \simeq \left(\frac{w}{w_e}\right)^{p-1}.$$

depends upon several assumptions, and discussions of yield models usually do not mention all of them. The value of the experiential complexity factor, ξ; the fault probability ratio, $\Phi(w)/\Phi(w_e)$, which depends upon the power, p, with which the defect size distribution is assumed to decrease with defect size; the exponent b, with which the average number of faults per chip deviates from linearity with chip-area ratio; the value of the coupling coefficient, β, that affects the spatial dispersion of fatal defects; and the assumed value of the Y_o term all affect the expected value of the number of fatal defects per chip on the new product. The effect of these assumptions on the predicted yield will be traced next.

VI. Examination of the Yield Prescription

Equation (5) shows that the random-defect-limited yield prescription resulting from the compound Poisson distribution is of the form

$$(1 + \beta\lambda)^{-1/\beta} = \sum \left(\prod_{s=0,1,2,\cdots,n-1} (1 + s\beta)\right) \frac{(-\lambda)^n}{n!}, \quad n = 0, 1, 2, \cdots. \quad (15)$$

For a new product, the value predicted by this prescription is increased either as the magnitude of the coupling coefficient β is increased or as the value of λ is decreased. In the absence of a physical model for the spatial dispersion of fatal defects, the magnitude of β needs to be assumed. A value of 0.5 is convenient as it lies midway between assuming that the defects are distributed with equispatial probability and assuming that they are all concentrated in one location.

The value of λ is decreased if either the scale factor, σ, is decreased or the inferred value of λ_e, the average number of defects per chip of existing product, is decreased. The magnitude of σ is decreased if the complexity factor, ξ, is decreased, if the power, p, with which the defect size distribution is assumed to fall off is decreased, or if the exponent, b, with which the number of fatal defects per chip deviates from linearity with chip area is increased. None of these three quantities is known exactly and assumptions based upon experience and inference need to be made.

The magnitude of λ_e, inferred from the actual observed yield, $Y_{\text{obs},e}$, of existing product decreases if either β_e, the coupling coefficient assumed to describe the spatial dispersion of fatal defects on the existing product, is increased or the term $Y_{o,e}$ is decreased. Neither of these two terms is known and assumptions about their value need to be made. The many assumptions contained, but not explicitly mentioned, in most treatments of yield models, are shown by formally writing the predicted yield of new product as

$$Y_{\text{obs}} = f(Y_o, \beta, \xi, p, b, Y_{o,e}, \beta_e; Y_{\text{obs},e}), \qquad (16)$$

where the only term that can be directly determined is $Y_{\text{obs},e}$. It should be noted that this expression is *not* a prescription from which the expected yield can be calculated, but rather a formal representation of the multiplicity of variables that are inherent in all yield models, though not usually explicitly mentioned.

The choice of yield prescription affects the predicted yield in two ways. Of the seven adjustable parameters, those pertaining to the process complexity factor, the defect size distribution, the area scale factor, and the Y_o terms are implicitly dependent upon the choice of yield model in that they are usually inferred from yield data. As assumptions about the spatial dispersion of fatal defects are explicitly contained in the terms β and β_e, it is instructive to examine the yield prescription for two cases: one where it is assumed that fatal defects are dispersed with equispatial probability, i.e., that there is no coupling between them or $\beta = 0$, leading to the simple Poisson model, and one where it is assumed that the coupling coefficient is positive, i.e., $\beta \geq 0$. The latter leads to the negative binominal model, which is a particular type of compound Poisson model. The series expansion of (15) shows that the difference between its value for $\beta > 0$ and for $\beta = 0$ is given by

$$\Delta = \sum \delta_n, \qquad n = 2, 3, 4, \cdots \qquad (17a)$$

where

$$\delta_n = \left(\prod (1 + s\beta) - 1\right)(-\sigma\lambda_e)^n/n!, \qquad s = 0, 1, 2, \cdots, n - 1. \qquad (17b)$$

Assumptions about the spatial dispersion of defects on the new product are contained in the term β, assumptions about the spatial dispersion of defects on existing product are reflected in the value of λ_e, and assumptions about the process complexity, the defect size distribution, and the area factor are contained in the scale factor, α.

The magnitude of the difference Δ is less than the magnitude of the first term in its series expansion. In fact, for many cases, the difference between the yield predicted when $\beta = 0$ and $\beta > 0$ is negligibly small. This is shown in the next section, where predictions of the yield prescription with different values of β are compared with actual data.

VII. Comparison with Data

Fig. 5 shows the actual yield and that predicted via the Poisson model ($\beta = 0$) and the negative binomial model with an intermediate value of 0.5 for the coupling coefficient, β. In both cases, an intermediate value of $b = 0.5$ has been used in the exponent of the area scale factor. The agreement is not good for the larger pseudochips. When the value of b is decreased to 0.4 and with values of 0.4, 0.5, and 0.6 for β, the predictions of the negative binomial model are closer to the actual yield, as shown in Fig. 6. Decreasing the value of b in the area scale factor depresses the predicted yield, whereas increasing β increases the predicted yield slightly. By adjusting the values of β, β_e, and b, the magnitude of λ_e, i.e., of the average number of fatal defects per single array segment pseudochip inferred from its yield, can be altered. Fig. 7 shows the predictions of the simple Poisson model with $b = 0.45$ and the predictions of the compound Poisson model with $b = 0.4$ and $\beta = 0.4$. Both are in substantive agreement with the actual yield.

It is clear that there is more than one choice of coupling coefficient, and more than one choice of value for b in the area scale factor, that predicts yields in quite reasonable agreement with the actual yield. In this case, where the actual yield is known already, one can "tune" the parameters b and β, or even the term Y_o, to fit the data. However, the objective of this exercise is not to show which model can better **describe** known results, but to show that the **predictive** capability depends not only on the choice of model but also on the manner in which the average number of fatal defects per chip of existing product is scaled to the new product. In particular, there is sufficient flexibility in scale factor to allow one to neglect the coupling coefficient and to use the simple Poisson model and yet retain adequate predictive capability. A reason why the Poisson model, as used by some practitioners, may not have provided accurate predictions is that the average number of fatal defects per chip of new product may not have been scaled to reflect the area factor.

Examining the effect of the area scale factor shows that the choice of b that provides good predictions over one range of areas does not do so for another range of areas. In general one does not know *a priori* which is a good value of b to use. However, comparing the predictions based upon the yield of test structures of different areas with the actual yield of product provides yield practitioners with the experience needed to select a reasonable value of b.

It may be argued that the introduction of an area scale factor is at variance with the concept of defect density, which should be invariant with respect to area. In theory, defect density is defined as $D = \lim A \to 0\ N_D/A$, with N_D being the number of defects in a region of area A. In practice, the defect density is taken to be the ratio $\lambda/\Phi A$ [14], where λ is inferred via one yield prescription or another, and Φ is the (assumed) defect sensitivity of the chip of area A. Moreover, it is known [34]–[36] that regardless

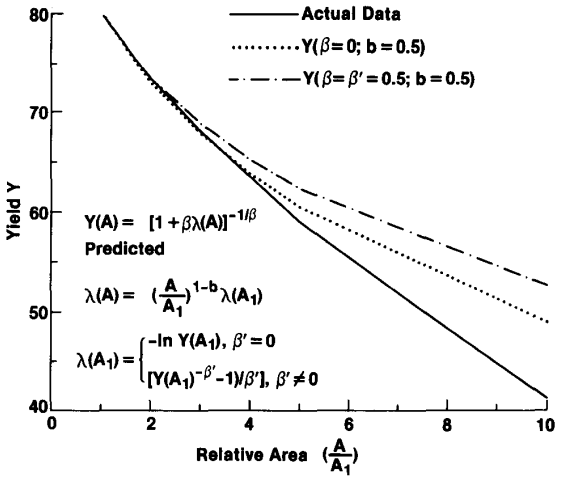

Fig. 5. Actual yield and yield predicted via the Poisson model with b = 0.5, and the negative binomial model with β = 0.5 and b = 0.5.

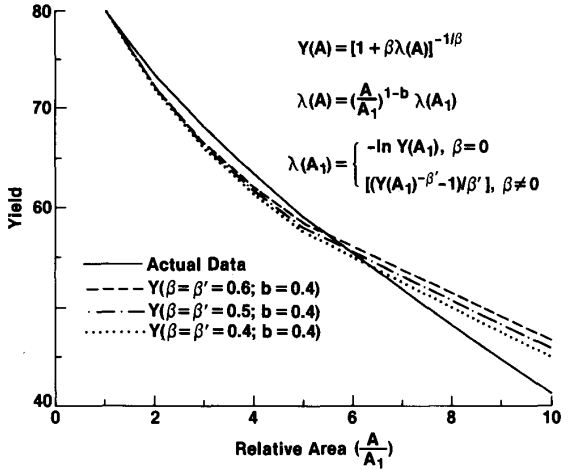

Fig. 6. Actual yield and yield predicted via the negative binomial model with b = 0.4 and β = 0.4, 0.5, and 0.6.

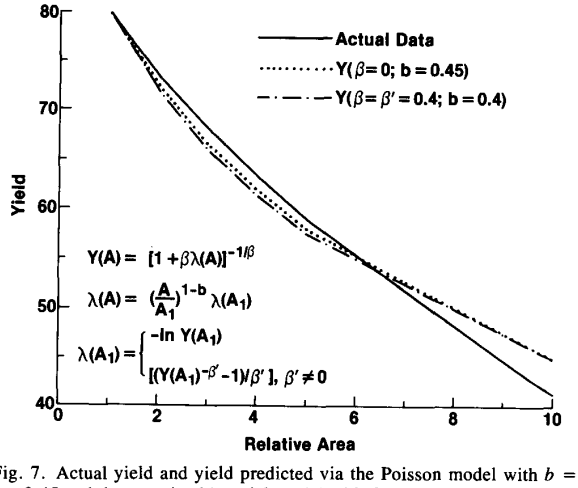

Fig. 7. Actual yield and yield predicted via the Poisson model with b = 0.45 and the negative binomial model with β = 0.4 and b = 0.4.

of the nature of the spatial dispersion of fatal defects, there is a pronounced radial dependence. Therefore, chips of successively larger area fabricated on the same wafer, as shown in the case of the pseudochips discussed earlier, experience an increased number of fatal defects, with the increase being somewhat less than that of the chip area. Thus the *wafer-averaged* number of fatal defects per chip for chips of different area does not increase linearly with chip area [37], as the concept of an invariant defect density would imply. Because of this, and because of the manner in which the defect density is obtained in practice, no basic physical concepts or definitions are violated by introducing an area scale factor.

In the example discussed, the only component in the scale factor is that of area. But in a more realistic prediction, where the new product is likely to have different minimum design features and may well be built in a different process if not technology, the process complexity and defect sensitivity scale factors will need to be used as well. It has been the experience of this practitioner, in a real manufacturing environment, that such a scaling enhances the predictive capability of the yield model used.

VIII. Discussion

To predict yield for a semiconductor product, essentially two pieces of information are needed: the average number of fatal defects per chip and a prescription from which to calculate the probability of obtaining a defect-free chip, i.e., a yield equation. By examining how well a particular yield equation fits existing data or predicts future results, many workers have attempted to determine which yield equation is better. However, the results are also affected by the value used in the yield equation for the average number of fatal defects per chip, but not much attention appears to have been paid to this fact.

The form of the yield equation depends upon how defects are dispersed over the wafer and reflects the assumptions made about the nature of the interaction between defects. The interaction may be inhibitory or attractive, or there may be no interaction at all. The average number of fatal defects per chip of new product needs to be obtained from the yield of existing product or test structures and necessarily reflects assumptions about the spatial dispersion of defects on that product or test structure.

For the independent, sequential process steps typical of semiconductor device fabrication, in the absence of any physical model to the contrary, it is reasonable to assume that there is no interaction between defects, leading to the Poisson model. But applying the Poisson model to predict the yield of new products of larger area than existing product from which the average number of fatal defects per chip is inferred has consistently been reported to predict lower yield than has later been observed. This discrepancy, together with the observation that defects frequently occur in clusters, particularly near the periphery of the wafer, has led to the development of cluster models,

the predictions of which have been reported to agree better with later observed results. However it is not only the yield equation but also the average value of fatal defects per chip used in the yield equation that affects the predicted value of yield. This average value is not independently known ahead of time and must be scaled from the value **inferred** from the yield of existing product, with the scale factor reflecting differences in process complexity and chip area, as well as defect size and spatial distributions, between new and existing product. It is not clear that the previously reported predictions based on the Poisson model have indeed accounted for these differences. In particular, the use of the term "defect density," with its implications of area independence, may tend to suggest that the average number of fatal defects per new chip scales linearly with the ratio of areas of new to existing chip, an assumption that has been shown [15], [37] to not always be valid.

For a product yet to be built, neither the spatial dispersion of fatal defects nor the size distribution of all defects is known *a priori*, and both have to be **inferred** from the observed **net** yield of an existing product or test structure. These inferences involve assumptions pertaining to the yield loss caused by factors other than defects, assumptions pertaining to the nature of the interaction between defects, and assumptions pertaining to the form of the defect size probability density function.

IX. Conclusion

This paper shows that the accuracy of yield predictions depends just as much on the accuracy of the assumed average number of fatal defects per chip as upon the choice of yield model. It shows that the nature of the spatial dispersion of fatal defects is not known *a priori* and that *a posteriori* attempts at verification are obscured by assumptions about the factors used to scale the average number of defects per chip of existing product inferred from its observed net yield, as well as by assumptions about the defect size distribution. Determination of the defect size distribution in turn requires assumptions about the spatial dispersion of all defects. It is noted that the often used inverse cube size density function is convenient, as it is essentially neutral with respect to yield and permits the use of other factors to make minor adjustments to the average number of defects per chip. The predictive capabilities of two commonly used yield models, the simple Poisson and a compound Poisson also known as the negative binomial model, have been compared. For the example investigated it has been shown that when appropriate scaling rules are used, the predictions of the simple Poisson model are accurate for chips of area up to at least an order of magnitude larger than that of the product for which yield data exist, suggesting that the Poisson model which has fewer adjustable parameters may be more efficient to use.

It has been said [46] that the attributes of a good model are that it should contain few arbitrary parameters, should accurately describe a large class of existing observations, and should make definite predictions about the results of future observations. For the case presented in this paper, the Poisson model with the average number of defects per chip scaled as described appears to meet these criteria.

References

[1] T. J. Wallmark, "Design considerations for integrated electron devices," *Proc. IRE*, vol. 48, pp. 293-300, Mar. 1960.

[2] J. E. Price, "A new look at yield of integrated circuits," *Proc. IEEE*, vol. 58, pp. 1290-1291, Aug. 1970.

[3] S. M. Hu, "Some considerations in the formulation of IC yield statistics," *Solid State Electron.*, vol. 22, no. 2, pp. 205-211, Feb. 1979.

[4] T. Michalka, R. Varshney, and J. Meindl, "A discussion of yield modeling with defect clustering, circuit repair and circuit redundancy," *IEEE Trans. Semiconductor Manufacturing*, vol. 3, pp. 116-127, August 1990.

[5] B. Ciciani and G. Iazeolla, "A Markov chain-based yield formula for VLSI fault-tolerant chips," *IEEE Trans. Computer-Aided Design*, vol. 10, pp. 252-259, Feb. 1991.

[6] S. Kikuda, H. Miyamoto, S. Mori, M. Niiro, and M. Yamada, "Optimum redundancy selection based on failure related yield model for 64Mb DRAMs and beyond," in *ISSCC 1991 Dig. Tech. Papers*, Feb. 1991, p. 123.

[7] J. A. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Trans. Semiconductor Manufacturing*, vol. 3, pp. 60-71, May 1990.

[8] C. H. Stapper, "Fact and fiction in yield modeling," *Microelectron. J.*, vol. 20, nos. 1 and 2, pp. 129-151, Spring 1989.

[9] W. Maly, "Yield simulation—A comparative study," in *Proc. 1989 Int. Workshop Defect and Fault Tolerance in VLSI Systems* (Tampa, FL), Oct. 23-24, 1989.

[10] M. Rivier, "Random yield simulation applied to physical circuit design," in *Yield Modelling and Fault Tolerance in VLSI*, W. Moore, W. Maly, and A. Strojwas, Eds. Adam Hilger, 1988, pp. 111-124.

[11] W. Maly, A. J. Strojwas, and S. W. Director, "VLSI yield prediction and estimation: A unified framework," *IEEE Trans. Computer-Aided Design*, vol. CAD-5, pp. 114-130, Jan. 1986.

[12] M. B. Ketchen, "Point defect yield model for wafer-scale integration," *IEEE Circuits and Devices Magazine*, vol. 1, pp. 24-34, July 1985.

[13] A. Rogers, *Statistical Analysis of Spatial Dispersion. The Quadrat Method.* London, U.K.: Pion Ltd., 1974, ch. 2, pp. 12-18.

[14] A. V. Ferris-Prabhu, "Defects, faults and semiconductor yield," in *Defect and Fault Tolerance in VLSI Systems*, vol. 1, I. Koren, Ed. New York: NY: Plenum Press, 1989, pp. 33-46.

[15] A. V. Ferris-Prabhu, "A cluster-modified Poisson model for estimating defect density and yield," *IEEE Trans. Semiconductor Manufacturing*, vol. 3, pp. 54-59, May 1990.

[16] P. R. Montmort, "Essai d'analyse sur les jeux des hasards," 1714. Cited in J. Gurland, "Some applications of the negative binomial and other contagious distributions," *Amer. J. Public Health*, vol. 49, no. 10, pp. 1388-1399, 1959.

[17] R. A. Fisher, "The negative binomial distribution," *Annals of Eugenics*, vol. 11, pp. 182-187, 1941.

[18] G. E. Blackman, "Statistical and ecological studies in the distribution of species in plant communities. I. Dispersion as a factor in the study of changes in plant populations," *Annals of Botany, London* (New Series), vol. 6, pp. 351-370, 1942.

[19] H. S. Sichel, "The estimation of the parameters of a negative binomial distribution with special reference to psychological data," *Psychometrika*, vol. 16, pp. 102-127, 1951.

[20] J. Neyman and E. L. Scott, "A theory of the spatial distribution of galaxies," *Astrophys. J.*, vol. 116, pp. 144-163, 1952.

[21] T. Okabe, M. Nagata, and S. Shimada, "Analysis on yield of integrated circuits and a new expression for the yield," *Elec. Eng. Japan*, vol. 92, no. 6, pp. 135-141, Dec. 1972.

[22] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated circuit yield statistics," *Proc. IEEE*, vol. 71, pp. 453-470, Apr. 1983.

[23] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 2nd ed. New York, NY: Wiley, 1957, pp. 131-132 and pp. 59-60.

[24] R. B. Seeds, "Yield, economic, and logistic models for complex digital arrays," in *1967 IEEE Int. Convention Rec.*, part 6, Apr. 1967, pp. 61–66.
[25] V. Foard Flack, "Introducing dependency into IC yield models," *Solid-State Electron.*, vol. 28, no. 6, pp. 555–559, 1985.
[26] A. V. Ferris-Prabhu, "Fault probability and critical area in VLSI yield projection," IBM Tech. Rep. TR19.0562, 1982.
[27] A. V. Ferris-Prabhu, "Role of defect size distribution in yield modeling," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 1727–1736, Sept. 1985.
[28] A. V. Ferris-Prabhu, "Computation of the critical area in semiconductor yield theory," in *Proc. Electronic Automation Design Conf. (EDA84)*, Mar. 1984, pp. 171–173.
[29] A. V. Ferris-Prabhu, "Defect size variations and their effect on the critical area of VLSI devices," *IEEE J. Solid-State Circuits*, vol. SC-20, no. 4, pp. 878–880, 1985.
[30] S. Gandemer, B. C. Tremintin, and J-J. Charlot, "Critical area and critical level calculation in IC yield modeling," *IEEE Trans. Electron Devices*, vol. 35, pp. 158–166, Feb. 1988.
[31] J. Pineda de Gyvez and J. A. G. Jess, "Systematic extraction of critical areas from IC layouts," in *Proc. Int. Workshop Defects and Fault Tolerance in VLSI Systems* (Tampa, FL), Oct. 1989, pp. 27–39.
[32] S. Gandemer, B. C. Tremintin, and J-J. Charlot, "A method for determining critical areas and critical levels for IC yield estimation," in *Yield Modeling and Fault Tolerance in VLSI*, W. Moore, W. Maly, and A. Strojwas, Eds. Adam Hilger, pp. 101–110; also *IEEE Trans. Electron Devices*, vol. 35, pp. 158–166, Feb. 1988.
[33] S. W. Director, "Manufacturing-based simulation: An overview," *IEEE Circuits and Devices Magazine*, vol. 3, pp. 3–9, Sept. 1987.
[34] T. Yanagawa, "Yield degradation of integrated circuits due to spot defects," *IEEE Trans. Electron Devices*, vol. ED-19, no. 2, pp. 190–197, 1972.
[35] A. V. Ferris-Prabhu, L. D. Smith, H. Bonges, and J. K. Paulsen, "Radial yield variations in semiconductor wafers," *IEEE Circuits and Devices Magazine*, vol. 3, no. 2, pp. 42–47, Mar. 1987.
[36] C. L. Mallory, D. S. Perloff, T. F. Hassan, and R. M. Stanley, "Spatial yield analysis in integrated circuit manufacturing," *Solid State Technology*, vol. 26, no. 11, pp. 121–127, Nov. 1983.
[37] A. V. Ferris-Prabhu and M. Retersdorf, "The effect on yield of clustering and radial variations in defect density," in *Proc. 1989 Int. Workshop Defect and Fault Tolerance in VLSI Systems* (Tampa, FL), Oct. 23–24, 1989, pp. 40–50.
[38] C. H. Stapper, "Modeling of defects in integrated circuit photolithographic patterns," *IBM J. Res. Develop.*, vol. 28, no. 4, pp. 461–474, 1984.
[39] W. Maly, "Modeling of lithography related yield loss for CAD of VLSI circuits," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, no. 3, pp. 166–167, 1985.
[40] S. P. Billat, "Automatic defect detection on patterned wafers," *Semiconductor International*, pp. 116–119, May 1987.
[41] W. Maly, F. J. Ferguson, and J. P. Shen, "Systematic characterization of physical defects for fault analysis of MOS IC cells," in *Proc. 15th Int. Test Conf.*, 1984, pp. 390–399.
[42] W. Maly, M. Thomas, J. Chinn, and D. Campbell, "Characterization of type, size and density of spot defects in the metallization layer," in *Yield Modeling and Fault Tolerance in VLSI*, W. Moore, W. Maly and A. Strojwas, Eds. Philadelphia, PA: Adam Hilger, 1988, pp. 71–91.
[43] A. V. Ferris-Prabhu, "Yield implications and scaling laws for submicrometer devices," *IEEE Trans. Semiconductor Manufacturing*, vol. 1, pp. 49–61, May 1988.
[44] W. Maly, M. Thomas, and J. Chinn, "Double-bridge test structure for the evaluation of type, size and density of spot defects," Tech. Rep. CMUCAD-87-2, Carnegie Mellon University, 1987.
[45] A. V. Ferris-Prabhu, "Forecasting semiconductor yield," in *Proc. Int. Conf. Computers, Systems and Signal Processing* (Bangalore, India), Dec. 1984, paper R46.13.
[46] S. W. Hawking, *A Brief History of Time*. New York: Bantam Books, 1988, p. 9.

Albert V. Ferris-Prabhu (SM'74) has a Ph.D. in physics and both an M.S. and a B.S. in engineering.

He is with IBM in Essex Junction, VT, and also serves as Adjunct Professor at the University of Vermont. His work in magnetism, computational physics, nonvolatile semiconductor memories, reliability, diffusion kinetics, and semiconductor device yield modeling has been reported in over a hundred publications.

Dr. Ferris-Prabhu is a fellow of the American Association for the Advancement of Science and a member of the American Physical Society, the New York Academy of Sciences, Tau Beta Pi, and Sigma Xi.