An Overview of Manufacturing Yield and Reliability Modeling for Semiconductor Products

WAY KUO, FELLOW, IEEE, AND TAEHO KIM

This paper presents an overview of yield, reliability, burn-in, cost factors, and fault coverage as practiced in the semiconductor manufacturing industry. Reliability and yield modeling can be used as a foundation for developing effective stress burn-in, which in turn can warranty high-quality semiconductor products. Yield models are described and their advantages and disadvantages are discussed. Both yield–reliability relationships and relation models between yield and reliability are thoroughly analyzed in regard to their importance to semiconductor products.

Keywords—*Fault coverage, package level burn-in, semiconductor yield, wafer level burn-in, yield–reliability relation.*

I. INTRODUCTION

The historical breakthrough invention of the first integrated circuit (IC) was made by Kilby in 1958; the first commercial monolithic IC came on the market in 1961, the metal–oxide–semiconductor (MOS) IC in 1962, and the complementary MOS (CMOS) IC in 1963. The path of continued advancement of IC's is marked by distinct periods of small scale integration (SSI), medium scale integration (MSI), large scale integration (LSI), very large scale integration (VLSI), and ultra large scale integration (ULSI) [1]. Today, the microelectronics industry has evolved into the world's largest manufacturing business [2]. Table 1 [3] traces the development of IC technology and the associated growth in the number of transistors that can be integrated in a chip of dynamic random access memory (DRAM). After the year 2000, the IC industry will enter the super large scale integration (SLSI) era with over 10^9 transistors for 4G (and over) DRAM's.

The three major types of materials making up circuits are metals, insulators, and semiconductors, among which the basic difference is the magnitude of the energy gap [4]. Metallic materials form the conducting element. The insulator is the dielectric which provides isolation between transistors or between metals; it also serves as part of transistor structure. Table 2 depicts projections for the future development of IC devices and technologies. Nevertheless, reliability and yield will continue to be problems for the IC industry in the future.

Table 1
The Progressive Trend of IC Technology

Integration level	Year	Number of transistors	DRAM integration
SSI	1950s	less than 10^2	
MSI	1960s	$10^2 \sim 10^3$	
LSI	1970s	$10^3 \sim 10^5$	4K, 16K, 64K
VLSI	1980s	$10^5 \sim 10^7$	256K, 1M, 4M
ULSI	1990s	$10^7 \sim 10^9$	16M, 64M, 256M
SLSI	2000s	over 10^9	1G, 4G and above

Table 2
The Projected Trend of IC Devices

Device Type	1987	1992	2000
CMOS	39%	73%	82%
BiCMOS	0%	2%	6%
NMOS	24%	4%	less than 1%
PMOS	less than 1%	0%	0%
Bipolar (analog)	20%	14%	9%
TTL	12%	4%	1%
ECL	4%	2%	less than 1%
GaAs	less than 1%	less than 1%	1%

A. Behavior of Failures

Systems and materials begin to wear out during use, and various mechanisms can contribute to failure. Therefore, failure needs to be defined within specific bounds under specific tolerance limits. Early failures may come from poor design, improper manufacturing, or inadequate use. It is also known that failures result from the aging process; material fatigue, excessive wearout, environmental corrosion, and undesirable environment can contribute to this process.

Manuscript received June 15, 1998; revised March 1, 1999. This work was supported in part by the Texas Advanced Technology Program under Grant ATP-036327-138, the NSF Project DMI-9400051, and an IBM Headquarters Manufacturing Project.

W. Kuo is with the Zachry Engineering Center, Texas A&M University, College Station, TX 77843 USA.

T. Kim is with the Telecommunications Network Lab, Korea Telecom, Taejon 305-390 Korea.

Publisher Item Identifier S 0018-9219(99)05753-9.

A study of many systems during their normal life expectancies has led to the conclusion that failure rates follow a certain basic pattern. It has been found that systems exhibit a high failure rate during their initial period of operation, called the infant mortality period (usually one year for IC's). The operating period that follows the infant mortality period has a lower failure rate and is called the useful life period. At this period, the rate tends to remain constant (about 40 years for IC's) until the beginning of the next phase, called the aging period. Failures during the last period are typically due to aging or cumulative damage. Typical failure rate behavior follows a distribution known as the bathtub curve.

Most electronic devices exhibit a decreasing failure rate (DFR) in their early life; this results from weak individuals that have shorter lives than the normal (stronger) ones. The weak devices may come from improper operations by workers, a contaminated environment, a power surge of the machines, defective raw materials, ineffective incoming inspection, or inadequate shipping and handling. If the weak devices are released to customers or are used to assemble modules or systems, many of these defects will cause early failures; from our experience, quite a few failures can be observed in the first year for "immature" products. This early-stage high hazard rate is called infant mortality because the product is in its infancy and is not mature enough to be released. Note that infant mortality is viewed for the whole lot instead of for a single device. A single device will either fail or pass a test, whereas the failure rate of a lot may follow a decreasing pattern.

Generally, the mechanisms of semiconductor failures are classified into three main areas [5], [6]: electronic stress failures; intrinsic failures; and extrinsic failures. Electrical stress failures are user related, and the cause of such failures is generally misuse. Electrical overstress (EOS) and electrostatic discharge (ESD), due to poor design of equipment or careless handling of components, are major causes of electrical stress failures, which can contribute to the aging of components and the possibility of intrinsic or extrinsic failures. Since ESD is an event-related failure, it is not possible to do a screening test against it. A major problem of ESD damage is the formation of latent defects which are extremely difficult to detect.

Failures inherent to the semiconductor die itself are defined as intrinsic. Intrinsic failure mechanisms tend to be the result of the wafer fabrication which is the front-end of the manufacturing process. Intrinsic failures including crystal defects, dislocations and processing defects, gate oxide breakdown, ionic contamination, surface charge spreading, charge effects, piping, and dislocations are important examples of intrinsic failure mechanisms. Time-dependent oxide breakdown occurs at weaknesses in the oxide layer due to poor processing or uneven oxide growth. Failures of MOS devices due to oxide breakdown during device operational life are very high because it is impossible to screen most defective devices before they reach the market. It is important that any defective gate oxides be detected at the final testing stage. Contamination is introduced by the environment, human contact, processing materials, and packaging.

Extrinsic failures result from device packaging, metallization, bonding, die attachment failures, particulate contamination, and radiation of semiconductor manufacturing. That is, extrinsic conditions affecting the reliability of components vary according to the packaging and interconnection processes of semiconductor manufacturing. As technologies mature and problems in the manufacturer's fabrication lines are ironed out, intrinsic failures are reduced, thereby making extrinsic failures all the more important to device reliability.

B. Removing Infant Mortalities Through Burn-In

Accelerated life tests that subject units to higher than usual levels of stress (e.g., voltage, temperature, humidity, pressure, and loading) are used to speed up the deterioration of materials or electronic components so that analysts are able to collect failure information more quickly.

About 40% of the microelectronics failures are reportedly due to temperature; vibration is the second highest factor, which accounts for 27% of the total failures; moisture accounts for 19%; sand and dust, 6%, salt, 4%, altitude, 2%, and shock, 2%. In other words, temperature is the most critical factor for component failure; this is especially true for semiconductors. Burn-in, a screening technique performed by applying high temperature and voltage early on product life cycle to remove latent defects, is found to be useful for highly IC systems [7]–[9]. By running test patterns, defective items can be found and removed. Burn-in time is the most important variable in burn-in experiments since it is directly related to cost [10]–[12].

Because the infant mortality of semiconductor products is high in terms of failure rate and long in terms of mortality period, burn-in at the factory has been widely exercised. According to Kuo and Kuo [9], the key questions for effectively exercising burn-in are the following.

1) How far should we go to reduce infant mortality by burn-in?
2) What are the savings from burn-in?
3) Under what environmental conditions should burn-in be performed?
4) Should burn-in be accomplished at the system, subsystem, or component level?
5) Who should be in charge of burn-in—the vendor, buyer, or a third party?
6) What would be the expected life after burn-in? How does it differ from the expected life without burn-in?
7) Is burn-in always necessary and economical?
8) Are there any side effects of burn-in?
9) How will the industry benefit from burn-in data?

As in MIL-STD-280A [13], several levels in a system have been defined: Chien and Kuo [14] and Whitbeck and Leemis [15] apply burn-in at the three-level (component, subsystem, and system level) and the two-level (component and system) system, respectively. Extremely

high system reliability can be achieved by burning in at all levels, although the component-level burn-in is generally performed by the vendors. For example, a 4M DRAM used in a personal computer (PC) can be viewed as a component. Sixteen 4M DRAM's are assembled on a printed circuit board (PCB) called a single in-line memory module (SIMM) to save space and to meet the motherboard specifications; the SIMM is then treated as a subsystem. Most major computer manufacturers require their DRAM and SIMM suppliers to perform burn-in and other environmental as well as electrical tests, to ensure quality of the incoming components. Finally, SIMM's are put on the motherboards for system-level tests; one frequently used test is to continuously open and close many windows and repeatedly execute selected programs or software to verify that the systems (PC's) under evaluation work successfully.

The importance and related costs of burn-in tests are discussed by Kuo [16]. Chien and Kuo [17] introduce an optimal burn-in strategy at different levels. In practice, burn-in, which may also be called the high temperature operating life (HTOL) test, is required by all semiconductor manufacturers for almost all products. Leemis and Beneke [18] provide a review of burn-in models and methods.

One other important issue in system reliability is incompatibility [19]. The incompatibility factor, which exists not only at the component level but also at the subsystem and the system level, is used to incorporate reliability loss due to poor manufacturability, workmanship, and design strategy. Chien and Kuo [17] propose a nonlinear model to: 1) estimate the optimal burn-in times for all levels; 2) determine the number of redundancies for each subsystem; and 3) model the incompatibility removal process.

Chien and Kuo [14] present a nonparametric approach that easily estimates the optimal system burn-in time without going through complex parameter estimation and curve fitting. However, this technique can only be applied when abundant failure data exist, which is not the case for new or expensive products. Hence, the Bayesian concept should be absorbed into the burn-in models when only limited data are collected because the Bayesian approach can handle the following three critical issues: 1) high testing costs of IC's; 2) the incorporation of experts' opinions; and 3) the reflection of degree of belief. The Dirichlet distribution, which is a natural conjugate prior for a multinomial likelihood and is a multivariate generalization of the beta distribution [20], is one of the famous models used in nonparametric Bayesian analysis.

In the IC industry, samples used for tests can be wafers, bare dice, or packaged devices. The package level tests use packaged devices such as samples. Presently, almost all burn-ins are done at the package level with the sample called device under tests (DUT). Chien and Kuo [21] use DUT to denote the sample put-in test. They extend the model developed by Mazzuchi and Singpurwalla [22] and apply their ideas on burn-in analysis to determine the system burn-in time.

C. New Techniques for Reliability Improvement

From the manufacturing standpoint, today's process technologies for deep-submicron devices are gradually approaching the physical limits. With current technologies, it is difficult to achieve high performance, high packaging density, and high reliability all at the same time [23]. In addition, a manufacturing factory requires a high initial investment and extremely high operation cost. As a consequence, developing new techniques to reduce costs becomes urgent. From a reliability point of view, current accelerated life tests and end-of-line failure analysis (FA) become less effective as the chip size is miniaturized [24]. The simple FA method of sampling the output of a manufacturing line must be shifted to new methods in order to better understand and control the input variables at each point in the manufacturing process [23]. The requirement of new techniques leads to the development of built-in reliability (BIR), wafer level reliability (WLR), qualified manufacturing line (QML), and physics-of-failure (POF) approaches [4], [25], [26].

To minimize reliability testing effort and to achieve target failure rates, reliability structures and high manufacturing yield must be taken into consideration when products are designed. Hu [25] defines BIR as a methodology or philosophy for manufacturing highly reliable IC's, not by measuring the output at the end of production, but by controlling input variables that impact product reliability. The BIR approach thus achieves the reliability goal through the elimination of all possible defects from the design phase of the product. Although this approach requires high initial cost compared with reliability improvement through enhancing reliability screen tests, reliable products will result with low overall costs. Generally, the BIR approach is effective only beyond the crossover point, since product costs increase due to large testing costs at that point. The basic idea of BIR is not new. However, the systematic use of it, and the recognition of its benefits, has only recently been reported. Some useful tools for BIR are statistical process control (SPC), WLR, intelligent burn-in, in-line testing, and circuit reliability simulation.

Another trend in the semiconductor industry is to apply WLR tests to screening and reliability analysis because the traditional reliability approaches may not support enough test time or test parts to resolve failure rates as low as 10 FIT's (1 FIT $=$ 1 failure per 10^9 device hours). WLR is the highly accelerated stressing test performed at the wafer level and on the test structure [27]. Because the testing is performed at the wafer level to reduce the time and expense of packaging, WLR is significantly different from traditional approaches and represents a transition from the end-of-line concept toward the concept of BIR. There are some examples of WLR implementation into a production line or testing methods [27]–[32]. According to Turner [33], the purpose of WLR is not to predict a lifetime, but to detect the variation sources that might affect reliability. To achieve the objectives of the WLR approach, WLR needs fast and highly accelerated wafer level tests (called WLR

fast or stressed tests) which are designed to address each specific reliability failure mechanism. However, Crook [24] and Turner [33] point out limitations of the WLR fast test. Since the WLR fast test is performed at the end of the manufacturing line and is not sensitive enough to detect process drifts, the WLR fast test is not always an effective process control monitor for detecting variable drifts out of specification and for providing quick feedback [24]. Further, it can only be applied with a full understanding of the limitations of the stresses, according to Turner [33]. Another disadvantage is that at higher stress levels, the failure mode may be physically different from what would occur under normal use conditions [27].

Recently, under pressure to qualify small quantities of highly reliable circuits, the U.S. Department of Defense (DOD) changed its approach to IC reliability from the qualified product concept to qualified manufacturing line (QML) [25]. QML is another evolutionary step devised for the purpose of developing new technologies for earlier marketing, improving circuit reliability and quality, and doing so at reduced costs. In QML, the manufacturing line is characterized by running test circuits and standard circuit types [34]. Similar to BIR, understanding failure mechanisms and performing failure analysis are critical elements in implementing the QML concept. Therefore, the QML approach places a heavy emphasis on documentation. QML is another resolution for the recognition of the impracticality of qualifying individual products and the belief that reliability can be built into all products by a qualified manufacturing line [25].

The concept of POF has been widely used in engineering fields, where the chance of testing is restricted due to variation of sample size, product cost, and time-to-market. Since the traditional approaches, which are based for data acquisition and curve-fitting to standard reliability models, cannot provide timely feedback any more, most semiconductor manufacturers apply POF to electronic products. If we know fundamental mechanical, electrical, chemical, and thermal mechanisms related to failures, it is possible to prevent failures in new as well as existing products before they occur. For this reason, Schlund *et al.* [35] develop the POF model to deal with time-dependent dielectric breakdown (TDDB).

II. Burn-In of Semiconductors

Starting with the growth of the crystal and proceeding to packaging, the manufacturing process for microcircuits is completely integrated. Yield and reliability are the driving forces for the success of any manufacturing scheme for a new technology. Yield must be maximized for each processing step while at the same time maintaining failure-free operation in excess of 10^7 h [4]. Several test steps are required in order to ensure reliability of final products and customer satisfaction. Currently, the wafer acceptance test (WAT), wafer probe (WP), burn-in, final test (FT), and quality control (QC) test are widely used. The relationship between principal IC manufacturing processes, reliability, and yield is depicted in Fig. 1.

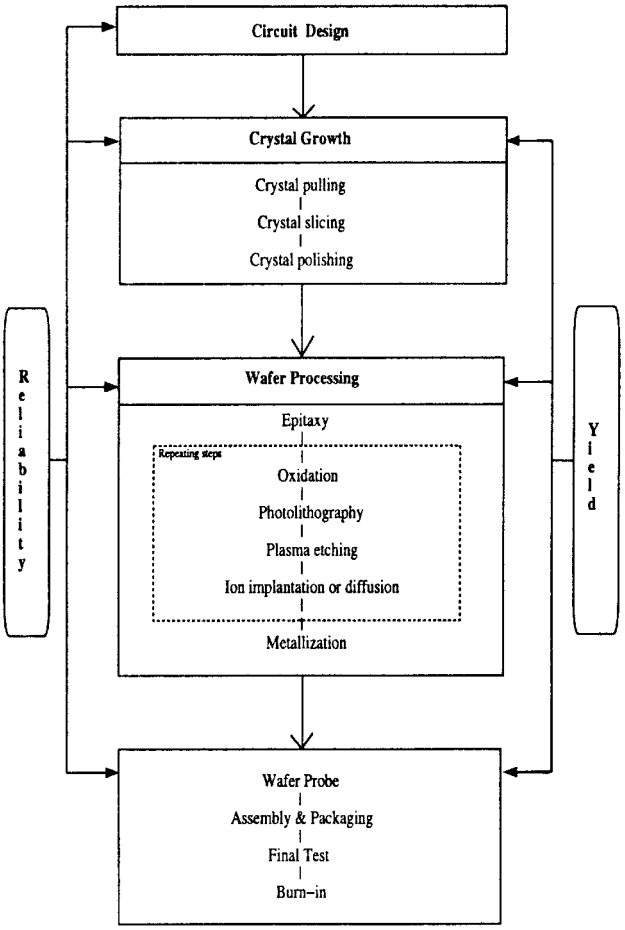


Fig. 1. Reliability's influence on the IC manufacturing process and yield (reprinted from [5] with permission from Kluwer Academic Publishers).

A. Burn-In Concepts

The burn-in test that subjects devices to higher than usual levels of stress such as voltage, temperature, and others is a technique used to speed up the deterioration of materials or electrical components so that analysts can collect information more promptly [5]. The test results have to be adjusted according to some time transformation models to provide predictions on the performance of the component in its normal use condition. The time transformation model can be chosen so that the relationship between the parameters of the failure distribution and the stressed condition is known.

Temperature is one of the most used physical mechanisms of the failure deterioration. If subscripts 1 and 2 are referred to normal conditions and accelerated conditions, respectively, and η is the time transformation factor, then the relationship between the time to failures under normal conditions, t_1, and accelerated conditions, t_2, can be expressed by

$$t_1 = \eta t_2. \tag{1}$$

Based on the Arrhenius equation and an activation energy of 0.4 eV, the η values for different temperature are given in Fig. 2. Relationship similar to (1) for the other stress factors are also available in Kuo *et al.* [5].

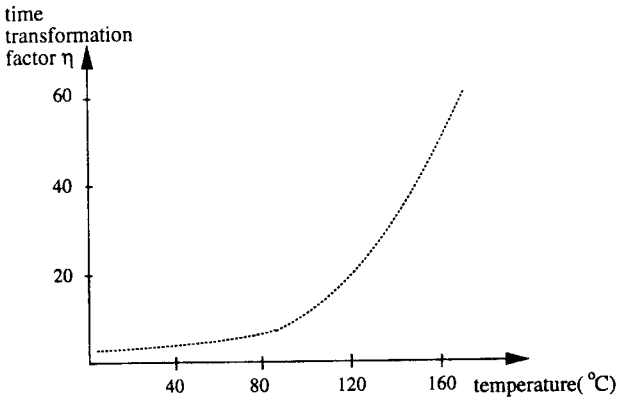


Fig. 2. The time transformation factor for different temperature at the activation energy of 0.4 eV.

B. Various Test Steps

WAT is an electrical test and is done at the wafer level right before WP. WP is often called chip probing (CP) or wafer sorting and its objective is to identify good bare dice on the wafer. Packaged dice which successfully passed the burn-in test will be ready for FT. During the FT, the full functionality of the product is checked. Usually, the test items in FT are similar but more complicated than the ones in WP. Many IC makers arrange burn-in between two FT stages. The FT stages before and after burn-in are sometimes called the pre- and post-burn-in tests, respectively; these two tests provide important information on the burn-in failure rate. The QC test is done on a sampling basis at the last stage before products are shipped to customers. Usually, visual inspection is an important part in a QC test. Most semiconductor products must go through WP, burn-in, and FT.

Assembled good chips which have passed function tests are put into special burn-in boards. These burn-in boards are then transferred to the burn-in chamber, where the chips are stressed to accelerate failure mechanisms. In general, it is known that burn-in is very effective in weeding out infant mortality failures [5], although burn-in can occasionally reduce manufacturing yields. The accelerated conditions, such as voltage, temperature, and burn-in time are critical factors determining the cost-effective burn-in. The need for burn-in depends upon the status of the product. Typically, new products require more extensive burn-in until the processes are sufficiently stable. A thorough analysis of cost-benefit burn-in is given in [5] and [9]; a first report from a system viewpoint on burn-in options appears in [16], and an optimal decision making model on the conceptual system burn-in is given in [36].

C. Burn-In Conditions and Types

During burn-in, IC's are tested under maximum electrical conditions with a typical temperature of 125 °C for 48, 96, 160, or 240 h, depending on the failure mechanism. To select a realistic burn-in method for an IC, we must know some basic conditions related to the IC [37], such as internal construction and fabrication of the chip, circuit function, circuit layout, number of actually activated and stressed circuit nodes, the fault coverage, possible failure modes and mechanisms, accelerating factors, and others. Hamilton [38] illustrates burn-in requirements for burn-in systems of more complex devices and test environments. For better results, parametric, nonparametric, and Bayes approaches are suggested in [14], [17], and [21].

Among burn-in approaches, four burn-in methods are particularly effective for semiconductor devices [5], [39]: steady-state burn-in (SSBI); static burn-in (SBI); dynamic burn-in (DBI); and test during burn-in (TDBI). It is known that SSBI and SBI are not effective for complex devices since external biases and loads may not stress internal nodes [39]. However, DBI places active signals on IC's which can propagate to internal nodes. TDBI is similar to DBI except it includes cycling with a functional test pattern. By conducting TDBI, manufacturers are able to monitor burn-in tests in real time [40].

When failures which are not temperature-dependent, like gate oxide breakdown, are not well detected by the normal burn-in, high voltage is often applied during the burn-in process. Many memory IC manufacturers are using high-voltage stress tests, SBI with reverse bias, and DBI to detect gate oxide defects [37].

There are three burn-in types based on levels of product [41], [42]: package level burn-in (PLBI); die level burn-in (DLBI); and wafer level burn-in (WLBI). PLBI is the conventional burn-in technology. DLBI serves for the burn-in of single IC die and WLBI for the entire wafer. Conventional burn in is sometimes carried out for packaged chips. Its primary advantage is to assure the reliability of final products. When parts that fail during the conventional burn-in must be scraped or abandoned, after they have gone through many process steps, the total product cost is likely to increase. In addition to the reduction of cost, the strong demand for known good die (KGD) is another motivation for developing more efficient burn-in technology. Conventional burn-in cannot support the burn-in of bare die. DLBI is the extension of the PLBI and uses most of the equipment and process of PLBI except die carrier and die handling capability. The integrity and cost of the carrier and handling process are the dominating factors in the decision making of DLBI. One advantage of DLBI is that it can provide burned-in and tested KGD. WLBI simultaneously puts stresses on every die of the entire wafer before performing burn-in. Since the burn-in process is performed at an earlier stage of product flow, WLBI can remove initial reliability failures earlier at reduced cost. Requirement of smaller size and lighter weight information system is a trend for multimedia era. However, the mounting technology in electronics systems has been advanced; therefore it is time for manufacturers to apply KGD to the market. WLBI is essential to produce KGD. Another advantage of WLBI is the fast feedback of yield and defect data, which makes manufacturing processes more proactive to fault corrections. Though the idea for applying burn-in at the wafer level may have originated from the need to provide conditioned (or burned-in) KGD, a successful

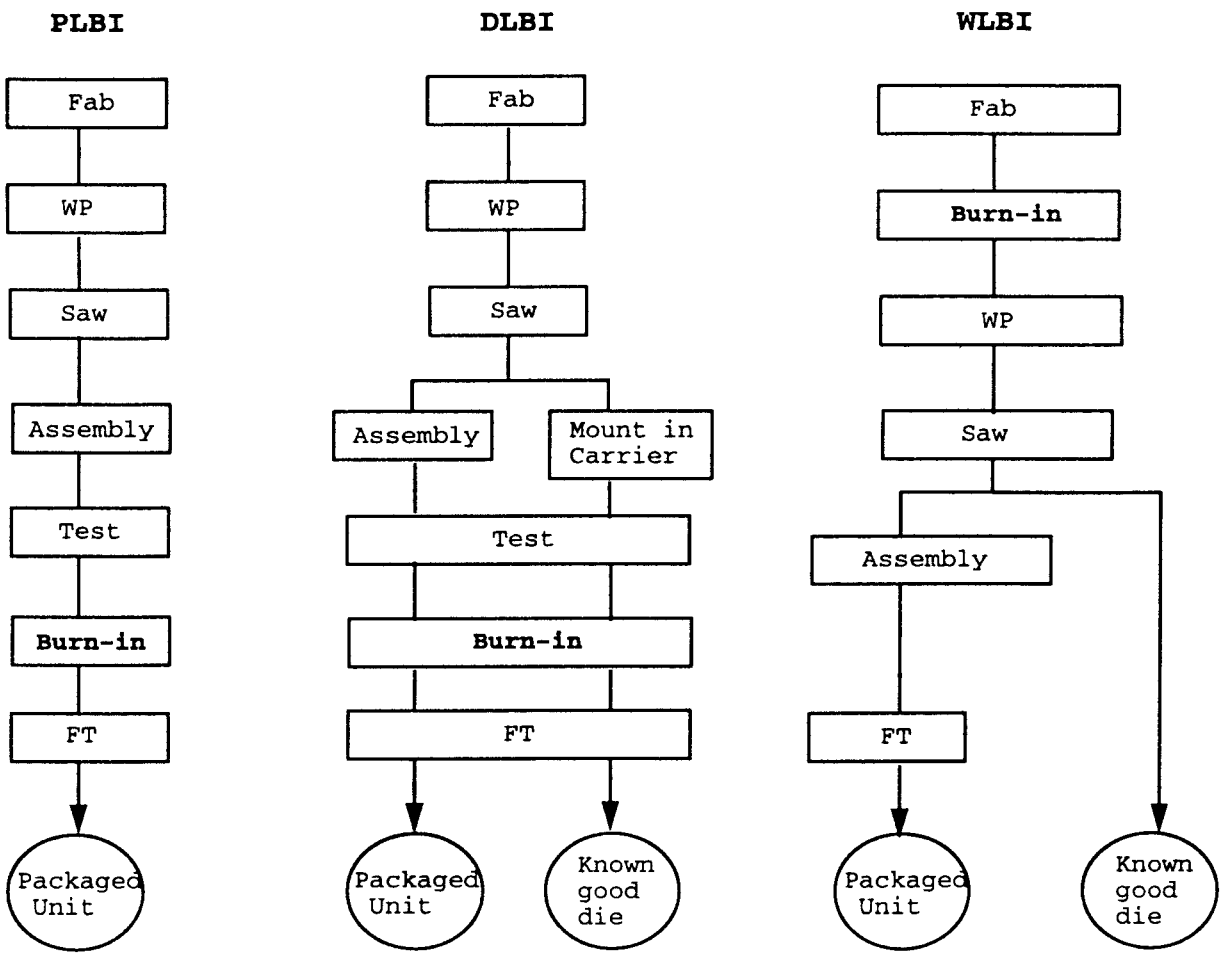


Fig. 3. Comparison of three burn-in flows.

WLBI results in considerable cost reduction for all IC products.

There exist some possible implementations of WLBI [41], [43]–[45]. Flynn and Gilg [46] present feasibility criteria for WLBI. However, building a whole wafer probing (or full wafer burn-in) capability is still a major technical challenge. The burn-in flows of three types are compared in Fig. 3 [41], [42]. The high initial cost of WLBI is a major concern of WLBI implementation. However, initial cost can be reduced by equipment cost reduction and better equipment centralization and utilization [41].

III. Modeling Yield

A. Yield and Reliability

Among the performance indexes for successful IC manufacturing, manufacturing yield is regarded as the most important one. Yield is usually defined as the ratio of the number of usable items after the completion of production processes to the number of potentially usable items at the beginning of production [47].

Since yield is a statistical parameter, yield functions at different manufacturing stages are multiplied in order to attain the total yield. Yield is not only a function of chip area but also a function of circuit design and layout. The total yield is a measure of good chips per wafer normalized by the number of chip sites per wafer. By determining the probabilities of failure and the critical areas for different defect types, it is possible to control and manage the yield of IC's [48]. Another way to control yield is to monitor defects. The number of defects produced during the manufacturing process can be effectively controlled by introducing test points at crucial times rather than throughout the assembly line [26]. This can significantly enhance the yield of the manufacturing process, improve the reliability of the outgoing product, and finally increase quality of the overall system.

Yield and reliability are two important factors affecting the profitability of semiconductor manufacturing. However, the correlation between them has not been clearly identified. There are three parameters that significantly affect the yield and reliability of IC's: design-related parameters such as chip area and gate oxide width; process-related parameters such as defect distribution and density; and operation-related parameters such as temperature and voltage. In general, reliability depends on all three parameters, whereas yield is affected by design and process-related parameters only. Therefore, we can conjecture that yield contains part of the information needed to predict reliability and that yield and reliability are correlated with each other. The basis

for the yield–reliability relationship and the relation model originates from this point of view.

Frost and Poole [49] develop a series model to determine the intrinsic reliability of IC's which shows the wearout-limited reliability based on defect concepts. Stevenson and Nachlas [50] use the physics-of-failure approach to derive the interrelationship between imperfections and the ultimate reliability of IC's. Jensen [51] shows that there exists a strong correlation between yield and reliability by surveying published papers and addressing yield models. He also argues that size and location of defects determine whether the defects are yield related or reliability related.

The presence of defects in IC's affects the yield as well as the reliability. Bruls [52] studies the reliability aspect of defects and devises the single-fault probability because he observes that the number of defects in a mature process is limited to one or a few and a single defect usually influences the reliability of an IC. Prendergast [53] points out a linear relationship between yield and reliability and suggests that this relationship can be effectively used to screen unreliable products. Another validation of the strong relationship between yield and reliability is presented by Van der Pol *et al.* [54]. Their research shows that a strong measurable relationship exists between the number of failures in the field (as well as in life tests), the yield due to the adoption of the WLR, and the use of reliability related design rules [54]. Thus, the root causes of reliability failures are the same as those of yield failures, and the manufacturing yield depends upon the number of defects found during the manufacturing process, which in turn determines reliability.

In order to reduce the cycle time and cost, rapid identification of yield losses and early elimination of the causes for losses are critical. El-Kareh *et al.* [55] emphasize that the process of reducing the chip size should be accompanied by improvement of yield in order to impact productivity.

IC device yields depend on many factors such as chip area, circuit design, circuit layout, and others. It is desirable to explain mathematically the overall yield and to control effectively and manage yields by determining the failure probabilities and the critical areas for each defect type [48].

B. Yield Component

Overall yield can be broken down into several components depending on the process grouping or the purpose of application. Here are four key yield components that are commonly used in semiconductor manufacturing: wafer process yield ($Y_{\rm wp}$); wafer probe yield ($Y_{\rm cp}$); assembly yield ($Y_{\rm ap}$); and final test yield ($Y_{\rm ft}$). According to one survey of Integrated Circuit Engineering Corp. (ICE) [56], the average values of wafer process yield and assembly yield are higher than those of wafer probe yield and final test yield. A schematic sequence of these yields and the typical average yield at each stage are presented in Fig. 4.

Sometimes, the line yield is interchangeably used with the wafer process yield and defined as the ratio between the numbers of wafers started and completed over a given production period. Cunningham *et al.* [57] subdivide the yield of a semiconductor into line yield, die yield, and final test yield. This yield categorization is very similar to Ferris-Frabhu's [47]. Generally, wafer fabrication processes directly affect wafer process yield (or line yield) and wafer probe yield (or die yield), and packaging processes influence assembly yield and final test yield.

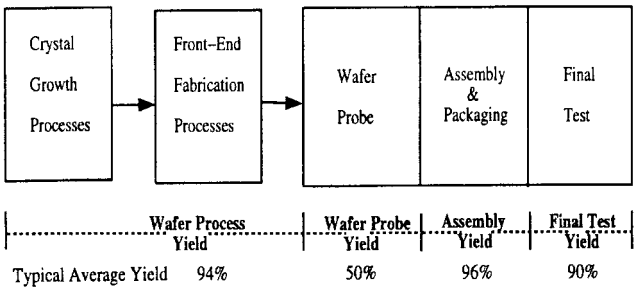


Fig. 4. Typical yield components commonly seen in semiconductor manufacturing.

The overall yield is defined as the product of yields from several consecutive processes, or [47], [55]

$$Y = Y_{\rm wp} \times Y_{\rm cp} \times Y_{\rm ap} \times Y_{\rm ft}. \tag{2}$$

Wafer process yield and wafer probe yield are two important factors influencing the productivity of semiconductor manufacturing. Most semiconductor industries focus on improving the low wafer probe yield especially because it is the bottle neck of the overall yield. To remain competitive, one must attain a high level of wafer probe yield.

C. Defects and Critical Area

For yield projection, it is useful to categorize defects as random or nonrandom defects [47], [55], [58]. Random defects are defects that occur by chance. Particles that cause shorts and opens or local crystal defects are random defects. Nonrandom defects include gross defects and parametric defects.

Defects which cause circuit failures are called faults or fatal defects [47], [59], [60]. The distinction between defects and faults plays an important role in calculating yield based on the defect density and chip area. Another parameter that affects yield is defect clustering.

The defect size distribution varies depending on process lines, process time, learning experience gained, and other variables. It is usually accepted that there is a certain critical size at which the density function peaks and then decreases on either side of the peak [64], [65]. Though there exist some distribution functions that behave like this, it is not easy to handle them analytically. Therefore, it is assumed that the defect size probability density function (pdf) is defined by a power law for defects smaller than the critical size and by an inverse power law for defects larger than the critical size [47]. Let x_0 be the critical size of the defect that is most likely to occur. The defect size pdf is defined as follows [66]:

$$s(x) = \begin{cases} c x_0^{-q-1} x^q, & 0 \le x \le x_0 \\ c x_0^{p-1} x^{-p}, & x_0 \le x \le \infty \end{cases} \tag{3}$$

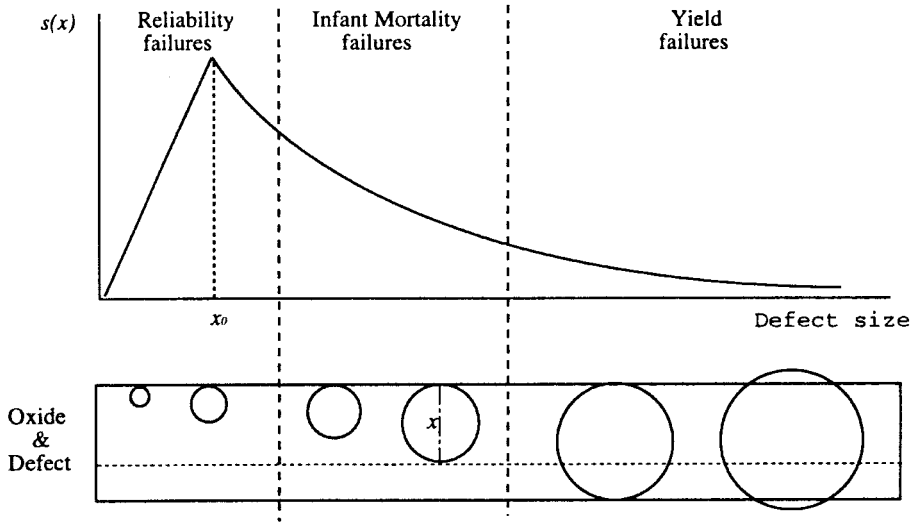


Fig. 5. The defect size pdf and related oxide problems (reprinted from [5] with permission from Kluwer Academic Publishers).

where $p \neq 1$, $q > 0$, and $c = (q+1)(p-1)/(q+p)$. It is experimentally shown that x_0 must be smaller than the minimum width or spacing of the defect monitor [65]. Defects smaller than x_0 cannot be resolved well by the optical monitoring [66]. Since very small defects are assumed to increase linearly with defect size to a point x_0, Stapper [65], [66] indicates that using values of $q = 1$ and $p = 3$ for the spatial distribution agrees reasonably well with experimental data. There are other observations for defect size distributions such as Rayleigh [67], lognormal [68], and gamma [69] distributions. A typical $s(x)$ versus defect size distribution curve for (3) is shown in Fig. 5, where the circles are the defects due to oxidation.

The critical area is an area where the center of a defect must fall to create a fault [65], [66]. That is, if a defect occurs in the critical area, then it causes a fault. Let $A_c(x)$ be a critical area of defect size x. The average critical area A_c is obtained with the integral form of

$$A_c = \int_0^\infty A_c(x)s(x)\,dx.$$

The average defect density of all sizes and average defect density of size x are defined as D_0 and $D(x)$, respectively. From the definition, the relationship between them is

$$D(x) = D_0 s(x).$$

Therefore, the average number of faults caused by defects, μ, is obtained by

$$\mu = A_c D_0.$$

D. Yield Models

A yield model is used to bridge from monitor to product, to bridge from product to product, or to predict yield before committing to a product [63]. That is, it is used to estimate the future yield of a current or new product and yield loss from each of the process steps. The Wallmark's model [70] is known as one of the earliest yield models. Among the models developed after this, the Poisson yield model and negative binomial yield model are most frequently used.

The Poisson model assumes that the distribution of faults is random and the occurrence of a fault at any location is independent of the occurrence of any other fault. For a given μ, the probability that a chip contains k defects is

$$P_k = \frac{e^{-\mu}\mu^k}{k!}, \qquad k = 0, 1, \ldots .$$

Since the yield is equivalent to the probability that the chip contains no defect

$$Y = P_0 = e^{-\mu} = e^{-A_c D_0}. \tag{4}$$

The Poisson yield model is widely used, but it sometimes gives a lower predicted yield than what is observed [47].

If the defect density is a random variable, the yield model is determined by the distribution that the defect density follows. The negative binomial model assumes that the likelihood of an event occurring at a given location increases linearly with the number of events that have already occurred at that location [71].

Assume that the defect density follows a gamma distribution

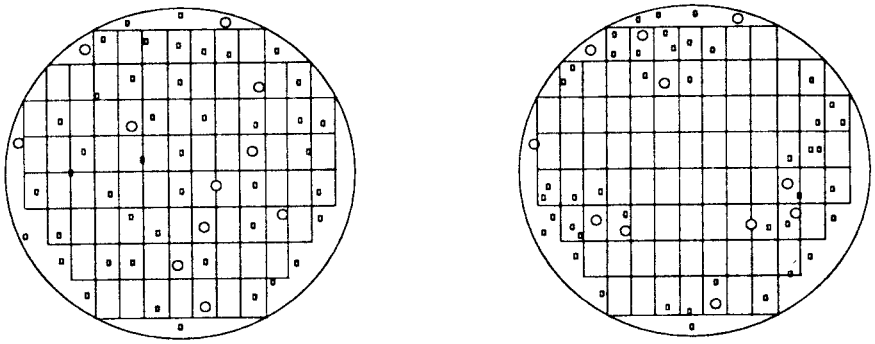
$$f(D) = \frac{1}{\Gamma(\alpha)\beta^\alpha} D^{\alpha-1}e^{-D/\beta}.$$

Then, the probability that one chip contains k defects follows negative binomial distribution

$$\begin{aligned} P_k &= \int_0^\infty \frac{e^{-A_c D}(A_c D)^k}{k!} \frac{D^{\alpha-1}e^{-D/\beta}}{\beta^\alpha \Gamma(\alpha)}\,dD \\ &= \frac{\Gamma(\alpha+k)(A_c\beta)^k}{k!\Gamma(\alpha)(1+A_c\beta)^{\alpha+k}}. \end{aligned}$$

Therefore, the yield model is given by

$$Y = P_0 = \left(1 + \frac{A_c D_0}{\alpha}\right)^{-\alpha}. \tag{5}$$



(a) (b)

Fig. 6. Comparison of two degrees of defect clustering given the same average defect density: (a) low clustering and (b) high clustering.

The clustering factor α determines the degree of clustering of the model. If α is equal to one, then (5) is equivalent to the Seed's yield model of (6). If α goes to ∞, then (5) gives the same result as the Poisson model of (4), implying no clustering. The practical range of α is 0.3 to 5.0. Stapper [72], [73], reports that this model fits well with actual yield data. Stapper [60]–[62] explains the effects of clustering on yield. For the same average defect density, clustering usually gives higher chip yield [63]. Fig. 6 shows configurations of two different degrees of defects clustering. The left one with lower clustering factor has lower yield even though both configurations contain the same number of defects. In Fig. 6, the low clustering situation resembles more to the Poisson model and the high clustering situation more resembles the negative binomial model of smaller α value.

If we assume that defect density follows a normal distribution which is approximated by a triangular distribution (i.e., Simpson distribution), then Murphy's yield model is obtained

$$Y = \left(\frac{1 - e^{-A_c D_0}}{A_c D_0}\right)^2.$$

The predicted yields of this model agree well with actual yields within a tolerance [74].

The assumption that defect density is exponentially distributed gives another yield model which is expressed by

$$Y = \frac{1}{1 + A_c D_0}. \tag{6}$$

Seed's yield model generally gives us higher yields than the actual observations [47]. Price [75] derived the same result by considering the total number of ways indistinguishable defects can be distributed among chips.

If the defect density is uniformly distributed over the interval [0, $2D_0$], then the yield is given by

$$Y = \frac{1 - e^{-2A_c D_0}}{2A_c D_0}.$$

This model predicts yield higher than the observed yield [74].

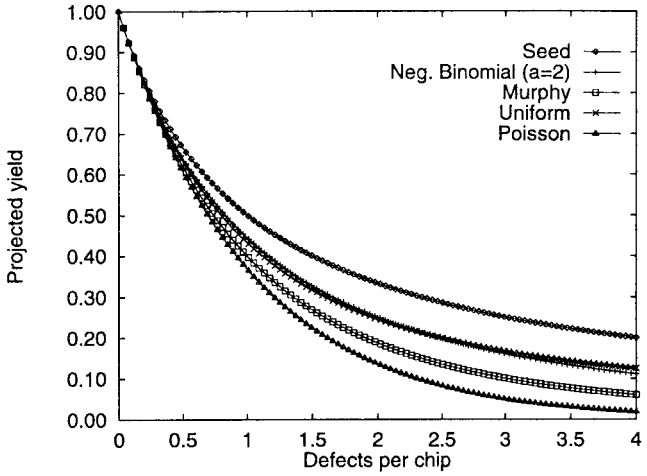


Fig. 7. The comparison of five yield models.

Okabe [76] presents another yield model which is based on Erlang distribution

$$Y = \frac{1}{(1 + A_c D_0/x)^x}$$

where x is the number of mask levels. It is structurally similar to the negative binomial yield model, but the derivation is different. It is reported that this yield model does not agree well with data [73].

Fig. 7 shows a comparison of yield models. As mentioned above, Seed's yield model and the Poisson yield model give the highest and the lowest projected yields, respectively. In Fig. 7, the x axis and y axis refer to μ and Y, respectively.

E. Different Approaches to Yield Modeling

Berglund [77] represents a variable defect size (VDS) yield model including both conventional small defects and much larger parametric or area defects. To do this, (4) can be modified as

$$Y = \exp\left[-\int A_c(x) D(x)\, dx\right]. \tag{7}$$

Assume that the larger defects are circular in shape with diameter x for a die of length L and width W, the total critical area sensitive to such larger defects is [77]

$$A_c(x) = LW + (L+W)x + \pi x^2/4.$$

Let Y_0 be the yield loss due to the defects of small size and Y_p the yield loss due to the defects which are comparable or larger than the die size. Berglund [77] shows that (7) is written in the product form of two exponential terms: $Y = Y_0 \times Y_p$. Therefore, the die-area-independent yield loss term Y_0 can be viewed as the gross yield, and the additional die-area-dependent term Y_p accounts for the added yield loss around the edges of the larger parametric defects. Berglund [77] concludes that by selecting appropriate values for some parameters, the VDS model will satisfactorily match most experimental data of yield versus die area that can also be matched by defect clustering models.

It is generally believed that yield is a function of chip area and that larger chips give lower yields. However, there are some cases in which the yields scatter in a wide range for chips with the same areas, because of the variation in circuit density. Stapper [48] presents a circuit count approach to yield modeling which includes the number of circuits n_j and the average number of random faults μ_j per circuit type j. The negative binomial yield model of this approach is given by

$$Y = Y_0 \left\{ 1 + \sum_j \frac{n_j \mu_j}{\alpha} \right\}^{-\alpha}$$

where Y_0 is the gross yield and α is a cluster factor.

To analyze and compare the yield of products from different semiconductor manufacturing facilities, Cunningham *et al.* [57] present a common yield model. The first step needed is to select the technological and organizational factors influencing the yields of different manufacturing processes. They select 18 candidate factors to build a model, apply a linear regression model to a sample of yield data from 72 die types in separate processes, and conclude that die size, process age, and photo link are significant variables. The resulting absolute yield model with R^2 (coefficient of determination) $= 0.6$ is given by [57]

$$Y = e^Z (1 + e^Z)^{-1}$$
$$Z = 0.33 - 0.80 X_1 + 0.34 \log X_2 + 0.39 X_3$$

where

- X_1 the die size variable refers to the area in cm^2;
- X_2 the process age variable refers to the time span in months between the first and last yield data supplied;
- X_3 the photo link variable, which is 1 if the photolithography system is linked and -1 otherwise.

However, the accuracy of an absolute yield model depends upon detailed information collected.

Michalka *et al.* [59] suggest a yield model to illustrate the effect of nonfatal defects and repair capabilities on yield calculations. Assume a die having both core and support areas where defects randomly occur. The support area yield is defined as the probability that there are no fatal defects in the support area

$$Y_s = \int_0^\infty e^{-DA_s} f(D)\, dD \tag{8}$$

where A_s is the support critical area. The core area yield includes the chance of defects being repaired. To do this, we need one more assumption: that fatal defects can be independently repaired with probability P_{rep}; however, no repair is possible in the support area. Let $Y_c(i)$ be the core yield given that there are i defects in the die. Then, the core area yield is [59]

$$Y_c = \sum_{i=0}^{\infty} \left\{ \int_0^\infty \frac{(DA_{t,c})^i e^{-DA_{t,c}}}{i!} f(D)\, dD \right\} Y_c(i) \tag{9}$$

where $A_{t,c}$ is the core area. From (8) and (9), the die yield is the product form of the support area yield and core area yield

$$Y = Y_s Y_c.$$

The productivity of a wafer is defined as the number of circuits available per wafer after fabrication [47]. All parameters except defect density are invariant after the design is fixed. The defect density is not a design parameter but results from the processes of fabrication. Based on an existing reference product, Ferris-Prabhu [47] suggests a method to predict the productivity of a new product, q quarters after the start of normal production, which is given by

$$P_q(s) = n_s N(s) Y_q(s)$$

where n_s is the number of circuits per square chip of edge s, $N(s)$ is the number of square chips per wafer, and $Y_q(s)$ is the predicted yield for a new product after q quarters.

Dance and Jarvis [78] present a performance-price improvement strategy using yield models and an application of yield models to accelerate learning. Using yield models to accelerate the progress of a learning curve reduces learning cycle time to deliver required manufacturing technology within the time frame set by the competition. They present four major improvement techniques to accelerate learning [78]: fine-grain yield models; short-loop defect monitors; equipment particulate characteristic; and yield confidence intervals. Other yield models used in various companies are well summarized in [48]. All the yield models are used to serve as planning tools. Depending on the applications and product history, specific models can be selected. Fig. 7 provides a guideline for such a comparison.

Table 3
DRAM and Pentium Cost Analysis (in 1994)

Product	4M DRAM	16M DRAM	16M DRAM	Pentium P54C
Feature size	0.6μ	0.5μ	0.35μ	0.6μ
Wafer size	150mm	200mm	200mm	200mm
Tested wafer cost	$600	$1,140	$1,410	$1,500
Die size	54.8mm^2	116.1mm^2	100mm^2	163mm^2
Total dice available / wafer	254	212	253	148
Defect density	0.5/cm^2	1.0/cm^2	0.6/cm^2	1.5/cm^2
Probe yield	80%	35%	58%	15%
Number of good dice	203	74	146	22
Factory cost /die	$4.14	$23.25	$12.63	$153.31
Average selling price / die	$12	$60	$27	$700
Approx. revenue/wafer start	$2,140	$3,120	$3,430	$11,200
Revenue/sq. in. started	$70	$62	$68	$223
Gross margin	65%	61%	53%	78%

Table 4
Wafer Processing Cost (in 1994)

Cost factor	Wafer size			
	100mm	125mm	150mm	200mm
Facility cost	$75M	$175M	$400M	$750M
Raw wafer cost	$11	$20	$30	$100
Depreciation per wafer	$54	$127	$302	$558
Wafer processing cost	$124	$207	$425	$972
Wafer processing yield	90%	95%	98%	98%
Yielded wafer processing cost	$138	$218	$434	$992

IV. Cost Factors

Manufacturing cost is almost 54% of the cost per good wafer produced by U.S. semiconductor manufacturers [79]. In general, manufacturing cost includes direct labor cost, material cost, spare part cost, maintenance cost, production control cost, facility cost, utility cost, and so on. Sometimes, manufacturing cost means wafer processing cost because assembly and final testing may be performed at different sites. Table 3 shows an example of cost analysis for DRAM and pentium chips [56].

A. Wafer Processing Cost

Wafer processing cost consists of direct labor cost (3%), raw wafer cost (7%), direct factory overhead cost (25%), and indirect factory overhead cost (65%) [56]. Table 4 shows wafer processing cost factors for four wafer sizes [56].

B. Wafer Probe Cost

Wafer probe cost adds about 5% additional cost to the yielded wafer processing cost shown in Table 4. Factors that affect the wafer probe cost are test time, number of dice to be tested, probe yield, test equipment costs, number of parallel test sites, and overhead costs [56]. Usually, wafer probe cost is high in the development stage of devices.

Table 5
An Example of Package Cost of DRAM and Pentium (in 1994)

Product	4M DRAM	16M DRAM	16M DRAM	Pentium P54C
Feature size	0.6μ	0.5μ	0.35μ	0.6μ
Package cost / die	$0.30	$0.40	$0.40	$20.00
Assembly yield	98%	95%	97%	98%

Table 6
Final Test Cost and Yield (in 1994)

Product	Final test cost ($)	Final test yield (%)
8-bit MPU	0.10	95
20,000 gate array	0.80	90
1M DRAM	0.30	93
4M DRAM	0.40	90
16M DRAM	0.80	75
4K GaAs SRAM	1.00	80
32-bit MPU (386)	2.50	90
32-bit MPU (P54C)	25.00	75

C. Assembly and Packaging Cost

Assembly and packaging costs are dependent upon the package price, labor cost, assembly yield, equipment cost, and overhead costs. Table 5 shows an example of package cost for DRAM's [56].

D. Final Test Cost

Final test cost depends on the level of testing and the complexity of devices. Some estimated final test costs and the final test yields are shown in Table 6 [79].

V. Fault Coverage and Occurrence

The defect level is defined as the percentage of defective circuits passing all phases of a manufacturing test [80] or the probability that any given IC has untested defects [81]. Thus, the defect level represents the proportion of a product which may fail because of extrinsic failure (or infant mortality) [86]. Let D_L be the defect level of the IC. Then it is given by [82]

$$D_L = 1 - Y^{1-T} \tag{10}$$

where Y and T are yield and fault coverage, respectively. The fault coverage is defined as the ratio of the number of detected faults and the number of faults assumed in the fault list, a measure of how many defects within the IC are tested. One minus defect level $(1 - D_L)$, called the reliable fraction, quality level, or sometimes reliability, represents the probability that an IC has no reliability defects.

The basic assumption of (10) is that all faults have equal probability of occurrence, which implies no clustering. That is, the faults are uniformly distributed. Corsi [80] extends this for nonequiprobable faults, using a generalized weighted fault coverage T

$$T = \frac{\sum_{j=1}^{m} A_{cj} D_{0j}}{\sum_{i=1}^{n} A_{ci} D_{0i}}$$

where m and n are the number of faults tested and the total number of faults assumed, respectively. This relationship is useful to estimate the defect level (or reliable fraction) after a test or to determine how much testing is necessary to obtain an assigned defect level (or reliable fraction).

Seth and Agrawal [83] combine fault coverage with fault occurrence probability in order to find a relationship between fault coverage and product quality. The fault occurrence probability is defined as the probability with which the fault will occur on a chip. Their attempt is to find a fault occurrence probability for individual faults instead of a distribution for them. They call the product of these two probabilities the absolute failure probability of a chip.

Let N be the total number of test vectors applied. After application of N test vectors, the true yield is given by [83]

$$Y = 1 - \frac{1}{c}\sum_{i=1}^{N} c_i - \frac{2N+1}{N}\frac{1}{c}\sum_{i=1}^{N} c_i \frac{i(i+1)}{(N+i)(N+i+1)}$$

where c is the total number of chips tested and c_i is the number of chips that fail exactly at test vector number i, and the estimated yield is also given by

$$Y_n = Y + \left(1 - Y - \frac{1}{c}\sum_{i=1}^{N} c_i\right)\frac{N+1}{N+n+1} + \frac{1}{c}\sum_{i=1}^{N} c_i \frac{i(i+1)}{(n+i)(n+i+1)}.$$

Therefore, the defect level is presented as

$$D_L = \frac{Y_n - Y}{Y_n}.$$

Since (10) does not provide good results when faults are correlated, Maxwell and Aitken [84] present another description for defect level

$$D_L = \frac{(1-T)(1-Y)\exp[-(n_0-1)T]}{Y + (1-T)(1-Y)\exp[-(n_0-1)T]} \tag{11}$$

where n_0 is the average number of faults on a die. Willing and Helland [81] present a mathematical model to develop fault coverage guidelines for complex electronic systems. Their model extends (11) based on probabilistic relationships between yield, fault coverage, and defect level and finds reliability to be a function of fault coverage and yield.

VI. Yield–Reliability Relation Model

A. Yield–Reliability Relationship

In the past, most attempts to assure high IC reliability employed product testing, life testing, or accelerated stress tests of the entire circuit. Because product testing is getting more expensive, more time consuming, and less capable of effectively identifying the causes of parametric and functional failures of IC's, the development of new technologies is needed. These new technologies will make it possible to remove wearout failures during the operational life.

The degree of manufacturing success is measured by yield, which is defined as the average ratio of devices on a wafer that pass the tests. Since yield is a statistical parameter and implies a probability function, yield functions are multiplied in order to attain the total yield. The total wafer yield is a measure of good chips per wafer normalized by the number of chip sites per wafer. Generally, since yields can be measured in various ways, the overall yield is calculated by the product of elements of yield such as the line yield, the WP yield, the assembly yield, and the FT yield.

Parameters that affect yield are defects, and the number of defects produced during the manufacturing process can be effectively controlled by introducing test points at crucial times rather than throughout the assembly line [26]. This not only improves the reliability of the outgoing product but also significantly enhances the yield of the manufacturing process, thus increasing the quality of the overall system. Test points are effective only at critical areas, and their random distribution in the process was observed not to yield the desired results of high quality and minimal defects density. There is another way to control the yield. Since IC device yields are not only a function of chip area but also a function of circuit design and layout, by determining the probabilities of failure and critical areas for different defect types, it is possible to control and manage the yield of IC's [48].

Schroen [85] suggests a new system for studying reliability by utilizing test structures sensitive to specific failure mechanisms. By stressing these structures, more accurate information about the reliability of a circuit can be obtained in a shorter time than by the use of traditional methods. Schroen also regarded this method as a means of reducing dependence on costly and time consuming burn-in testing.

As was previously stated, yield and reliability are two important factors affecting the profitability of semiconductor manufacturing. However, the correlation between them has not been clearly identified.

B. Yield–Reliability Relation Model

Some reliability prediction models describe the defect level or the reliable fraction of products as a function of yield. Most models are based on the relationship of the device degradation and the long-term reliability. These models can only be used to estimate the defect level after a final test or to interrelate failures with the ultimate reliability [80], [82], [86], [87]. If one wants to identify the effects of stresses or conditions causing the infant mortality failures, it is necessary to relate the reliability model with defect reliability physics and to describe that as the function of yield. However, only two relation models have been reported until now. Huston and Clarke's model [86] uses the critical area for the yield and the reliability to model the relationship. In their model, for the given yield Y the reliability R is given by

$$R = Y^{A_r/A_c} \tag{12}$$

where A_r and A_c are reliability and yield critical areas, respectively. In order to use the model of (12), it is necessary to calculate the reliability critical area based on defect reliability physics. Using a least square regression, they provide 0.3 as an estimate of A_r/A_c.

Kuper *et al.* [87] and Van der Pol *et al.* [54] use the same model for the yield-reliability relation and present the experimental data to show the correlation. Their relation model is given by

$$R = (Y/M)^{\alpha} \tag{13}$$

where M ($M > 0.9$) is a parameter for clustering effects and edge exclusions and α is the ratio between the density of reliability defects D_r and the density of yield defects D_y ($\alpha = D_r/D_y$). One assumption of (13) is that the density of reliability defects is a fraction of the density of yield defects. They suggest using the same α for similar products in a given technology and apply (13) to five different IC's in order to verify the existence of a strong relationship between yield and failure occurring in the early life time.

In general, reliability is the ability or capability of the product to operate properly without failure, which is defined as the cumulative probability function at time t for a given time under the operating conditions.

Note that the models in (12) and (13) are not related to time t. Thus, the reliability terms used in the two relation models are not defined at a specific time t, but the average ratio of devices working properly in the early life time.

Kuo *et al.* [5] and Kim *et al.* [88] show a different relation model which is defined at time t based on the POF concept. Let $R(t)$ and $c(t)$ be the reliability at time t and the time-dependent constant, respectively. Then, since reliability is presented as a function of yield and time

$$R(t) = g(Y, t) = Y^{c(t)}.$$

Their model concentrates on the gate oxide reliability and provides a possible way to interrelate yield and burn-in. Since reliability and yield are strongly related, the decision to burn-in or not burn-in can be made by observing yield. This is another way to avoid time-consuming burn-in.

VII. Conclusions

As was pointed out in a recent manuscript [5], since most microelectronics components have an infant mortality period of about one year under ordinary operating conditions, the reliability problem in the infant mortality period becomes extremely important. One purpose for applying burn-in to products is to guarantee high reliability of the end products. In addition, we take lessons from early failed products for which design modifications can be made for the future products. We update the design and manufacturing processes in order to enhance both the manufacturing yield and the product reliability. If and when this purpose is achieved, screening products becomes unnecessary. However, microelectronics products using new technology come to the market place almost daily; therefore, information obtained from screening is valuable for a limited number of manufacturing processing updates using the existing technology. Beyond that, once the existing technology becomes obsolete, the products using new technology need to be evaluated to meet the quality and reliability standards again. Information obtained from burn-in on current products can serve as prior knowledge for burn-in on the design of products due to new technology. Unless we can forecast the exact causes for design and manufacturing flaws of future products, stress burn-in will still serve the screening purpose. In particular, IC's for high consequence of failures need to subject themselves to the full screening procedure before they are assembled into a dedicated system.

A process that produces a product of high yield implies that high ratio of the number of usable items at the completion of the process to the number of potentially usable items at the beginning of production. The yield of the specific process that the manufacturing engineers of semiconductors often refer to is presented as Y_0 in Fig. 8. Assume the time at completion of this process is zero, as indicated in Fig. 8. Beyond time zero, the failure-free operation of a device is known as reliability. However, a lower Y_0 will affect reliability which is a function of time. The trend toward smaller Y_0 as a function of time may be caused by

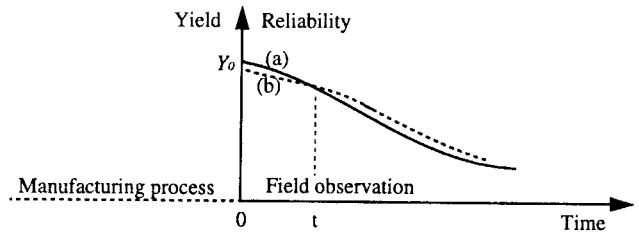


Fig. 8. The decreasing function of yield with respect to time being the reliability problem.

intrinsic, extrinsic, or wearout failures. Therefore, the yield and reliability of microelectronics manufacturing products are highly related, but high manufacturing yield due to one manufacturing process does not necessarily imply high reliability of the products from that manufacturing process in the field. For example, in Fig. 8, curve (b) shows products that have a lower yield than products of curve (a) at the completion of the manufacturing process but curve (a) exhibits high reliability after the field observation time t. The functional relationship between reliability, which is time dependent, and yield, which is quality dependent, deserves special attention in future studies. According to Tang [89], the probability of a defective IC depends on the process defect density and die area and does not depend on the specific IC. Therefore, we can estimate the failure rate for a new IC using data from an IC with similar technology. However, a larger die generally has a higher burn-in failure rate than a smaller die due to more opportunities for defects. Also, an IC with small geometry and a complex wafer process is more prone to defects. In addition, cost, wafer size, and burn-in effectiveness will have direct impact on manufacturing yield. Tradeoffs among these factors are essential in order to guarantee high reliability in semiconductor products.

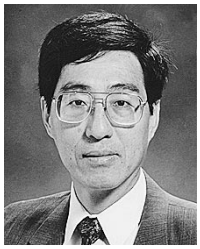
ACKNOWLEDGMENT

The authors would like to acknowledge suggestions made by the reviewers. Figs. 1 and 5 are reprinted from [5] with permission from Kluwer Academic Publishers.

REFERENCES

[1] A. G. Sabnis, *VLSI Electronics Microstructure Science V.22 VLSI Reliability.* San Diego, CA: Academic, 1990.
[2] D. E. Swanson, "Forty years and looking forward," *Semicond. Int.*, vol. 11, no. 1, p. 13, Jan. 1988.
[3] E. R. Hnatek, *Integrated Circuit Quality and Reliability*, 2nd ed. New York: Marcel Dekker, 1995.
[4] A. Christou, *Integrating Reliability into Microelectronics Manufacturing.* Chichester, U.K.: Wiley, 1994.
[5] W. Kuo, W. T. K. Chien, and T. Kim, *Reliability, Yield, and Stress Burn-in.* Norwell, MA: Kluwer, 1998.
[6] A. Amerasekera and D. S. Campbell, *Failure Mechanisms in Semiconducter Devices.* New York: Wiley, 1987.
[7] M. Campbell, "Monitored burn-in improves VLSI IC reliability," *Comput. Design*, vol. 24, pp. 143–146, Apr. 1985.
[8] D. L. Denton and D. M. Blythe, "The impact of burn-in on IC reliability," *J. Environmental Sci.*, vol. 29, no. 1, pp. 19–23, Jan./Feb. 1986.
[9] W. Kuo and Y. Kuo, "Facing the headaches of early failures: A state-of-the-art review of burn-in decisions," *Proc. IEEE*, vol. 71, pp. 1257–1266, Nov. 1983.
[10] D. Chi and W. Kuo, "Burn-in optimization under reliability & capacity restrictions," *IEEE Trans. Reliability*, vol. 38, pp. 193–198, June 1989.
[11] K. Chou and K. Tang, "Burn-in time and estimation of change-point with Weibull-exponential mixture distribution," *Decision Sci.*, vol. 23, no. 4, pp. 973–990, July/Aug. 1992.
[12] D. G. Nguyen and D. N. P. Murthy, "Optimal burn-in time to minimize cost for products sold under warranty," *IIE Trans.*, vol. 14, no. 3, pp. 167–174, 1982.
[13] The Naval Publications and Forms Center, "Definitions of item levels, item exchangeability, models and related terms," The Naval Publications and Forms Center, Philadelphia, PA, MIL-STD-280A, 1969.
[14] W. T. K. Chien and W. Kuo, "A nonparametric approach to estimate system burn-in time," *IEEE Trans. Semiconduct. Manufact.*, vol. 9, pp. 461–466, Aug. 1996.
[15] C. W. Whitbeck and L. M. Leemis, "Component vs system burn-in techniques for electronic equipment," *IEEE Trans. Reliability*, vol. 38, pp. 206–209, June 1989.
[16] W. Kuo, "Reliability enhancement through optimal burn-in," *IEEE Trans. Reliability*, vol. R-33, pp. 145–156, June 1984.
[17] W. T. K. Chien and W. Kuo, "Modeling and maximizing burn-in effectiveness," *IEEE Trans. Reliability*, vol. 44, pp. 19–25, Mar. 1995.
[18] L. M. Leemis and M. Beneke, "Burn-in models and methods: A review," *IIE Trans.*, vol. 22, no. 2, pp. 172–180, June 1990.
[19] W. Kuo, "Incompatibility in evaluating large-scale systems reliability," *IEEE Trans. Reliability*, vol. 43, pp. 659–660, Dec. 1994.
[20] M. Haim and Z. Porat, "Bayes reliability modeling of a multi-state consecutive k-out-of-n: F system," in *Proc. Annu. Reliability and Maintainability Symp.*, 1991, pp. 582–586.
[21] W. T. K. Chien and W. Kuo, "A nonparametric Bayes approach to decide system burn-in time," *Naval Res. Logistics*, vol. 44, no. 7, pp. 655–671, 1997.
[22] T. A. Mazzuchi and N. D. Singpurwalla, "A Bayesian approach to inference for monotone failure rates," *Statistics Probability Lett.*, vol. 3, no. 3, pp. 135–141, June 1985.
[23] E. Takeda, K. Ikuzaki, H. Katto, Y. Ohji, K. Hinode, A. Hamada, T. Sakuta, T. Funabiki, and T. Sasaki, "VLSI reliability challenges: From device physics to wafer scale systems," *Proc. IEEE*, vol. 81, pp. 653–674, May 1993.
[24] D. L. Crook, "Evolution of VLSI reliability engineering," in *Proc. Int. Reliability Physics Symp.*, 1990, pp. 2–11.
[25] C. Hu, "Future CMOS scaling and reliability," *Proc. IEEE*, vol. 81, pp. 682–689, May 1993.
[26] J. A. Shideler, T. Turner, J. Reedholm, and C. Messck, "A systematic approach to wafer level reliability," *Solid State Technol.*, vol. 38, no. 3, p. 47, Mar. 1995.
[27] T. A. Dellin, W. M. Miller, D. G. Pierce, and E. S. Snyder, "Wafer level reliability," in *Proc. SPIE Microelectronics Manufacturing and Reliability*, vol. 1802, Jan. 1993, pp. 144–154.
[28] A. Papp, F. Bieringer, D. Koch, H. Kammer, A. Kohlhase, A. Lill, A. Preussger, A. Schlemm, and M. Schneegans, "Implementation of a WLR—Program into a production line," in *1995 IRW Final Rep.*, 1996, pp. 49–54.
[29] S. Garrard, "Production implementation of a practical WLR program," in *1994 IRW Final Rep.*, 1995, pp. 20–29.
[30] T. E. Kopely, K. Young, R. Rakkhit, S. Chan, and P. Marcoux, "Wafer level hot-carrier measurements for building-in reliability during process development," *1994 IRW Final Rep.*, 1995, pp. 57–59.
[31] L. N. Lie and A. K. Kapoor, "Wafer level reliability procedures to monitor gate oxide quality using V ramp and J ramp test methodology," in *1995 IRW Final Rep.*, 1996, pp. 113–121.
[32] O. D. Trapp, Ed., *Proc. 1991 Int. Wafer Level Reliability Workshop*, Lake Tahoe, CA., Oct. 13–16, 1991.
[33] T. E. Turner, "Wafer level reliability: Process control for reliability," *Microelectron. Reliability*, vol. 36, nos. 11/12, pp. 1839–1846, 1996.
[34] J. M. Soden and R. E. Anderson, "IC failure analysis: Techniques and tools for quality and reliability improvement," *Proc. IEEE*, vol. 81, pp. 703–715, May 1993.
[35] B. Schlund, C. Messick, J. Suehle, and P. Chaparala, "A new physics-based model for time-dependent-dielectric-breakdown," in *Integrated Reliability Workshop, Final Rep. Int.*, 1995, pp. 72–80.
[36] T. Kim and W. Kuo, "Optimal burn-in decision making," submitted for publication.

[37] E. R. Hnatek, "A realistic view of VLSI burn-in," *Evaluation Eng.*, vol. 28, no. 2, pp. 80–89, 1989.
[38] H. E. Hamilton, "An overview—VLSI burn-in considerations," *Evaluation Eng.*, vol. 31, no. 2, pp. 16–20, Feb. 1992.
[39] D. Romanchik, "Why burn-in IC's?," *Test & Measurement World*, vol. 12, pp. 85–88, Oct. 1992.
[40] ——, "Burn-in: Still a hot topic," *Test & Measurement World*, vol. 12, pp. 51–54, Jan. 1992.
[41] D. Gralian, "Next generation burn-in development," *IEEE Trans. Comp., Packag, Manufact. Technol. B*, vol. 17, pp. 190–196, May 1994.
[42] B. Vasquez and S. Lindsey, "The promise of known-good-die technologies," in *MCM'94 Proc.*, 1994, pp. 1–6.
[43] A. Martin, J. Suehle, P. Chaparala, P. O'Sullivan, A. Mathewson, and C. Messick, "Assessing MOS gate oxide reliability on wafer level with ramped/constant voltage and current stress," in *Integrated Reliability Workshop, Final Rep. Int.*, 1995, pp. 81–91.
[44] A. D. Singh, "On wafer burn-in strategies for MCM die," in *Proc. Int. Conf. Exhibition Multichip Modules*, Apr. 1994, pp. 255–260.
[45] D. B. Tuckerman, B. Jarvis, L. Chang-Ming, P. Patel, and M. Hunt, "A cost-effective wafer-level burn-in technology," in *Proc. Int. Conf. Exhibition Multichip Modules*, Apr. 1994, pp. 34–40.
[46] W. G. Flynn and L. Gilg, "A pragmatic look at wafer-level burn-in: The wafer-level known-good die consortium," in *IECEM'96 Proc.*, 1996, pp. 287–292.
[47] A. V. Ferris-Prabhu, *Introduction to Semiconductor Device Yield Modeling*. Boston, MA: Artech House, 1992.
[48] 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.
[49] D. F. Frost and K. F. Poole, "A method for predicting VLSI-device reliability using series models for failure mechanisms," *IEEE Trans. Reliability*, vol. R-36, pp. 234–242, 1987.
[50] J. L. Stevenson and J. A. Nachlas, "Microelectronics reliability predictions derived from components defect densities," in *Proc. Annu. Reliability and Maintainability Symp.*, 1990, pp. 366–371.
[51] F. Jensen, "Yield, quality and reliability—A natural correlation?," in *Reliability'91*, R. H. Matthews, Ed. London: Elsevier Applied Science, 1991, pp. 739–750.
[52] E. M. J. G. Bruls, "Reliability aspects of defect analysis," *IEEE/ETC*, pp. 17–26, 1993.
[53] J. G. Prendergast, "Reliability and quality correlation for a particular failure mechanism," in *Proc. Int. Reliability Physics Symp.*, 1993, pp. 87–93.
[54] J. Vander Pol, F. Kuper, and E. Ooms, "Relation between yield and reliability of integrated circuits and application to failure rate assessment and reduction in the one digit fit and ppm reliability era," *Microelectron. Reliability*, vol. 36, nos. 11/12, pp. 1603–1610, 1996.
[55] B. El-Kareh, A. Ghatalia, and A. V. S. Satya, "Yield management in microelectronic manufacturing," in *Proc. 45th Electronic Components Conf.*, 1995, pp. 58–63.
[56] Integrated Circuit Engineering Corp., *Cost Effective IC Manufacturing 1995*. Scottsdale, AZ: ICE, 1995.
[57] S. P. Cunningham, C. J. Spanos, and K. Voros, "Semiconductor yield improvement: Results, and best practices," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, pp. 103–109, May 1995.
[58] C. H. Stapper, F. M. Armstrong, and K. Saji, "Integrated circuit yield statistics," *Proc. IEEE*, vol. 71, pp. 453–470, Apr. 1983.
[59] T. L. Michalka, R. C. Varshney, and J. D. Meindl, "A discussion of yield modeling with defect clustering, circuit repair, and circuit redundancy," *IEEE Trans. Semiconduct. Manufact.*, vol. 3, pp. 116–127, Aug. 1990.
[60] C. H. Stapper, "The effects of wafer to wafer defect density variations on integrated circuit defect and fault distributions," *IBM J. Res. Develop.*, vol. 29, no. 1, pp. 87–97, Jan. 1985.
[61] ——, "On yield, fault distributions and clustering of particles," *IBM J. Res. Develop.*, vol. 30, no. 3, pp. 326–338, May 1986.
[62] ——, "Large-area fault clusters and fault tolerance in VLSI circuits: A review," *IBM J. Res. Develop.*, vol. 33, no. 2, pp. 162–173, Mar. 1989.
[63] A. Ghatalia and B. El-Kareh, "Yield management in microelectronic manufacturing," Short Course Notes of The National Alliance for Photonics Education in Manufacturing, Austin, TX, Oct. 1996.
[64] 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, pp. 878–880, Aug. 1985.
[65] C. H. Stapper, "Modeling of integrated circuit defect sensitivities," *IBM J. Res. Develop.*, vol. 27, no. 6, pp. 549–557, Nov. 1983.
[66] ——, "Modeling of defects in integrated circuit photolithographic patterns," *IBM J. Res. Develop.*, vol. 29, no. 1, pp. 461–475, Jan. 1985.
[67] W. Maly, "Modeling of lithography related yield loss for CAD of VLSI circuits," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, pp. 166–177, July 1985.
[68] C. Kooperberg, "Circuit layout and yield," *IEEE J. Solid-State Circuits*, vol. 23, no. 4, pp. 887–892, Aug. 1988.
[69] Z. Stamenkovic and N. Stojadinovic, "New defect size distribution function for estimation of chip critical area in integrated circuit yield models," *Electron. Lett.*, vol. 28, no. 6, pp. 528–530, 1992.
[70] T. J. Wallmark, "Design considerations for integrated electron devices," *Proc. IRE*, vol. 48, pp. 293–300, Mar. 1960.
[71] A. V. Ferris-Prabhu, "Models for defects and yield," in *Defect and Fault Tolerance in VLSI Systems*, I. Koren, Ed. Booknews, 1989, pp. 33–46.
[72] C. H. Stapper, "Defect density distribution for LSI yield calculations," *IEEE Trans. Electron Devices*, vol. ED-20, pp. 655–657, July 1973.
[73] ——, "Fact and fiction in yield modeling," *Microelectron. J.*, vol. 20, nos. 1/2, pp. 129–151, 1989.
[74] B. T. Murphy, "Cost-size optima of monolithic integrated circuit," *Proc. IEEE*, vol. 52, pp. 1537–1545, Dec. 1964.
[75] J. E. Price, "A new look at yield of integrated circuits," *Proc. IEEE*, vol. 58, pp. 1290–1291, Aug. 1970.
[76] T. Okabe, M. Nagata, and S. Shimada, "Analysis on yield of integrated circuits and a new expression for the yield," *Elect. Eng. Japan*, vol. 92, no. 6, pp. 135–141, Dec. 1972.
[77] C. N. Berglund, "A unified yield model incorporating both defect and parametric effects," *IEEE Trans. Semiconduct. Manufact.*, vol. 9, pp. 447–454, Aug. 1996.
[78] D. Dance and R. Jarvis, "Using yield models to accelerate learning curve progress," *IEEE Trans. Semiconduct. Manufact.*, vol. 5, pp. 41–45, Feb. 1992.
[79] Semiconductor Industry Association, *1978–1993 Industry Data Book*, 1994.
[80] F. Corsi and S. Martino, "Defect level as a function of fault coverage and yield," in *Proc. European Test Conf.*, 1993, pp. 507–508.
[81] W. Willing and A. Helland, "Establishing ASIC fault-coverage guidelines for high-reliability systems," in *Proc. Annu. Reliability and Maintainability Symp.*, 1998, pp. 378–382.
[82] T. W. Williams and N. C. Brown, "Defect level as a function of fault coverage," *IEEE Trans. Comput.*, vol. C-30, pp. 987–988, Dec. 1981.
[83] S. C. Seth and V. D. Agrawal, "On the probability of fault occurrence," in *Defect and Fault Tolerance in VLSI Systems*, I. Koren, Ed. New York: Plenum, 1989, pp. 47–52.
[84] P. Maxwell and R. Aitken, "Test sets and reject rates: All fault coverages are not created equal," *IEEE Design Test Comput.*, vol. 10, pp. 42–51, Mar. 1993.
[85] W. H. Schroen, "Process testing for reliability control," in *Proc. Int. Reliability Physics Symp.*, 1978, pp. 81–87.
[86] H. H. Huston and C. P. Clarke, "Reliability defect detection and screening during processing—Theory and implementation," in *Proc. Int. Reliability Physics Symp.*, 1992, pp. 268–275.
[87] F. Kuper, J. van der Pol, E. Ooms, T. Johnson, R. Wijburg, W. Koster, and D. Johnston, "Relation between yield and reliability of integrated circuits: Experimental results and application to continuous early failure rate reduction programs," in *Proc. Int. Reliability Physics Symp.*, 1996, pp. 17–21.
[88] T. Kim, W. Kuo, and W. T. K. Chien, "A relation model of yield and reliability for gate oxide failures," in *Proc. 1998 Annu. Reliability and Maintainability Symp.*, Anaheim, CA, Jan. 19–22, 1998, pp. 428–433.
[89] S. Tang, "New burn-in methodology based on IC attributes, family IC burn-in data, and failure mechanism analysis," in *Proc. Annu. Reliability and Maintainability Symp.*, 1996, pp. 185–190.



Way Kuo (Fellow, IEEE) received the B.S. degree in nuclear engineering from National Tsing-Hua University, Taiwan, in 1972 and the M.S. and Ph.D. degrees in industrial engineering and the M.S. degree in statistics from Kansas State University, Manhattan, KS, in 1977, 1979, and 1980, respectively.

He is a Professor and Head of the Department of Industrial Engineering at Texas A&M University, College Station. Previously, he was a Reliability Specialist at Bell Laboratories, Visiting Scientist at the Oak Ridge National Laboratory, Senior NRC Associate at the Naval Postgraduate School, and Senior Engineer of Ames Lab. In 1991, he was the Senior Fulbright Scholar in Reliability and Quality in Lisbon, Portugal, and Glasgow, U.K. He has coauthored the texts *Optimization of Systems Reliability,* (New York: Dekker, 1985), *Software Measurement: A Visualization Toolkit for Project Control & Process Improvement,* (Englewood Cliffs, NJ: Prentice Hall, 1998), *Reliability, Yield, and Stress Burn-In,* (Norwell, MA: Kluwer, 1998), and *Fundamentals and Applications of Reliability Optimization* (Cambridge, U.K.: Cambridge Univ. Press, 1999).

Dr. Kuo was elected as an Academician of the International Academy for Quality in 1996. He is a Halliburton Professor of Engineering at Texas A&M and a Fellow of the Institute of Industrial Engineering (IIE) and the American Society for Quality (ASQ). He is the Editor of IEEE TRANSACTIONS ON RELIABILITY.

Taeho Kim received the B.S. and M.S. degrees in industrial engineering from Seoul National University, Seoul, Korea, in 1983 and 1985, respectively, and the Ph.D. degree in industrial engineering from Texas A&M University, College Station, in 1998.

In 1986, he joined Korea Telecom, Taejon, Korea, where he is currently a Senior Researcher. His fields of interest include semiconductor reliability and yield, stress burn-in and its optimization, software reliability, and telecommunications network planning and service quality.

Dr. Kim is a member of Phi Kappa Phi.