J. Webster (ed.), *Wiley Encyclopedia of Electrical and Electronics Engineering*

SEMICONDUCTOR DEVICE MANUFACTURE YIELD AND RELIABILITY MODELING

The historical breakthrough invention of the first integrated circuit (*IC*) was made by Jack 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 ICs 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). Table 1 (2) 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*). In the first years of the new century, the IC industry will enter the super-large-scale integration (*SLSI*) era with over 10^9 transistors for 4-Gbit (and over) DRAMs.

Table 2 (3) summarizes the projection of the world market growth according to the World Semiconductor Trade Statistics (*WSTS*) press release. It is expected, during the following years into 2004, that the market will grow at a compound rate of 15.6%. The most critical point in this evolving market is for failure analysis and reduction to keep pace with IC technology development, since the reliability and yield will continue to be problems for the IC industry.

Behavior of Failures

Systems and materials begin to wear out during use, and various mechanisms can contribute to failures. Therefore, failures need to be confined 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 other factors in the environment can contribute to this process.

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 ICs). The operating period that follows the infant mortality period has a lower failure rate and is called the *useful life* period. In this period (about 40 years for ICs), the failure rate tends to remain constant until the beginning of the next phase, called the *aging* period. Failures during the aging period are typically due to cumulative damage. Typically the failure rate 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 faulty 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 not actually mature enough to be released. Note that infant

Table 1. The Progressive Trend of IC Technology

Integration level	Decade	Number of transistors	DRAM (bits) integration
SSI	1950s	$<10^2$	
MSI	1960s	10^2–10^3	
LSI	1970s	10^3–10^5	4 K, 16 K, 64 K
VLSI	1980s	10^5–10^7	256 K, 1 M, 4 M
ULSI	1990s	10^7–10^9	16 M, 64 M, 256 M
SLSI	2000s	$>10^9$	1 G, 4 G, and above

Table 2. Industry Sales Expectations for IC Devices

	Billions of Dollars			Percentage Growth		
Device Type	2000	2001	2002	00/99	01/00	02/01
Discretes	16.9	15.3	16.8	29.5	−9.8	9.8
Optoelectronics	9.8	9.7	10.8	69.7	−1.1	11.1
Bipolar digital	1.1	0.6	0.5	7.9	−44.3	−19.9
Analog	30.5	27.7	31.4	38.2	−9.3	13.3
MOS micro	61.5	52.7	59.3	19.0	−14.3	12.5
MOS logic	34.6	30.9	34.7	49.3	−10.6	12.3
MOS Memory	49.2	38.9	46.8	52.5	−20.9	20.2
Total	204.4	176.8	201.4	36.8	−13.5	13.9

mortality is defined for the whole lot rather than 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 failure are classified into three main areas (4, 5): electrical 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 in the semiconductor die itself are called *intrinsic*. Intrinsic failure mechanisms tend to be the result of the wafer fabrication, which is the front end of the manufacturing process. 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 frequent, because it is impossible to screen most such 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 during semiconductor manufacture. Thus, extrinsic conditions affecting the reliability of components vary according to the packaging and interconnection processes. As technologies mature, intrinsic failures are reduced, thereby making extrinsic failures all the more important for device reliability.

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 microelectronics failures are reportedly due to temperature; vibration is the second highest factor, accounting for 27%; moisture accounts for 19%; sand and dust, 6%, salt, 4%, altitude, 2%, and shock, 2%. Thus, temperature is the most critical factor for component failure; this is especially true for semiconductors (4). *Burn-in*, a screening technique performed by applying high temperature and voltage early in the product life cycle to remove latent defects, is found to be useful for highly integrated circuit systems (6,7,8). 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 (9,10,11). Regarding optimal burn-in decision, see Ref. 12.

Because the infant mortality of semiconductor products is high in failure rate and long in mortality period, burn-in at the factory has been widely practiced. According to Kuo and Kuo (8), the key questions for effective burn-in are:

(1) How much should infant mortality be reduced by burn-in?
(2) Under what environmental conditions should burn-in be performed?
(3) Should burn-in be accomplished at the system, subsystem, or component level? What are the strategies to perform burn-in?
(4) Who should be in charge of burn-in—the vender, the buyer, or a third party?
(5) What should be the expected life of device after burn-in? How does it differ from the expected life without burn-in?
(6) Is burn-in always necessary and economic?
(7) What are the savings from burn-in?
(8) Are there any side effects of burn-in?
(9) How will the industry benefit from burn-in data?
(10) What laws of physics should be considered in conducting burn-in?

As described 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 on three levels (component, subsystem, and system) and on two levels (component and system), respectively. Extremely high system reliability can be achieved by burning in at all levels; in that case the component-level burn-in is generally performed by the vendor. For example, a 4-Mbit DRAM used in a personal computer (*PC*) can be viewed as a component. Sixteen 4-Mbit DRAMs are assembled on a printed circuit board (*PCB*) called a *SIMM* (single in-line memory module) 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, SIMMs are put on the motherboards for system-level tests; one frequently used test is to continual open and close many windows and repeatedly execute selected programs or software to verify that the systems (PCs) 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, comprises 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 approach should be incorporated 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 ICs, (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 best-known 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 as samples. Presently, almost all burn-ins are done at the package level, and the sample is called the *DUT* (device under test). Chien and Kuo (21) use DUT in that sense. 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.

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 factory requires a high initial investment and has 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 replaced by new methods in order to better understand and control the input variables at each point in the manufacturing process (23). This requirement leads to the development of built-in reliability (*BIR*), wafer-level reliability (WLR), qualified manufacturing line (QML), and physics of failure (POF) approaches (25,26,27).

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 (26) defines BIR as a methodology or philosophy for manufacturing highly reliable ICs, not by measuring the output at the end of production, but by controlling input variables that affect 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 screening tests, the resulting reliable products will lead to low overall costs. Generally, the BIR approach is effective only beyond a certain crossover point, when reliability improvement offsets the large testing costs.

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 ***FITs*** (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 (28). 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 already some examples of WLR implementation in a production line or testing methods (28,29,30,31,32,33).

According to Turner (34), 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) that are designed to address each specific reliability failure mechanism. However, Crook (24) and Turner (34) point out limitations of the WLR fast test. Since it 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, according to Turner (34), it can only be applied with a full understanding of the limitations of the stresses. Another disadvantage is that at higher stress levels, the failure mode may be physically different from what would occur under normal use conditions (28).

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 QML (26). 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 (35). As in BIR, understanding failure mechanisms and performing failure analysis are critical in implementing the QML concept. Therefore, the QML approach places a heavy emphasis on documentation. QML is another response to 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 (26).

The concept of POF has been widely used in engineering fields, where the opportunity for testing is restricted by 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, can no longer provide timely feedback, 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. (36) develop the POF model to deal with time-dependent dielectric breakdown (*TDDB*).

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 (25). 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.

Burn-in Concepts. A burn-in test that subjects devices to higher than usual levels of stress such as voltage or temperature is a technique used to speed up the deterioration of materials or electrical components so that analysts can collect information more promptly (4). The test results have to be adjusted according to some time transformation models to provide predictions of 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.

Exposure to elevated temperature is one of the most-used physical mechanisms for failure deterioration. If subscripts 1 and 2 refer to normal conditions and accelerated conditions, respectively, and η is the time

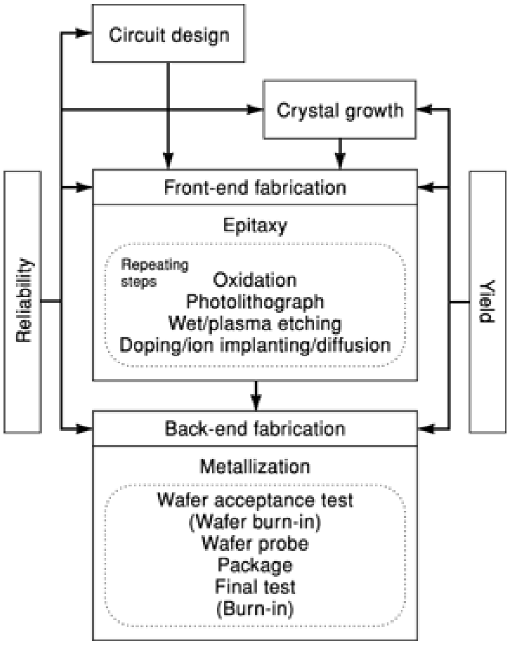

Fig. 1. Influence of reliability and yield on the IC manufacturing process.

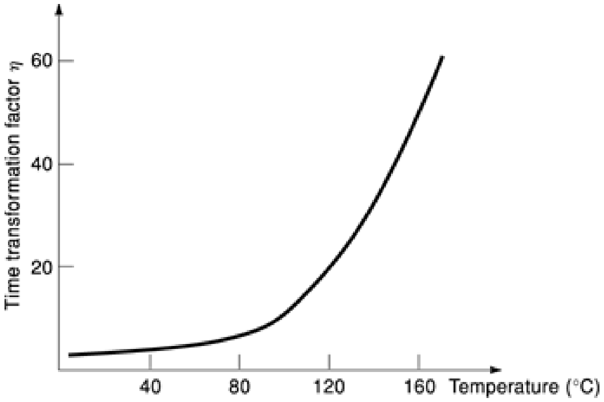

Fig. 2. The time transformation factor for different temperatures at activation energy 0.4 eV.

transformation factor, then the relationship between the time to failure under normal conditions, t_1, and accelerated conditions, t_2, can be expresses 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 temperatures are given in Fig. 2. Relationship similar to Eq. (1) for the other stress factors are also described in Kuo et al. (4).

Various Test Steps. The WAT is an electrical test and is done at the wafer level right before the WP. The WP is often called chip probing (*CP*) or wafer sorting, and its objective is to identify good bare dice on the wafer. Packaged dice that have successfully passed the burn-in test will be ready for the FT. During the FT, the full functionality of the product is checked. Usually, the test items in the FT are similar to but more complicated than the ones in the 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 that 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 (4), although it can occasionally reduce manufacturing yields. The accelerating 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 cost–benefit analysis of burn-in is given in Refs. 8 and 4; a first report from a system viewpoint on burn-in options appears in Ref. 16, and an optimal-decision-making model of the conceptual system burn-in is given in Ref. 37.

Burn-in Conditions and Types. During burn-in, ICs 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 (38), 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, and accelerating factors. Hamilton (39) 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 Refs. 14, 17 and 21.

Among burn-in approaches, four are particularly effective for semiconductor devices (4, 40): 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 (40). However, DBI places active signals on ICs, 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 (41).

When failures that are not temperature-dependent are not well detected by the normal burn-in, high voltage is often applied during the burn-in. Many memory IC manufacturers are using high-voltage stress tests, SBI with reverse bias, and DBI to detect gate oxide defects (38).

There are three burn-in types based on product levels (42, 43): 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 a 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 scrapped 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 dies (*KGDs*) is another motive for developing more efficient burn-in technology. Conventional burn-in can not support the burn-in of bare die.

DLBI is an extension of 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 to use DLBI. One advantage of DLBI is that it can provide burned-in and tested KGDs.

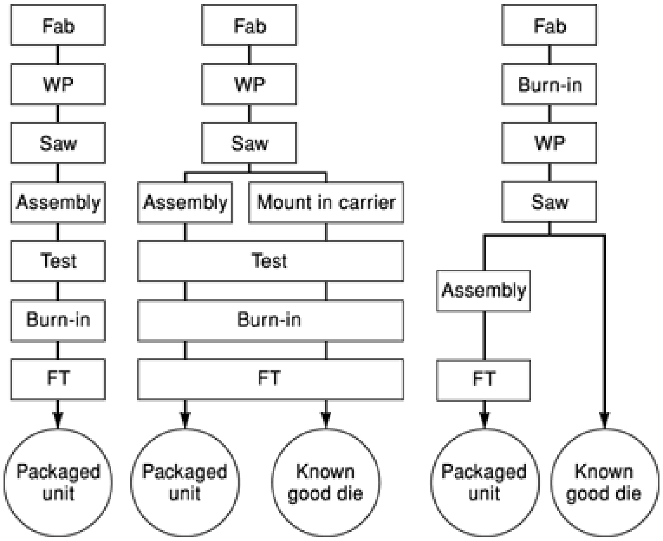

Fig. 3. Comparison of three burn-in flows.

WLBI simultaneously puts stresses on every die of the entire wafer before performing burn-in. Since the burn-in is performed at an earlier stage of product flow, WLBI can remove initial reliability failures earlier at reduced cost.

Demand for smaller-size and lighter-weight information systems is a trend in the multimedia era. However, the mounting technology in electronics systems is mature; therefore it is time for manufacturers to supply KGDs to the market. WLBI is essential for producing KGDs.

Another advantage of WLBI is the fast feedback of yield and defect data, which makes manufacturing processes more proactive in fault correction. Though the idea of applying burn-in at the wafer level may have originated in the need to provide conditioned (or burned-in) KGDs, a successful WLBI results in considerable cost reduction for all IC products.

There already exist some experimental implementations of WLBI (42, 44,45,46). Flynn and Gilg (47) 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 the three types are compared in Fig. 3 (42, 43). The high initial cost of WLBI is a major concern for its implementation. However, the initial cost can be reduced by equipment cost reduction and better equipment centralization and utilization (42).

Modeling Yield

Yield and Reliability. Among the performance indices 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 (48).

Since yield is a statistical parameter, yield functions at different manufacturing stages are multiplied in order to attain the total yield. Yield is a function not only of chip area but also of circuit design and layout. The total yield is the number 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 ICs (49).

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 (27). 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 classes of parameters that significantly affect the yield and reliability of ICs: design-related parameters such as chip area, gate oxide thickness, and junction depth; manufacturing-related parameters such as defect distribution and density; and operation-related parameters such as temperature and voltage. In general, reliability depends on all three classes of 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 yield–reliability relation and its modeling are based on this point of view.

Frost and Poole (40a) developed a series model to determine the intrinsic reliability of ICs, which assumes wearout-limited reliability determined by defects. Stevenson and Nachlas (50) used the POF approach to derive the relation between imperfections and the ultimate reliability of ICs. Jensen (51) showed that there exists a strong correlation between yield and reliability by surveying published papers and addressing yield models. He also argued that the size and location of defects determine whether the defects are yield-related or reliability-related.

The presence of defects in ICs affects the yield as well as the reliability. Bruls (52) studied the reliability aspect of defects and calculated the single-fault probability, because he observed that the number of defects in a mature process is limited to one or a fews and a single defect usually influences the reliability of an IC. Prendergast (53) pointed out a linear relationship between yield and reliability and suggested 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 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 of 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 improve productivity.

IC device yields depend on many factors such as chip area, circuit design, and circuit layout. It is desirable to explain the overall yield mathematically and to effectively control and manage yields by determining the failure probabilities and the critical areas for each defect type (49).

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_{wp}), wafer probe yield (Y_{cp}), assembly yield (Y_{ap}), and final test yield (Y_{ft}). According to one survey by *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 term “line yield” is used interchangeably with “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 (48). Generally, wafer fabrication processes directly affect wafer process yield

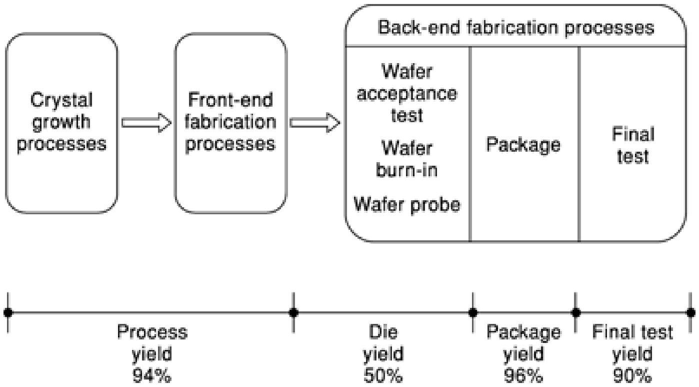

Fig. 4. Typical yield components commonly seen in semiconductor manufacturing.

(or line yield) and wafer probe yield (or die yield), and packaging processes influence assembly yield and final test yield.

The overall yield is defined as the product of yields from the several consecutive processes, or (48, 55)

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

The wafer process yield and wafer probe yield are the two most important factors influencing the productivity of semiconductor manufacturing. Because the wafer probe yield is the bottleneck of the overall yield, one must attain a high wafer probe yield to remain competitive.

Defects and Critical Area. For yield projection, it is useful to categorize defects as random or nonrandom (48, 55, 58). Random defects are defects that occur by chance. Particles that cause shorts and opens and local crystal defects are random defects. Nonrandom defects include gross defects and parametric defects.

Defects that cause circuit failures are called faults or fatal defects (48, 59, 60). The distinction between defects and faults plays an important role in calculating the yield based on the defect density and chip area. Another parameter that affects yield is defect clustering.

The defect size distribution depends on the process line, process time, learning experience, and other variables. It is usually accepted that there is a certain critical size at which the density function peaks, and decreases on either side of the peak (61, 62). 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 given by a power law for defects smaller than the critical size and by an inverse power law for defects larger than the critical size (48). Let x_0 be the critical size of the defect that is most likely to occur. The defect size pdf is defined below (63):

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

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 (62). Defects smaller than x_0 cannot be resolved well by optical monitoring (63). Since very small defects are assumed to increase linearly with defect size to a point x_0, Stapper (62, 63) indicates that using values of $q = 1$ and $p = 3$ for the spatial distribution agrees reasonably well with experimental data. There are other proposals for defect size distributions, such as the Rayleigh (64),

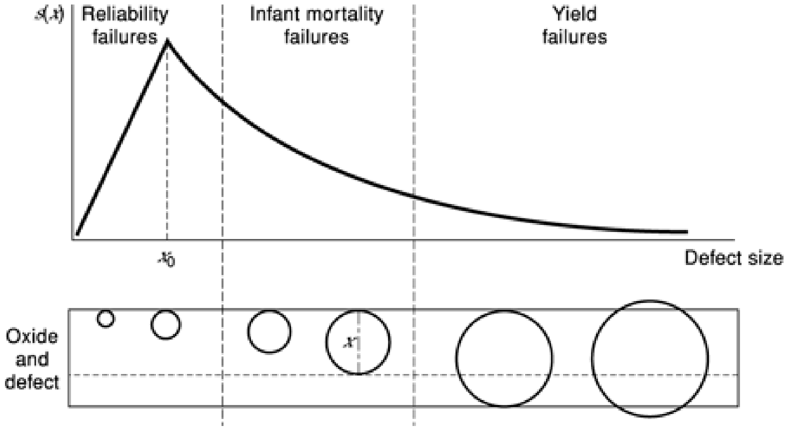

Fig. 5. The defect-size pdf and related oxide problems.

lognormal (65), and gamma (66) distributions. A typical distribution curve of $s(x)$ versus defect size for Eq. (3) is shown in Fig. 5, where the circles represent defects due to oxidation.

A *critical area* is an area where the center of a defect must fall to create a fault (62, 63). That is, if a defect occurs in a 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 in the integral form

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

The average defect density of all sizes and average defect density of size x are denoted by 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 as

$$\mu = A_c D_0$$

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 (67). That is, it is used to estimate the future yield of a current or new product and the yield loss from each of the process steps. Wallmark's model (68) is one of the earliest yield models. Among the models developed since 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 number of faults caused by defects μ, the probability that a chip contains k defects is

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

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 (48).

If the defect density is a random variable, the yield model is determined by its distribution. 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 (69).

Assume that the defect density follows a gamma distribution:

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

where α and β are the shape and scale parameters, respectively. Then the probability that one chip contains k defects follows the negative binomial distribution

$$P_k = \int_0^{\infty} \frac{e^{-A_c D}(A_c D)^k}{k!} \frac{D^{\alpha-1} e^{-D/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad dD = \frac{\Gamma(\alpha+k)(A_c\beta)^k}{k!\Gamma(\alpha)(1+A_c\beta)^{\alpha+k}}$$

Therefore, the yield model is given by

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

The clustering factor α determines the degree of clustering of the model. If α is equal to 1, then Eq. (5) is equivalent to Seed's yield model in Eq. (6) below. If α goes to ∞, then Eq. (5) gives the same result as the Poisson model in Eq. (4), implying no clustering. The practical range of α is 0.3 to 5.0. Stapper (70, 71) reports that this model fits actual yield data well. Stapper (60, 72, 73) explains the effects of clustering on yield. For the same average defect density, clustering usually gives a higher chip yield (67). Figure 6 shows configurations of two different degrees of defect 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 the Poisson model, and the high-clustering situation resembles more a negative binomial model of smal α value.

If we assume that defect density follows a normal distribution that 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 tolerance (74).

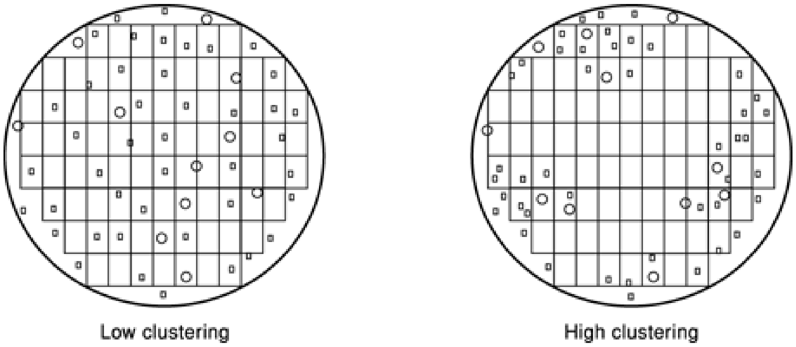

Fig. 6. Comparison of two degrees of defect clustering for the same average defect density.

The assumption that the defect density is exponentially distributed gives seed's yield model, which is expressed by

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

Seed's model generally gives higher yields than the actual observations (48). 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 a yield higher than the observed yield (74).

Okabe et al. (76) present another yield model, which is based on the 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 (71).

Figure 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.

Different Approaches to Yield Modeling. Berglund (77) presents a variable defect size (*VDS*) yield model including both conventional small defects and much larger parametric or area defects. To do this, Eq. (4) can be modified as

$$Y = \exp\left[- \int A_c(x) D(x)\, dx \right]$$

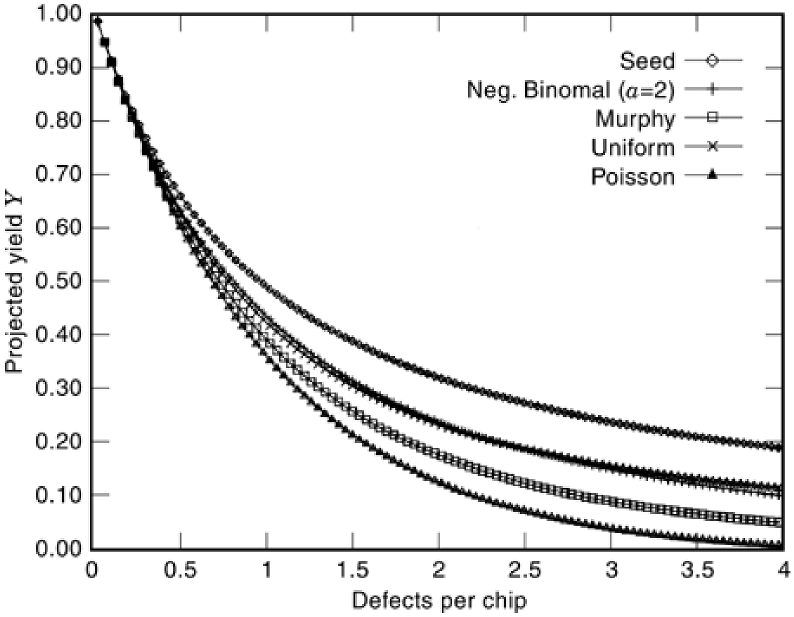

Fig. 7. Comparison of five yield models.

Assuming 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 that are comparable to or larger than the die size. Berglund (77) shows that Eq. (7) can be written as the product of two exponential factors: $Y = Y_0 Y_\mathrm{p}$. Here the die-area-independent yield loss factor Y_0 can be viewed as the gross yield, and the additional die-area-dependent factor Y_p represents the added yield loss around the edges of the larger parametric defects. Berglund (77) concludes that on 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 over a wide range for chips with the same areas, because of the variation in circuit density. Stapper (49) presents a circuit-count approach to yield modeling that 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 popular yield model. According to this 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 (coefficient of determination) $R^2 = 0.6$ is given by (57)

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

where

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

The accuracy of an absolute yield model depends upon the 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 critical support area. The core-area yield includes the chance of defects being repaired. To find it, 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 Eqs. (8) and (10), the die yield is the product 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 (48). All parameters except the 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 (48) suggests a method to predict the productivity of a new product, q quarters after the start of normal production, which is given by

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

Table 4. Wafer Processing Cost for Four Sizes (in 1994)

Cost Factor	100 mm	125 mm	150 mm	200 mm
Facility cost	\$75 M	\$175 M	\$400 M	\$750 M
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

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 their application to the acceleration of learning. Using yield models to accelerate the progress of a learning curve reduces the learning cycle time and thus helps 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, the equipment particulate characteristic, and yield confidence intervals. Other yield models used in various companies are well summarized in Ref. 49.

All the yield models are used as planning tools. Depending on the applications and product history, specific models can be selected. Figure 7 provides a guideline for such a comparison.

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 only, because assembly and final testing may be performed at different sites. Table 3 shows an example of cost analysis for DRAM and Pentium chips (80).

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%) (80). Table 4 shows wafer processing cost factors for four wafer sizes (80). The wafer processing cost divided by the wafer process yield is the yielded wafer processing cost. In Table 4, for the size 200 mm, the yield wafer processing cost is \$992. If we assume the total number of dice available per wafer is 162 for 16 Mbit DRAM (Table 3), the yielded wafer processing cost per die is \$6.12.

Wafer Probe Cost. The wafer probe cost adds about 5% 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, the wafer probe cost is high in the development stage of devices. The yielded wafer probe cost is calculated by

$$\frac{\text{wafer probe cost}}{\text{wafer probe yield}} + (1 - \text{wafer probe yield}) \times (\text{wafer processing cost})$$

Table 5. An Example of Package Cost of DRAM and Pentium (in 1998)

Product	16 Mbit DRAM	64 Mbit DRAM	Pentium P54CS
Feature size	0.35 μm	0.35 μm	0.35 μm
Package cost	\$0.40	\$0.50	\$25.75
Assembly yield	99%	99%	99%

In this case, the wafer probe cost is

$$\frac{6.12 \times 0.05}{0.40} + 0.60 \times 6.12 = \$4.44$$

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 DRAMs (80). The yielded assembly and packaging cost is obtained by

$$\frac{\text{assembly and packaging cost}}{\text{assembly yield}} + (1 - \text{assembly yield}) \\ \times (\text{yielded wafer processing cost} + \text{yielded wafer probe cost})$$

In this example, we have

$$\frac{0.5}{0.99} + 0.01 \times (6.12 + 4.44) = \$0.61$$

Final Test Cost. The 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 (80): The yielded final test cost is calculated by

$$\frac{\text{final test cost}}{\text{final test yield}} + (1 - \text{final test yield}) \times (\text{yielded wafer processing cost} + \text{yielded wafer probe cost} \\ + \text{yielded assembly and packaging cost})$$

In this case, the yielded final test cost is

$$\frac{0.6}{0.95} + 0.05 \times (6.12 + 4.44 + 0.61) = \$1.19$$

Therefore, the total cost for a 200 mm 16 M bit DRAM is

$$6.12 + 4.44 + 0.61 + 1.19 = \$12.36$$

Table 6. Final Test Cost and Yield (in 1998)

Product	16 Mbit DRAM	64 Mbit DRAM	Pentium P54CS
Feature size	0.35 μm	0.35 μm	0.35 μm
Final test cost	\$0.60	\$1.20	\$35.00
Final test yield	95%	85%	70%

Fault Coverage and Occurrence

The defect level is defined as the percentage of defective circuits passing all phases of a manufacturing test (81), or the probability that any given IC has untested defects (82). Thus, the defect level represents the proportion of a product that may fail because of extrinsic failure (or infant mortality) (83). Let D_{L} be the defect level of the IC. Then it is given by (84)

$$D_{\mathrm{L}} = 1 - Y^{1-T} \tag{10}$$

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

The basic assumption of Eq. (10) is that all faults have equal probability of occurrence, which implies no clustering. That is, the faults are uniformly distributed. Corsi (81) extends this to 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 (85) combined 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 that the fault will occur on a chip. Their attempt was to find a fault occurrence probability for individual faults instead of a distribution for them. They called 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 (85)

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

where c is the total number of chips tested and c_i is the number of chips that fail exactly at test vector 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

$$D_{\mathrm{L}} = \frac{Y_n - Y}{Y_n}$$

Since Eq. (10) does not provide good results when faults are correlated, Maxwell and Aitken (86) have presented another relation for the defect level:

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

where n_0 is the average number of faults on a die. Willing and Helland (82) present a mathematical model to develop fault coverage guidelines for complex electronic systems. Their model extends Eq. (11) with probabilistic relationships between yield, fault coverage, and defect level, and finds reliability to be a function of fault coverage and yield.

Yield–Reliability Relation Models

Yield–reliability Relation. 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 ICs, the development of new technologies is needed. These new technologies will make it possible to avoid wearout failures during the operational life.

The degree of manufacturing success is measured by the *yield*, which is defined as the average fraction 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 overall wafer yield is a measure of good chips per wafer normalized by the number of chip sites per wafer. The overall yield is calculated as the product of factors such as the line yield, the WP yield, the assembly yield, and the FT yield.

To maximize the yield, 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 (27). 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 in critical areas, and their random distribution in the process was observed not to yield the desired results of high quality and minimal defect density. There is another way, however, 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 ICs (49).

Schroen (87) 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.

Yield–reliability Relation Models. 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 between 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 (81, 84, 83, 88). If one wants to identify the effects of stresses or conditions causing the infant mortality failures, it is necessary to relate the reliability model to defect reliability physics and to describe that as the function of yield. Only two relation models have been reported so far. Huston and Clarke's model (83) uses the critical area for the yield and the reliability to model the relationship. In their model, for a given yield Y, the reliability R is given by

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

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

Kuper et al. (88) and Van der Pol et al. (54) use the same model for the yield–reliability relation and present experimental data to show the correlation. They express the model as

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

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 Eq. (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 Eq. (13) to five different ICs in order to verify the existence of a strong relationship between yield and failure occurring early in the lifetime.

In general, reliability is the ability of the product to operate properly without failure and is defined as the cumulative probability function at time t for a given time under the operating conditions. Note that the models in Eqs. (12) and (13) are not related to t. Thus, the reliabilities used in the two relation models are not defined at a specific time t, but are average fractions of devices working properly early in their lives.

Kuo et al. (4) and Kim et al. (89,90,91) use a different relation model, which is defined at time t and based on the POF concept. Let $R(t)$ and $c(t)$ be the reliability at time t and a time-dependent constant, respectively. Then, the 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 to burn in can be made by observing the yield. This is another way to avoid time-consuming burn-in.

There are recent developments in predicting the relationship between reliability and yield of semiconductor manufacture. Some of them include improvement of yield modeling (92, 93), new methods for yield enhancement (94, 95), and discussions on correlation between reliability and yield for certain production chips (96, 97).

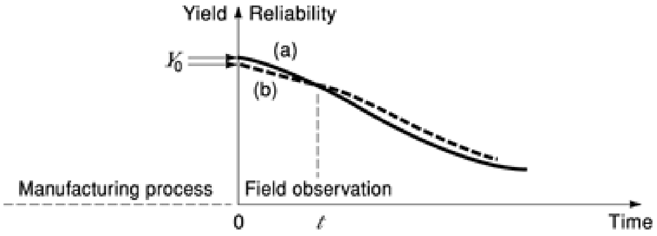

Fig. 8. The decreasing function of reliability with respect to time for different yields.

Conclusions

Because 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 (4). In practice, many modern devices have lifetimes close to 5 years. 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 marketplace 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, ICs for applications with drastic consequences of failure need to be subjected to a full screening procedure before they are assembled into a dedicated system.

A high yield means a 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 a specific process that manufacturing engineers in the semiconductor industry often refer to is presented as Y_0 in Fig. 8. Assume the time at completion of this process is zero, as indicated in the Fig. 8. Beyond time zero, the probability of failure-free operation of a device is called the *reliability*. In general, the lower Y_0 is, the lower will be the reliability as a function of time. The decrease of reliability may be caused by intrinsic, extrinsic, or wearout failures. Therefore, the yield and reliability of microelectronics manufacturing products are highly related, but high manufacturing yield does not necessarily imply high reliability of the products of 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 higher 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 (98), 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, because it presents 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 a direct effect on manufacturing yield. Tradeoffs among these factors are essential in order to guarantee high reliability in semiconductor products.

Acknowledgments

This chapter is based largely, with updated information, on “An Overview of Manufacturing Yield and Reliability Modeling for Semiconductor Products,” by W. Kuo and T. Kim, which appeared in *Proceedings of the IEEE*, Vol. 87, No. 8, pp. 1329–1344, Aug. 1999 (99), © 1999 IEEE. This material is based in part upon work supported by the Texas Advanced Technology Program under grant ATP-036327-138, and NSF project DMI-9908207.

Acronyms and Symbols

BIR = Built-in reliability
CMOS = Complementary metal oxide semiconductor
CP = Chip probing
DBI = Dynamic burn-in
DFR = Decreasing failure rate
DLBI = Die level burn-in
DOD = U.S. Department of Defense
DRAM = Dynamic random access memory
DUT = Device under test
EOS = Electrical overstress
ESD = Electrostatic discharge
FA = Failure analysis
FIT = Failure per 10^9 device-hours
FT = Final test
HTOL = High-temperature operating life
IC = Integrated circuit
ICE = Integrated Circuit Engineering Corp.
KGD = Known good die
LSI = Large-scale integration
MOS = Metal oxide semiconductor
MSI = Medium-scale integration
PC = Personal computer
PCB = Printed circuit board
pdf = Probability density function
PLBI = Package-level burn-in
POF = Physics of failure
QC = Quality control
QML = Qualified manufacturing line
SBI = Static burn-in
SIMM = Single in-line memory module
SLSI = Super-large-scale integration
SPC = Statistical process control
SSBI = Steady-state burn-in
SSI = Small-scale integration
TDBI = Test during burn-in
TDDB = Time-dependent dielectric breakdown
ULSI = Ultra-large-scale integration
VDS = Variable defect size

VLSI = Very large-scale integration
WAT = Wafer acceptance test
WLBI = Wafer-level burn-in
WLR = Wafer-level reliability
WP = Wafer probe
WSTS = World Semiconductor Trade Statistics
Y_{wp} = Wafer process yield (line yield)
Y_{cp} = Wafer probe yield (die yield)
Y_{ap} = Wafer assembly yield
Y_{ft} = Final test yield

BIBLIOGRAPHY

1. A. G. Sabnis *VLSI Electronics Microstructure Science, Vol. 22, VLSI Reliability*, San Diego, CA: Academic Press, 1990.
2. E. R. Hnatek *Integrated Circuit Quality and Reliability*, 2 ed., New York: Marcel Dekker, 1995.
3. *World Semiconductor Trade Statistics (WSTS), Press Release, May 2001*, Spring Forecast Session, May 15–18, 2001 [Online]. Available WWW: http://www.wsts.org
4. W. Kuo W. T. K. Chien T. Kim *Reliability, Yield, and Stress Burn-in*, Norwell, MA: Kluwer Academic, 1998.
5. A. Amerasekera D. S. Campbell *Failure Mechanisms in Semiconductor Devices*, New York: Wiley, 1987.
6. M. Campbell Monitored burn-in improves VLSI IC reliability, *Computer Design*, **24** (4): 143–146, April 1985.
7. D. L. Denton D. M. Blythe The impact of burn-in on IC reliability, *J. Environ. Sci.*, **29** (1): 19–23, Jan./Feb. 1986.
8. W. Kuo Y. Kuo Facing the headaches of early failures: A state-of-the-art review of burn-in decisions, *Proc. IEEE*, **71**: 1257–1266, 1983.
9. D. Chi W. Kuo Burn-in optimization under reliability & capacity restrictions, *IEEE Trans. Reliab.*, **38**: 193–198, 1989.
10. K. Chou K. Tang "Burn-in time and estimation of change-point with Weibull–exponential mixture distribution," *Decision Sci.*, **23** (4): 973–990, 1992.
11. D. G. Nguyen D. N. P. Murthy Optimal burn-in time to minimize cost for products sold under warranty, *IIE Trans.*, **14** (3): 167–174, 1982.
12. W. Kuo *et al. Optimal Reliability Design: Fundamentals and Applications*, Cambridge UK: Cambridge University Press, 2000.
13. MIL-STD-280A, *Definitions of Item Levels, Item Exchangibility, Models and Related Terms*, Philadelphia: The Naval Publications and Forms Center, 1969.
14. W. T. K. Chien W. Kuo A nonparametric approach to estimate system burn-in time, *IEEE Trans. Semicond. Manuf.*, **9**: 461–466, 1996.
15. C. W. Whitbeck L. M. Leemis Component vs system burn-in techniques for electronic equipment, *IEEE Trans. Reliab.*, **38**: 206–209, 1989.
16. W. Kuo Reliability enhancement through optimal burn-in, *IEEE Trans. Reliab.*, **R-33**: 145–156, 1984.
17. W. T. K. Chien W. Kuo Modeling and maximizing burn-in effectiveness, *IEEE Trans. Reliab.*, **44**: 19–25, 1995.
18. L. M. Leemis M. Beneke Burn-in models and methods: A review, *IIE Trans.*, **22** (2): 172–180, 1990.
19. W. Kuo Incompatibility in evaluating large-scale systems reliability, *IEEE Trans. Reliab.*, **43**: 659–660, 1994.
20. M. Haim Z. Porat Bayes reliability modeling of a multistate consecutive K-out-of-n: f system, *Annual Reliability and Maintainability Symp.*, 1991, pp. 582–586.
21. W. T. K. Chien W. Kuo A nonparametric Bayes approach to decide system burn-in time, *Naval Res. Logist.*, **44** (7): 655–671, 1997.
22. T. A. Mazzuchi N. D. Singpurwalla A Bayesian approach for inference for monotone failure rates, *Statist. Probab. Lett.*, **37**: 135–141, 1985.
23. E. Takeda *et al.* VLSI reliability challenges: From device physics to wafer scale systems, *Proc. IEEE*, **81**: 653–674, 1993.
24. D. L. Crook Evolution of VLSI reliability engineering, *Proc. International Reliability Physics Symp.*, 1990, pp. 2–11.

25. A. Christou *Integrating Reliability into Microelectronics Manufacturing*, Chichester: Wiley, 1994.
26. C. Hu Future CMOS scaling and reliability, *Proc. IEEE*, **81**: 682–689, 1993.
27. J. A. Shideler *et al.* A systematic approach to wafer level reliability, *Solid State Technol.*, **38** (3): 47, 48, 50, 52, 54, March 1995.
28. T. A. Dellin *et al.* Wafer level reliability, *SPIE Microelectronics Manufacturing and Reliability, Proc. Int. Soc. Opt. Eng.*, 1992, pp. 144–154.
29. A. P. Bieringer *et al.* Implementation of a WLR-program into a production line, *1995 IRW Final Report*, 1996, pp. 49–54.
30. S. Garrard Production implementation of a practical WLR program, *1994 IRW Final Report*, 1995, pp. 20–29.
31. T. E. Kopely *et al.* Wafer level hot-carrier measurements for building-in reliability during process development, *1994 IRW Final Report, IEEE Int. Integrated Reliability Workshop*, 1995, pp. 57–59.
32. L. N. Lie A. K. Kapoor Wafer level reliability procedures to monitor gate oxide quality using V ramp and J ramp test methodology, *1995 IRW Final Report, IEEE Int. Integrated Reliability Workshop*, 1996, pp. 113–121.
33. O. D. Trapp (ed.) *1991 International Wafer Level Reliability Workshop*, Lake Tahoe, CA, 1991.
34. T. E. Turner Wafer level reliability: Process control for reliability, *Microelectron. Reliab.*, **36** (11/12): 1839–1846, 1996.
35. J. M. Soden R. E. Anderson IC failure analysis: Techniques and tools for quality and reliability improvement, *Proc. IEEE*, **81**: 703–715, 1993.
36. B. Schlund *et al.* A new physics-based model for time-dependent dielectric breakdown, *Proc. Int. Reliability Physics Symp.*, 1996, pp. 84–92.
37. T. Kim W. Kuo Optimal burn-in decision making, *J. Quality Reliab. Int.*, **14** (6): 417–423, 1998.
38. E. R. Hnatek A realistic view of VLSI burn-in II, *Evaluation Eng.*, **28** (2): 80, 82–86, 89, 1989.
39. H. E. Hamilton An overview—VLSI burn-in considerations, *Evaluation Eng.*, **31** (2): 16, 18–20, 1992.
40. D. Romanchik Why burn-in ICs ? *Test & Measurement World*, **12** (10): 85–86, 88, Oct. 1992. D. F. Frost K. F. Poole A method for predicting VLSI-device reliability using series models for failure mechanisms, *IEEE Trans. Reliab.*, **R-36**: 234–242, 1987.
41. D. Romanchik Burn-in: Still a hot topic, *Test & Measurement World*, **12** (1): 51–52, 54, Jan. 1992.
42. D. Gralian Next generation burn-in development, *IEEE Trans. Compon. Packag. Manuf. Technol. B, Adv. Packag.*, **17**: 190–196, 1994.
43. B. Vasquez S. Lindsey The promise of known-good-die technologies, *MCM '94 Proc.*, 1994, pp. 1–6.
44. A. Martin *et al.* Assessing MOS gate oxide reliability on wafer level with ramped/constant voltage and current stress, *1995 IRW Final Report, IEEE Int. Integrated Reliability Workshop*, 1996, pp. 81–91.
45. A. D. Singh On wafer burn-in strategies for MCM die, *Int. Conf. Exhibition Multichip Modules*, 1994, pp. 255–260.
46. D. B. Tuckerman *et al.* A cost-effective wafer-level burn-in technology, *Int. Conf. Exhibition on Multichip Modules*, 1994, pp. 34–40.
47. W. G. Flynn L. Gilg A pragmatic look at wafer-level burn-in: The wafer-level known-good-die consortium, *IECEM '96 Proc.*, 1996, pp. 287–292.
48. A. V. Ferris-Prabhu *Introduction to Semiconductor Device Yield Modeling*, Boston: Artech House, 1992.
49. C. H. Stapper R. J. Rosner Integrated circuit yield management and yield analysis: Development and implementation, *IEEE Trans. Semicond. Manuf.*, **8**: 95–102, 1995.
50. J. L. Stevenson J. A. Nachlas Microelectronics reliability predictions derived from components defect densities, *Annual Reliability and Maintainability Symp.*, 1990, pp. 366–371.
51. F. Jensen Yield, quality and reliability—a natural correlation? in R. H. Matthews (ed.), *Reliability '91*, London: Elservier Applied Science, 1991, pp. 739–750.
52. E. M. J. G. Bruls Reliability aspects of defect analysis, *IEEE/ETC*, 1993, pp.17–26.
53. J. G. Prendergast Reliability and quality correlation for a particular failure mechanism, *Proc. Int. Reliability Physics Symp.*, 1993, pp. 87–93.
54. J. Van der Pol F. Kuper 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. Reliab.*, **36** (11/12): 1603–1610, 1996.
55. B. El-Kareh A. Ghatalia A. V. S. Satya Yield management in microelectronic manufacturing, *Proc. 45th Electronic Components Conf.*, 1995, pp. 58–63.

56. Integrated Circuit Engineering Corp., *Cost Effective IC Manufacturing 1995*, Scottsdale, AZ, 1995.
57. S. P. Cunningham C. J. Spanos K. Voros Semiconductor yield improvement: Results, and best practices, *IEEE Trans. Semicond. Manuf.*, **8**: 103–109, 1995.
58. C. H. Stapper F. M. Armstrong K. Saji Integrated circuit yield statistics, *Proc. IEEE*, **71**: 453–470, 1983.
59. T. L. Michalka R. C. Varshney J. D. Meindl A discussion of yield modeling with defect clustering, circuit repair, and circuit redundancy, *IEEE Trans. Semicond. Manuf.*, **3**: 116–127, 1990.
60. C. H. Stapper The effects of wafer to wafer defect density variations on integrated circuit defect and fault distributions, *IBM J. Res. Devel.*, **29**: 87–97, 1985.
61. A. V. Ferris-Prabhu "Defect size variations and their effect on the critical area of VLSI devices," *IEEE J. Solid State Circuits*, **SC-20**: 878–880, 1985.
62. C. H. Stapper Modeling of integrated circuit defects sensitivities, *IBM J. Res. Devel.*, **27**: 549–557 (1983).
63. C. H. Stapper Modeling of defects in integrated circuit photolithographic patterns, *IBM J. Res. Devel.*, **28**: 461–475, 1984.
64. W. Maly Modeling of lithography related yield loss for CAD of ULSI circuits, *IEEE Trans. Comput.-Aided Design*, **CAD-4**: 166–177, 1985.
65. C. Kooperberg Circuit layout and yield, *IEEE J. Solid-State Circuits*, **23**: 887–892, 1988.
66. Z. Stamenkovic N. Stojadinovic New defect size distribution function for estimation of chip critical area in integrated circuit yield models, *Electron. Lett.* **28** (6): 528–530, 1992.
67. A. Ghatalia B. El-Kareh *Yield Management in Microelectronic Manufacturing*, Short Course Notes, Austin, TX: National Alliance for Photonics Education in Manufacturing, 1996.
68. T. J. Wallmark Design considerations for integrated electron devices, *Proc. IRE*, **48**: 293–300, 1960.
69. A. V. Ferris-Prabhu Models for defects and yield, in I. Koren (ed.), *Defect and Fault Tolerance in VLSI Systems*, New York: Plenum Press, 1989, pp. 33–46.
70. C. H. Stapper Defect density distribution for LSI yield calculations, *IEEE Trans. Electron Devices*, **ED-20**: 655–657, 1973.
71. C. H. Stapper Fact and fiction in yield modeling, *Microelectron. J.*, **20** (1/2): 129–151, 1989.
72. C. H. Stapper On yield, fault distributions and clustering of particles, *IBM J. Res. Devel.*, **30**: 326–338, 1986.
73. C. H. Stapper Large-area fault clusters and fault tolerance in VLSI circuits: A review, *IBM J. Res. Devel.*, **33**, 162–173, 1989.
74. B. T. Murphy Cost-size optima of monolithic integrated circuit, *Proc. IEEE*, **52**: 1537–1545, 1964.
75. J. E. Price A new look at yield of integrated circuits, *Proc. IEEE*, **58**: 1290–1291, 1970.
76. T. Okabe M. Nagata S. Shimada Analysis on yield of integrated circuits and a new expression for the yield, *Electrical Eng. Japan*, **92** (6): 135–141, 1972.
77. C. N. Berglund "A unified yield model incorporating both defect and parametric effects," *IEEE Trans. Semicond. Manuf.*, **9**: 447–454, 1996.
78. D. Dance R. Jarvis Using yield models to accelerate learning curve progress, *IEEE Trans. Semicond. Manuf.*, **5**: 41–45, 1992.
79. Semiconductor Industry Association, *1978–1993 Industry Data Book*, 1994.
80. Integrated Circuit Engineering Corp., *Cost Effective IC Manufacturing 1998–1999*, Scottsdale, AZ, 1997.
81. F. Corsi S. Martino Defect level as a function of fault coverage and yield, *Proc. European Test Conf.*, 1993, pp. 507–508.
82. W. Willing A. Helland Establishing ASIC fault-coverage guidelines for high-reliability systems, *Annual Reliability and Maintainability Symp.*, 1998, pp. 378–382.
83. H. H. Huston C. P. Clarke Reliability defect detection and screening during processing—theory and implementation, *Proc. International Reliability Physics Symp.*, 1992, pp. 268–275.
84. T. W. Williams N. C. Brown Defect level as a function of fault coverage, *IEEE Trans. Comput.*, **C-30**: 508–509, 1981.
85. S. C. Seth V. D. Agrawal On the probability of fault occurrence, in I. Koren (ed.), *Defect and Fault Tolerance in VLSI Systems*, New York: Plenum, 1989, pp. 47–52.
86. P. Maxwell R. Aitken Test sets and reject rates: All fault coverages are not created equal, *IEEE Design and Test of Computers*, 10 (1): 42–51, March 1993.
87. W. H. Schroen Process testing for reliability control, *Proc. Int. Reliability Physics Symp.*, 1978, pp. 81–87.
88. F. Kuper *et al.* Relation between yield and reliability of integrated circuits: Experimental results and application to continuous early failure rate reduction programs, *Proc. Int. Reliability Physics Symp.*, 1996, pp. 17–21.

89. T. Kim W. Kuo W. T. K. Chien A relation model of yield and reliability for gate oxide failures, *1998 Annual Reliability and Maintainability Symp.*, Anaheim, CA, 1998, pp. 428–433.
90. T. Kim W. Kuo Modeling manufacturing yield and reliability, *IEEE Trans. Semicond. Manuf.*, **12**: 485–492, 1999.
91. T. Kim W. Kuo W. T. K. Chien "Burn-in effect on yield," *IEEE Trans. Electron. Packag. Manuf.*, **23**: 293–299, 2000.
92. T. Chen M. J. Wang "Fuzzy set approach for yield learning modeling in wafer manufacturing," *IEEE Trans. Semicond. Manuf.*, **12**: 252–258, 1999.
93. M. Recio Strategy and tools for yield enhancement, *Proc. SPIE Proc. Int. Soc. Opt. Eng.*, **3743**: 122–129, 1999.
94. C. Jun *et al.* Simulation-based semiconductor chip yield model incorporating a new defect cluster index, *Microelectron Reliab.*, **39** (4): 451–456, 1999.
95. C. J. McDonald New tools for yield improvement in integrated circuit manufacturing: Can they be applied to reliability? *Microelectron. Reliab.*, **39** (6): 731–739, 1999.
96. P. W. Mason *et al.* Relationship between yield and reliability impact of plasma damage to gate oxide, *Int. Symp. on Plasma Processinduced Damage*, P2ID, 2000, pp. 2–5.
97. W. C. Riordan R. Miller J. Hicks Reliability versus yield and die location in advanced VLSI, *Microelectron. Reliab.*, **39** (6): 741–749, 1999.
98. S. Tang New burn-in methodology based on IC attributes, family IC burn-in data, and failure mechanism analysis, *Proc. Annual Reliability and Maintainability Symp.*, 1996, pp. 189–190.
99. W. Kuo T. Kim An overview of manufacturing yield and reliability modeling for semiconductor products, *Proc. IEEE*, **87**: 1329–1344, 1999.

READING LIST

D. L. Erhart *et al.* On the road to building-in reliability, *1995 IRW Final Report, IEEE Int. Integrated Reliability Workshop*, 1996, pp. 5–10.

M. Pecht A. Dasgupta Physics-of-failure: an approach to reliable product development, *1995 IRW Final Report*, 1996, pp. 1–4.

C. H. Stapper W. A. Klaasen The evaluation of 16-Mbit memory chips with built-in reliability, *Proc. Int. Reliability Physics Symp.*, 1992, pp. 3–7.

WAY KUO
TAEHO KIM
Texas A&M University