Derivation and Implication of a Novel DRAM Bit Cost Model

Haruo Nakatsuka, *Senior Member, IEEE*

***Abstract*—A model for the cost/performance of a large-scale integrated circuit (LSI) is derived using critical area with $1/x^3$ defect size distribution and common industry trends for device parameters and process parameters. The model predicts that dynamic random access memory bit cost will begin to increase sometime after 2005, if the current bit capacity increase rate of four times every three years remains effective. It is suggested to reduce the rate to two times every two years, which will ensure a bit cost reduction beyond 2010. However, if the defect density can be reduced faster than the past trend, a four times bit capacity increase every three years can still remain cost effective.**

***Index Terms*—Critical area, defect density, defect size distribution, yield model, DRAM cost trend, technology roadmap.**

I. Introduction

LARGE-SCALE INTEGRATED circuits (LSI) have been enhanced by the miniaturization of feature sizes, as well as the integration of an increasing number of transistors per chip. The rate of miniaturization and integration has been exponential. In the case of memory, for example, the rate of integration has been four times every three years. Other parameters, such as minimum feature size, chip size, bit cost, etc., have shown similar exponential characteristics. Such exponential rates have been intrinsic to the semiconductor technology trend, and are reflected in the technology road map used to specify future generations of technology development.

The technological barrier to realize gigabit level integration with sub-0.1-μm feature sizes is becoming more difficult to overcome, and in turn the cost of manufacturing and technology development is increasing dramatically. Chip yields will be deteriorated by the integration of larger numbers of elements with smaller feature sizes. The manufacturing equipment and facilities for such advanced LSI will become more sophisticated and more expensive. Under these circumstances, it will become difficult to maintain the exponentially improving cost/performance trend of the past decades. The industry is already deviating from this trend. For example, 256M dynamic random access memory (DRAM), which was supposed to be introduced in 1999, was not commercialized. Instead chip shrinks of 64M DRAM continued to be utilized in the industry.

In this paper, a model for the cost/performance of future LSI is proposed under the assumption that the industry continues to follow the past trend. Although the actual LSI cost consists of various components such as chip cost, packaging cost, test cost and the technology development overhead, only the chip cost is addressed in the following discussions since it constitutes the majority of the LSI cost. As for the technology development overhead, cost sharing models have been adopted in the industry as a solution [1].

Manuscript received November 28, 2000; revised December 18, 2001.

The author was with Toshiba Corporation, Tokyo, Japan. He is now with Applied Materials Japan Inc., Tokyo 100-0004, Japan (e-mail: Haruo_Nakatsuka@amat.com).

Publisher Item Identifier S 0894-6507(02)04463-9.

II. Modeling of the Semiconductor Technology Trend

A. Yield Model

The average number of defects on a chip has been assumed to be the product of the chip size and the defect density. With the concept of the critical area, however, the effective defect number λ is expressed as

$$\lambda = \int D(x) A_C(x)\, dx \tag{1}$$

where x is the size of the defect, $D(x)$ is the defect density as a function of x and $A_C(x)$ is the critical area as a function of x [2], [3].

For the defect size distribution, $1/x^p$ is widely accepted [3], [4], and $D(x)$ is formulated as

$$D(x) = D_C \frac{a}{x^p} \tag{2}$$

where D_C denotes a parameter representing the environment of the manufacturing line and a is a normalizing factor. If a is defined as

$$\int_F^\infty \frac{a}{x^p}\, dx = 1$$

D_C represents the number of particles larger than F.
$A_C(x)$ is dependent on the design of the LSI. For x smaller than F which is minimum feature size or a fraction of it depending on failure criteria, $A_C(x)$ is zero regardless of the design [2], [3] and hence λ can be expressed as

$$\lambda = D_C A_{\text{eff}},$$

where

$$A_{\text{eff}} = \int_F^\infty (a/x^p) A_C(x)\, dx \tag{3}$$

is the effective critical area.

Suppose the chip is shrunk by a factor of S in the feature size, $A_C(x)$ will be shrunk similarly in the x axis by the factor of S

and will be shrunk in the vertical axis by the factor of S^2 since the $A_C(x)$ represents the area. Then, λ is given as

$$\begin{aligned}\lambda &= D_C \int_{\mathrm{SF}}^{\infty} (a/x^p) S^2 A_C(x/S)\, dx \\ &= D_C S^{3-p} \int_F^{\infty} (a/z^p) A_C(z)\, dz \\ &= D_C S^{3-p} A_{\mathrm{eff}} \end{aligned} \tag{4}$$

where $z = x/S$.

On the other hand, if the chip is enlarged proportionally by increasing the number of transistors by a factor of X, $A_C(x)$ will increase in proportion to X for the same x value and hence

$$\begin{aligned}\lambda &= D_C \int_F^{\infty} (a/x^p) X A_C(x)\, dx \\ &= D_C X A_{\mathrm{eff}}. \end{aligned} \tag{5}$$

In the case where both of the above instances occur simultaneously, the effective average number of defects on a chip is given as

$$\lambda = D_C S^{3-p} X A_{\mathrm{eff}} \tag{6}$$

where S is the rate of the design rule shrink, X is the rate of the transistor count increase and A_{eff} is the original effective critical area before the chip shrink and the transistor count increase. This provides a generalized formula for the effective critical area. The parameter p depends on the environment of the LSI fabrication line. The data reported from various fabricators shows a p value of about 3 [3], [5]. In the case $p = 3$, (6) becomes simpler as

$$\lambda = D_C X A_{\mathrm{eff}}. \tag{7}$$

This indicates that the effective critical area does not change by chip shrink, which agrees with the results reported in a special case with parallel stripe patterns [6]. Computer calculated A_{eff} for some DRAMs and microprocessors shows fairly good agreement with (7) [7].

There have been various yield models reported in the literature [8]. Although (6) or (7) can be applied to any yield model, the Poisson model is assumed here since it fits relatively well to the yields at well controlled fabs. Then, the yield is given as

$$Y = Y_S \exp(-D_C S^{3-p} X A_{\mathrm{eff}}) \tag{8}$$

where Y_S denotes the yield associated with nonrandom-defect failures such as systematic failures and parametric failures. In the case of $p = 3$, (8) becomes

$$Y = Y_S \exp(-D_C X A_{\mathrm{eff}}). \tag{9}$$

In the following discussions, formula (9) is used with an assumption that $p = 3$ holds valid even beyond 0.1 μm.

B. Chip Area Trend

Integrating larger numbers of transistors on a chip increases LSI chip area. Implementing smaller feature sizes decreases it. As a first order approximation, the chip area A can be expressed as

$$A = knF^2$$

where k denotes a coefficient unique to each product type, such as DRAM, n is the number of transistors and F is the minimum feature size. The exponential trend for n and F can be expressed as

$$n = n_0 \nu^t \tag{10}$$

$$F = F_0 \sigma^t \tag{11}$$

where ν is the rate of the transistor count increase per year, σ is the feature size shrink rate per year, t is time in years and n_0 and F_0 denote the number of transistors and the minimum feature size at $t = 0$. Then, the chip size trend is given as

$$A = A_0 \nu^t \sigma^{2t} \tag{12}$$

where $A_0 = kn_0F_0^2$ is the chip area at $t = 0$.
σ^t and ν^t correspond to S and X in (6), respectively. As is known by Moore's law, the minimum feature size shrinks by two thirds every three years and DRAM bit capacity increases four times every three years. These correspond to $\sigma = 0.87$ and $\nu = 1.59$ respectively.

C. Defect Density Trend

Extensive effort has been made to reduce the defect density in the semiconductor industry. However, the trend of defect density reduction has not been explicitly reported. It is assumed here that D_C also follows the exponential trend and it is expressed as

$$D_C = D_{C0} \kappa^t \tag{13}$$

where κ is the rate of defect density reduction and D_{C0} denotes D_C at $t = 0$. D_0 data have been widely reported in the industry. However, D_0, which was originally an average defect density [9], is usually calculated backward from the yield data, assuming $\lambda = D_0 A$. Since the defect sensitive area is the critical area rather than the whole chip area, the calculated D_0 gives a smaller value than D_C, as $D_0 = (A_{\mathrm{eff}}/A) D_C$. Equations (6), (12) and (13) give the trend of D_0 as

$$D_0 = (A_{\mathrm{eff}}/A_0) D_{C0} \sigma^{(1-p)t} \kappa^t. \tag{14}$$

This means κ cannot be extracted from D_0 trend data without knowing σ and p. There are, however, some industry survey reports, where the D_0 from same technology fabricators were tracked [10]. Since σ was supposed to be 1 in those cases, κ was estimated from those D_0 data. The extracted κ value was 0.8, although some data presumably affected by systematic failures or abnormal yield were excluded.

Combining (8), (12) and (13), the yield is expressed as

$$Y = Y_S \exp(-\lambda_0 \sigma^{(3-p)t} \nu^t \kappa^t) \tag{15}$$

where $\lambda_0 = D_{C0} A_{\mathrm{eff}}$. In the case of $p = 3$, (15) becomes

$$Y = Y_S \exp(-\lambda_0 \nu^t \kappa^t). \tag{16}$$

D. Chip Cost Trend

Chip cost is given as the processed wafer cost divided by the number of good dice per wafer. The number of available dice per wafer N is approximated as

$$N = (\pi/4)(d - \sqrt{A})^2/A$$

where d is the wafer diameter. The wafer diameter has been enlarged one and a half times in two technology generations. Because of the delay in the introduction of 300-mm wafers, however, it is argued whether three generations per wafer size increase might best represent the trend in the future. Therefore, the wafer diameter trend is formulated as

$$d = d_0(1.5)^{t/3m} \tag{17}$$

where m is the number of technology generations before the introduction of the next wafer diameter. A three-year cycle time is assumed between the technology generations.

Good dice per wafer N_Y is given as

$$N_Y = \frac{(\pi/4)(d - \sqrt{A})^2}{A} Y. \tag{18}$$

The processed wafer cost consists of various components such as depreciation cost, material cost, labor cost, utility cost, etc. Dominant, however, is the depreciation cost of the equipment investment, which has been increasing exponentially from generation to generation. Therefore, the processed wafer cost W is assumed also to follow the exponential trend and is expressed as

$$W = W_0 Z^{t/3} \tag{19}$$

where Z is the rate of increase of wafer cost between generations and W_0 is the wafer cost at $t = 0$. The amount of capital investment for new generation fabricators has been in the range of one and a half to two times higher than that of previous generation fabricators. From such consideration, it is assumed that $Z = 1.7$.

Equations (18) and (19) give the chip cost C as

$$C = \frac{W_0 Z^{t/3} A}{(\pi/4)(d - \sqrt{A})^2 Y}. \tag{20}$$

E. Cost/Performance of LSI

One of the indices for economical merit of LSI is the cost per transistor C_n, which is related to bit cost for memory and gate cost for logic LSI. From (10), (12) and (20), C_n is given as

$$C_n = \frac{A_0 W_0 Z^{t/3} \sigma^{2t}}{n_0(\pi/4)(d - \sqrt{A})^2 Y}. \tag{21}$$

Taking the ratio over C_{n0}, which is the cost per transistor at $t = 0$,

$$C_n/C_{n0} = \frac{Z^{t/3}\sigma^{2t}}{\{(d - \sqrt{A})/(d_0 - \sqrt{A_0})\}^2 (Y/Y_0)} \tag{22}$$

gives the relative improvement of the transistor cost. To simplify (22), it was assumed that $d \gg \sqrt{A}$ and $d_0 \gg \sqrt{A_0}$. Further assumed was that Y_S eventually reaches the same yield for any technology node by the efforts of process improvement and design refinement. Then, in the case of $p = 3$, (22) can be expressed using (16) and (17) as

$$C_n/C_{n0} = \frac{\sigma^{2t} Z^{t/3}}{1.5^{2t/3m} \exp\{\lambda_0(1 - \nu^t \kappa^t)\}}. \tag{23}$$

This provides a model for normalized bit cost trend in the case of memory LSI or normalized gate cost trend in the case of logic LSI.

For the DRAM case, however, there is a minor contradiction in the chip size model (12). It holds valid for same generation DRAM's incorporating the same memory cell structure. In new generation DRAM's, however, the memory cell has been shrunk to a smaller size than that of the optical shrink, by introducing more densely packed cell structures. Whereas the industry trend shows a one and a half times chip size increase for new generation DRAMs [11], the model (12) predicts a 1.7 times increase in 3 years with $\nu = 1.59$ and $\sigma = 0.87$. Therefore, the chip size model was modified for DRAM as

$$A = A_0 \gamma^t \nu^t \sigma^{2t}. \tag{24}$$

where γ is the shrink rate corresponding to the cell structure change and is estimated to be 0.96 from the above discussion. In a strict sense, (4) and (5) do not hold valid if a more densely packed cell structure is introduced. As a first order approximation, however, this effect is neglected because the cell array portion is less sensitive to defects by use of redundancy repair. Therefore, formula (23) is modified for DRAM as

$$C_n/C_{n0} = \frac{\gamma^t \sigma^{2t} Z^{t/3}}{1.5^{2t/3m} \exp\{\lambda_0(1 - \nu^t \kappa^t)\}}. \tag{25}$$

TABLE I
PARAMETERS USED IN DRAM BIT COST MODEL

Parameter	Value	Parameter	Value
ν	1.59	Z	1.7
σ	0.87	m	2
κ	0.8	λ_0	0.2

III. DRAM Bit Cost Trend

The bit cost trend for DRAM is calculated using formula (25). Table I summarizes the parameters used in the calculation. It should be noted that the calculated data are significant only at node points; between which the chips are only shrunk without increasing the actual DRAM bit capacity. The technology node point is defined as the year when the new generation DRAM enters into volume production. The reference year at $t = 0$ was chosen at 1999 because it would have been the year when both 256M DRAM production and 300-mm wafer introduction had started at the same time if the past trend had continued.

The wafer diameter parameter, m, was set to be 2 because it has been the trend in past decades. According to (5), λ_0 for 256M DRAM in 1999 can be estimated as four times that of 64M DRAM which utilizes the 256M DRAM cell structure. Although the actual yield data are not available from the industry, it is generally recognized that 64M DRAM chip yield was well over 80% in 1999.

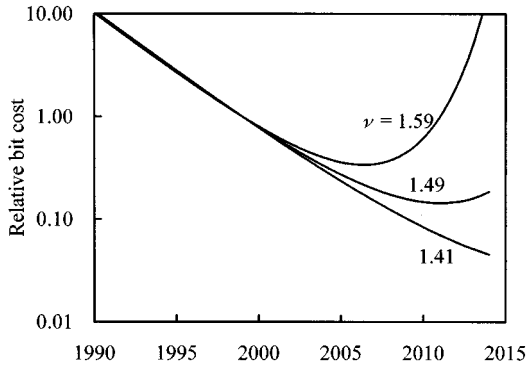

Fig. 1. Relative bit cost trend of DRAM with rate of bit capacity increase as parameter. ν denotes bit capacity increase rate per year. $\nu = 1.59, 1.49$ and 1.41 correspond to the bit capacity increase of four times per three years, four times per three and a half years and four times per four years or two times per two years, respectively. Reference year is 1999.

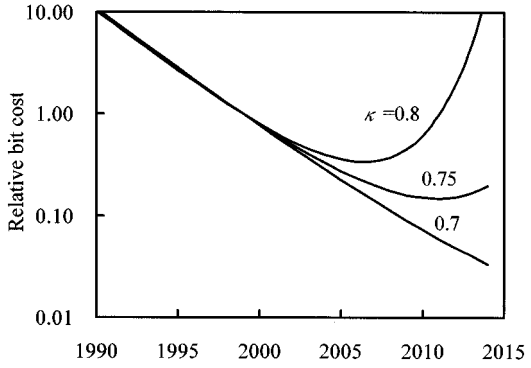

Fig. 2. Relative bit cost trend of DRAM with rate of defect density reduction as parameter. κ denotes the yearly reduction rate of the defect density parameter D_C. $\kappa = 0.8$ has been the industry trend. Bit capacity increase rate is four times in three years.

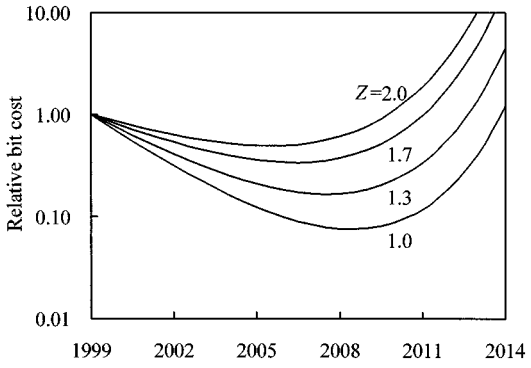

Fig. 3. Dependence of DRAM bit cost on rate of wafer cost increase. Z is a rate of the processed wafer cost increase between generations. $Z = 1$ corresponds to no wafer cost increase. Bit capacity increase rate is four times every three years.

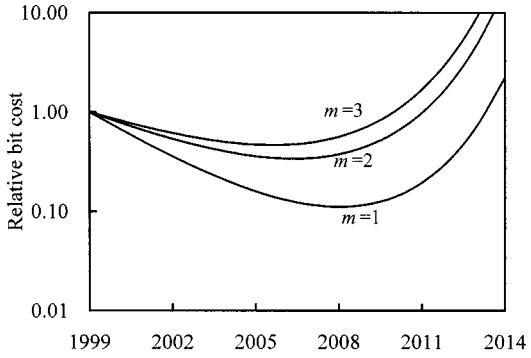

Fig. 4. Dependence of DRAM bit cost on timing of larger size wafer introduction. m is a number of technology generations before the introduction of next wafer diameter. $m = 3, 2$ and 1 correspond to nine-, six-, and three-year cycle time for larger diameter wafer introduction. Bit capacity increase rate is four times in three years.

Therefore, it would not be unrealistic to assume the random defect yield for 64M DRAM to be around 95%. This gives the value of 64M DRAM λ to be 0.05 and, hence, λ_0 for 256M DRAM is estimated to be 0.2. Those figures reflect the effect of the redundancy repair.

Fig. 1 shows the bit cost reduction trend with ν as a parameter. It is remarkable that the bit cost starts to increase around 2005 if the industry keeps the past trend of four times bit capacity increase every three years. With slower rates such as four times increase every four years, however, the bit cost reduction can be maintained beyond 2010. This rate corresponds to two times bit capacity increase every two years, which is coincidentally happening already in the industry where 128M DRAM was introduced instead of 256M DRAM.

The calculation was also made backward from the reference year. The results agree with the past trend regardless of the bit capacity increase rate, although rates other than four times every three years did not occur in reality. This means that the deviation from the past trend becomes significant when the feature size is shrunk beyond 0.1 μm.

The effects of other parameters were investigated. Fig. 2 shows the dependence of the bit cost trend on κ. If the defect density is reduced faster than the industry trend of $\kappa = 0.8$, the cost effectiveness will continue. With a defect density reduction rate of 30% per year, for example, the cost reduction extends toward year 2010. However, the feasibility of faster defect reduction is not clear at this time.

The cost dependence on Z and m are shown in Figs. 3 and 4, respectively. Fig. 3 indicates that the bit cost increase will occur even if the industry maintains the current wafer cost in the future. Introduction of larger size wafers does not prevent the bit cost increase as shown in Fig. 4. These results imply that investment strategy will have less influence on the future bit cost trend.

IV. Discussion

The results shown in Fig. 1 suggest that the industry should switch the integration strategy from four times bit capacity increase every three years to two times every two years. Because it demands a significant strategy change, the credibility of the above model should be ascertained.

Although some approximations and assumptions were made in the model, the most significant contributor to the bit cost increase in the future is the defect size distribution model (2) with $p = 3$, which leads to a larger λ value when the feature size is shrunk to much smaller dimensions. If the p value were much smaller, such as $p = 1$, however, the past trend could be maintained as shown in Fig. 5. One question is the extendibility of the $1/x^3$ distribution model to the sub-0.1-μm range. It was hypothetically assumed in earlier reports that defect density de-

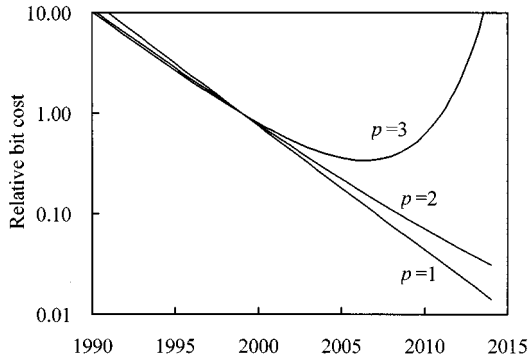

Fig. 5. Relative bit cost trend of DRAM with defect size distribution coefficient p as parameter. Defect size distribution of $1/x^p$ is assumed. $p = 3$ is widely observed in the manufacturing lines. Bit capacity increase rate is four times every three years.

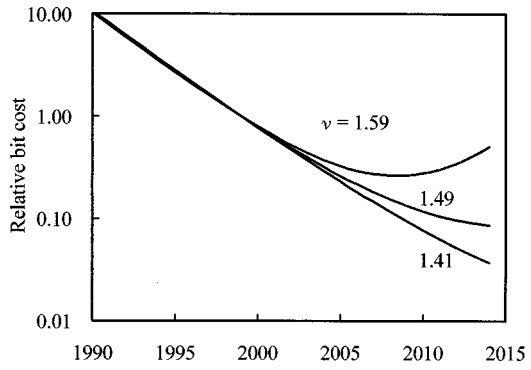

Fig. 6. Relative bit cost trend of DRAM with negative binominal yield model. The cluster parameter α is assumed to be 3.

creases linearly as the defect size becomes smaller than a certain size x_0 [2]. The value of x_0 was in the range of a few μm. With inspection technology advances, however, increasing numbers of particles, which follow the $1/x^3$ distribution, have been observed down into the 0.1 μm range [12], [13]. The total defect distribution on silicon chips has been reported to show the $1/x^3$ dependence down into the quarter μm range [14]. From those considerations, it would not be unreasonable to assume that the $1/x^3$ distribution extends beyond the sub-0.1-μm range.

In the actual manufacturing process, there are a multiple of defect sources, each of which may have a different size dependence. In such case, total defect density can be expressed as

$$D(x) = \sum_{i=1}^{n} D_{C_i} a_i x^{-p_i}. \tag{26}$$

This gives the yield formula as

$$Y = Y_S \Pi_{i=1}^{n} Y_i \tag{27}$$

where

$$Y_i = \exp\{-D_{C_i} S^{3-p_i} X A_{\text{eff}}(i)\}$$

which is similar to formula (8). Therefore, each Y_i gives a similar contribution to the bit cost trend as seen in Fig. 5. If certain Y_i with $p = 1$ or 2 became dominant in much smaller dimensions, the future bit cost trend would be more optimistic. However, it is not clear at this moment whether such defect sources exist nor whether they will become dominant beyond certain dimensions.

The assumption made in deriving (23) was that the nonrandom-defect yield Y_s reaches to the same level regardless of the technology node. However, systematic failures and parametric failures are becoming less easy to resolve as the feature size is shrunk to smaller dimensions. For example, lithography has become more sophisticated beyond the quarter-micrometer node, which makes yield improvement more difficult. With resolution enhancement techniques, particles smaller than the minimum mask size may cause failures due to interference. If these issues cannot be overcome at each technology node, the bit cost trend will become more pessimistic.

Another concern is the credibility of the yield model. In order to see the dependence on the yield model, the negative binominal yield model [2], which has also been widely used in the industry, was applied to the simulation. In this case, (25) is reformulated as

$$C_n/C_{n0} = \frac{\gamma^t \sigma^{2t} Z^{t/3} (1 + \lambda_0 \nu^t \kappa^t/\alpha)^\alpha}{1.5^{2t/3m} (1 + \lambda_0/\alpha)^\alpha} \tag{28}$$

where α is the cluster parameter. Results of the simulation using (28) are shown in Fig. 6. The same trends as in the case of the Poisson model are obtained although the bit cost increase is slightly moderated. Therefore, these yield models do not affect the future bit cost trend. However λ, the effective average number of defects on a chip, is a major factor to determine the trend.

V. Conclusion

A model for the future LSI cost trend has been derived assuming the exponential trend of device parameters and process parameters to continue in the future. Remarkable results were obtained for DRAM. The bit capacity increase rate of four times every three years would cause a bit cost increase sometime after 2005. It is proposed that the industry should switch to a new integration strategy of two times bit capacity increase every two years. This would ensure furthering bit cost reduction beyond 2010.

It is also suggested to accelerate the defect density reduction without exceeding the capital investment trend. If the defect density parameter D_C can be reduced at a much faster rate, or the defect size distribution parameter p can be reduced for defects in the sub-0.1-μm range, it can be an alternative solution to maintain the cost effectiveness of future LSI, although the feasibility is not clear at this time.

Acknowledgment

The author would like to thank S. Sugimoto of Semiconductor Process and Manufacturing Engineering Center, Toshiba, for his discussions on the effective critical area model.

References

[1] H. Nakatsuka, "DRAM development alliance," in *Proc. 5th Int. Symp. Semiconductor Manufacturing*, Oct. 1996, pp. 21–24.
[2] C. H. Stapper, "Modeling of integrated circuit defect sensitivities," *IBM J. Res. Develop.*, vol. 27, pp. 549–557, Nov. 1983.

[3] C. H. Stapper and R. J. Rosner, "Integrated circuit yield management and yield analysis: Development and implementation," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, pp. 95–102, May 1995.
[4] P. G. Parks and E. A. Burke, "The nature of defect size distribution in semiconductor process," in *Proc. IEEE/SEMI Semiconductor Manufacturing Science Symp. 1989*, 1989, pp. 131–135.
[5] C. Hess and L. H. Weiland, "Harp test structure to electrically determine size distributions of killer defects," *IEEE Trans. Semiconduct. Manufact.*, vol. 11, pp. 194–203, May 1998.
[6] C. H. Stapper, "Modeling of defects in integrated circuit photolithographic patterns," *IBM J. Res. Develop.*, vol. 28, pp. 461–475, July 1984.
[7] S. Sugimoto and H. Nakatsuka, "Yield analysis of LSI based on critical area model," *Trans. IEICE*, vol. J84-C, no. 9, pp. 888–893, Sept. 2001.
[8] C. H. Stapper, "Fact and fiction in yield modeling," *Microelectron. J.*, vol. 20, pp. 129–151, May 1989.
[9] B. T. Murphy, "Cost-size optima of monolithic integrated circuits," *Proc. IEEE*, vol. 52, pp. 1537–1545, Dec. 1964.
[10] R. C. Leachman, Ed., "Competitive Semiconductor Manufacturing Study, Third Report in the Results of the Main Phase," Univ. of California at Berkley, Aug. 1996.
[11] A. Nitayama, Y. Kohyama, and K. Hieda, "Future directions for DRAM memory cell technology," in *IEDM Technol. Dig.*, 1998, pp. 355–358.
[12] K. Okumura, "IC/LSI manufacturing," in *Proc. SEMI Technol. Symp. 1988*, 1988, pp. 14–18.
[13] K. Ichijo, S. Arasaki, T. Matsuda, N. Tsuda, and K. Tsubouchi, "0.1 μm particle detection with flow-cell-type laser particle counter," in *Proc. Fifth Int. Symp. Semiconductor Manufacturing*, 1996, pp. 188–191.
[14] C. Hess, D. Stashower, B. E. Stine, G. Verma, L. H. Weiland, K. Miyamoto, and K. Inoue, "Fast extraction of killer defect density and size distribution using a single layer short flow NEST structure," *Proc. IEEE 2000 Int. Conf. Microelectronics Test Structures*, vol. 13, pp. 57–62, Mar. 2000.

Haruo Nakatsuka (M'92–SM'01) received the B.S. and M.S. degrees in applied physics from Tokyo University, Tokyo, Japan, in 1964 and 1966, respectively.

He joined Toshiba Corporation in 1966. From 1969 to 1970, he was a visiting research associate at Stanford University, Stanford, CA. After managing several development programs on DRAMs and Microprocessors at Toshiba, he assumed Project Manager of IBM-Siemens-Toshiba DRAM development alliance from 1992 to 1993. He was Chief Research Director of Toshiba Research and Development Center from 1996 to 1999. He is currently with Applied Materials Japan, Inc.