Quantifying the Value of Ownership of Yield Analysis Technologies

Charles Weber, Vijay Sankaran, Kenneth W. Tobin, Jr., *Member, IEEE*, and Gary Scher

***Abstract*—A model based on information theory, which allows yield managers to determine optimal portfolio of yield analysis technologies for both the R&D and volume production environments, is presented. The information extraction per experimentation cycle and information extraction per unit time serve as benchmarking metrics for yield learning. They enable yield managers to make objective comparisons of apparently unrelated technologies. Combinations of four yield analysis tools—electrical testing, automatic defect classification, spatial signature analysis and wafer position analysis—are examined in detail to determine the relative value of ownership of different yield analysis technologies.**

***Index Terms*—Analysis, information theory, learning, management, ownership, value, yield.**

I. Introduction

COST-OF-OWNERSHIP models, which have been customized for the semiconductor industry over the last 15 years [1]–[6], give managers a very good idea of the true cost of technologies. However, managers of integrated circuits also need to assess the value of ownership (VoO) of technologies, in order to make optimal decisions on which technologies to purchase or develop.

Manufacturers of integrated circuits invest billions of dollars in process equipment, and they are interested in obtaining as rapid a return on their investment as possible. Rapid yield learning is thus becoming an increasingly important source of competitive advantage. The sooner a potentially lucrative circuit yields, the sooner the manufacturer can generate a revenue stream. Conversely, rapid identification of the cause of yield loss can restore a revenue stream and prevent the destruction of material in process [7], [8].

A series of studies has shown that shortening the defect learning cycles accelerates yield learning by increasing experimentation capacity. Defects must be detected, analyzed and eliminated within increasingly shorter time periods. Consequently, successful yield improvement tends to consist of a total systems approach that involves electrical testing, defect inspection and *in situ* fault detection. A defect reduction team can thus develop true yield management capability by correlating data obtained from methods with short data cycles to those extracted from methods with longer ones. Once defect databases become large enough, signals from short-cycle methods can foreshadow effects on final yield [8]–[14].

Yield managers have a large but expensive arsenal of yield improvement tools and methods at their disposal, whose data cycles can vary by orders of magnitude. Different tools perform different functions under different conditions, and some combinations of tools and methods work better than others do. Yield managers need to know which combination of tools works the most effectively and the most cost-effectively, in order to maximize the profitability of their operations. Yield managers require metrics that allow them to assess the value of apparently unrelated options. In layman's terms, they need to effectively compare apples and oranges.

This paper uses a model based on information theory in an attempt to create an objective method of comparing technology options for yield analysis. The information extraction per experimentation cycle and information extraction per unit time serve as benchmarking metrics for yield learning. Combinations of four yield analysis technologies—electrical testing (ET), automatic defect classification (ADC), spatial signature analysis (SSA) and wafer position analysis (WPA)—are examined in detail to determine an optimal yield management strategy for both the R&D and volume production environments.

II. Converting Data Into Knowledge

Yield learning characterizes the radical experience curves of the semiconductor industry, where enormous investments need to be recovered in a relatively short time [13], [14]. Yield learning is an iterative experimentation process, which is repeated until all sources of yield loss are detected, identified and eliminated, or until the cost of further experimentation exceeds the benefit of the knowledge gained [15], [16]. Yield learning can be accelerated by shortening the experimentation cycle, or by making each experimentation cycle more effective. The former option depends upon the engineering team's ability to reduce the design time of an experiment, the fabrication facility's ability to reduce the fabrication cycle time, the test area's ability to accelerate the data generation rate, and the engineering team's ability to increase the data analysis rate. The latter option depends upon how the data are analyzed, and to what extent they are converted into knowledge, where knowledge is defined as certain information.

Information theory provides and excellent metric for how effectively information is converted to knowledge—the entropy

Manuscript received February 1, 2001; revised February 6, 2002.

C. Weber is with MIT Sloan School of Management, Cambridge, MA 02142 USA (e-mail: WeberCM@ATTGlobal.net).

V. Sankaran is with Inventes Corporation, Austin, TX 78735 USA (e-mail: VJ@Rocketmail.com).

K. W. Tobin, Jr. is with Image Science and Machine Vision, Oak Ridge National Laboratories, Oak Ridge, TN 37831-6010 USA.

G. Scher is with Sleuthworks, Inc., Fort Collins, CO 80525 USA (e-mail: gms@sleuthworks.com).

Digital Object Identifier 10.1109/TSM.2002.804876

0894-6507/02$17.00 © 2002 IEEE

of the information source [17]–[20]. A source of information reveals an amount of information $I(X_i)$ whenever the source is in state X_i [21]. $I(X_i)$ is, therefore, known as the ***self-information*** and is given by[1]

$$I(X_i) = -\log_2 P(X_i) \text{ bits} \tag{1}$$

where $P(X_i)$ is the probability of occurrence of state X_i. ***Information entropy*** is defined as expectation of $I(X_i)$, or the average amount of self-information per state [17]. It is given by the random variable

$$\begin{aligned} H(X_i) &\equiv \langle I(X_i) \rangle = \sum_{i=1}^{m} P(X_i) \cdot I(X_i) \\ &= -\sum_{i=1}^{m} P(X_i) \cdot \log_2 P(X_i) \text{ bits/state.} \end{aligned} \tag{2}$$

Information entropy is at a maximum when all states are equiprobable or $P(X_i) = 1/m$, a situation that reflects maximum ignorance about the information source. Information entropy decreases from its maximum value as $P(X_i)$ concentrates into fewer states, approaching zero as the probability of one state approaches unity, and the probability of all other states approach zero. In other words, information entropy approaches zero as information becomes knowledge.

The ***relative entropy*** of a probability distribution $P(X_i)$ with respect to a second probability distribution $P(Y_i)$ is given by the Kullback–Leibler formula

$$H_{P(X_i)||P(Y_i)} = \sum_{i=1}^{m} P(X_i) \cdot \log_2 \frac{P(X_i)}{P(Y_i)} \text{ bits/state} \tag{3}$$

where the sum covers all possible states of the system [22], and $P(Y_i)$ plays the role of a reference measure. The relative entropy can thus be used to compare the final state of an experimentation cycle to the initial state or to benchmark the amount of information extraction performed by two different, possibly unrelated processes.

III. Problem Localization

Localizing the root cause of a problem to a particular process step or process technology constitutes the stated objective of yield analysis. It has also been shown to be most valuable stage of the problem-solving process. During problem localization, problem solvers that specialize in yield analysis or process integration, execute a sequence of trial-and-error procedures, which concentrate the probability of finding the root cause of a problem into fewer and fewer process steps. Once the odds of finding the root cause of a problem in a particular process step are close to 100%, the yield analysis and process integration specialists call in specialists in technologies associated with the culprit process step to continue the problem-solving process [23].

[1]The base of log $P(X_i)$ determines the units of information. The binary $\log_2 P(X_i)$ is given in "bits;" the decimal $\log_{10} P(X_i)$ is given in "hartleys"; and the natural $\log_e P(X_i)$ is expressed in "nats" [19].

To adequately model the problem-localization process, one has to define a discrete random variable X, whose discrete states X_i represent the individual steps of the semiconductor process. Problem localization consists of concentrating $P(X_i)$—the probability that the root cause of the problem can be found in process step X_i—into fewer and fewer process steps. If $P(X_i)$ concentrates into a distribution of X_i that exhibits only one peak, then the variance of $P(X_i)$ can be used as a metric for the extent of problem localization. However, if $P(X_i)$ exhibits multiple peaks, the variance of $P(X_i)$ is not a good metric for the extent of problem localization because $P(X_i)$ can concentrate without causing a significant reduction in the variance of $P(X_i)$. In contrast, information entropy $H(X_i)$ is an excellent metric for the extent of problem localization because it decreases as probability concentrates regardless of the number of peaks in the probability distribution.

The advantages of entropy as a metric for problem localization can be demonstrated in a modern semiconductor process, which typically consists of 500 steps [24] termed X_1 through X_{500}. If we assume that prior to the application of a diagnostic technology the source of a fault is equally likely to reside in steps X_{100}, X_{200}, X_{300} and X_{400}, and infinitesimally unlikely to come from any other process step, then $P(X_{100}) = P(X_{200}) = P(X_{300}) = P(X_{400}) = 0.25$, and $P(X_{\text{other}}) = 0$. If problem solvers conduct an experiment that eliminates X_{200} and X_{300} as possible sources of the fault, then the probability distribution concentrates to $P(X_{100}) = P(X_{400}) = 0.5$, and $P(X_{\text{other}}) = 0$. As Table I illustrates, the variance of the distribution does not decrease as probability concentrates (it actually increases slightly), whereas the information entropy does. Consequently, the variance does not accurately track the level of knowledge regarding the location of the root cause, whereas information entropy does.

Previous experience at solving problems of a certain kind has been shown to determine strategies for subsequent problem solving and experimentation [25], [26], implying that the problem-localization process does not necessarily start from equiprobability between process steps. In many instances, the odds that the root cause of a problem can be associated with a specific process step can and must be estimated from the outcome of previous experiments. For example, the probability that high resistance in a metal-3 line width structure results from metal-3 deposition or the lithography step that determines metal-3 line width can in principle be estimated from a factory's historical records. If previous experiments have shown that the photolithographic step associated with metal-3 line width was the cause of high resistance in the line width structure in 90% previous occurrences the problem, and in the remaining 10% of all occurrences the root cause of the problem not to be related to lithography, then historical data contains enough information for a problem solver to correctly identify the photolithography step associated with metal-3 as the culprit step in 90% of all occurrences of the problem. The problem solvers in charge of problem localization subsequently face the challenge of converting this information into knowledge by elevating the probability of finding the problem's root cause in a specific process step to 100%. They begin the problem-localization

TABLE I
INFORMATION ENTROPY, VARIANCE AND STANDARD DEVIATION BEFORE AND AFTER DIAGNOSTIC ACTIVITY

	Information Entropy	***Variance***	***Standard Deviation***
Units	(Bits per Process Step)	(# of Process Steps)^2	(# of Process Steps)
Numerical Value Before Diagnostic Activity	2.000	16667	129
Numerical Value After Diagnostic Activity	1.000	22500	150

process at low entropy, and they subsequently try to reduce the entropy to zero.

IV. The Value of Yield Analysis Technologies

Semiconductor yield problems have the potential to induce severe losses. For example, if one process tool in a factory disperses 100 defects with half a micrometer in diameter onto product wafers, it can "kill" a significant portion of the 200 chips wafers with a diameter of 200 millimeters typically contain. Let us assume that on average the tool kills 25 out of 200 chips on every wafer, and that these chips are microprocessors that sell for \$100 a piece. The tool will therefore cause \$2500 of damage in the time that it takes to process one wafer, which may be as short as one minute. If the tool is situated near the beginning of the line, it may damage wafers that will take nearly two months to reach the end of the line, when the problem is discovered. In the interim, the tool continues to reduce the profit of the operation by \$2500/minute, \$150 000 per hour or more than \$2.5 million per day. If the problem is discovered after the first damaged wafer exits the end of the line, more than \$100 million will have been lost.

Clearly, it is in every semiconductor manufacturer's interest to discover process excursions and tool contamination early, even if they do not actually impact the yield. The potential damage is so great that the manufacturer needs to proactively respond to every defect signal [23]. Semiconductor manufacturers have thus invested in expensive in-line inspection tools, which detect contamination that may or may not cause electrical faults within a few hours of critical process steps. Defect analysis tools and defect sourcing methods reduce the data generated by inspection tools and point to the process steps that are the likely culprits of potential yield loss.

In an economic environment governed by radical experience curves, the actual value of information extraction also depends on the time required to reduce the data, or the length of the experimentation cycle. A tool that identifies the source of an electrical fault with absolute certainty but requires the full VLSI process to be completed may be less valuable than a tool that can identify the source of a fault with significant probability within a few hours.

In the following subsections, we assume that all combinations of yield analysis technologies operate on the same semiconductor process, a modern ultra-large-scale integration (ULSI) process that consists of 500 process steps. Under these circumstances, the information extraction rates per experimentation cycle and the information extraction rate per unit time can serve as benchmarking metrics for yield learning. Combinations of four yield analysis technologies—electrical testing (ET), automatic defect classification (ADC), spatial signature analysis (SSA) and wafer position analysis (WPA)—are examined in detail to determine an optimal yield management strategy for both the R&D and volume production environments.

A. Electrical Testing and Wafer Position Data

Chips on product wafers are electrically tested for functionality shortly after they emerge from the fabrication facility. The wafers also typically contain microelectronic test structures in the scribe line between the chips that may reveal characteristics of the fabrication process when subjected to a parametric test [27]. Functional testing alone identifies defective chips, but both functional and parametric testing of product wafers can localize the source of an electrical fault to within a neighborhood of a few process steps. In doing so they dramatically reduce the entropy of the information source. For example, if the fact of chip failure were the only information available to a yield engineer, the engineer would have a $1/500 = 0.002$ chance of identifying the culprit process step in a process that consists of 500 such steps. If, however, the engineer had access to the information that the NMOS transistor threshold voltage was out of specification, then the engineer could most likely use his/her expertise to reduce the source of the electrical fault to a neighborhood of about 20 process steps. Without access to the history of the process, the engineer would have to assume that each step has a $1/20 = 0.05$ chance of being the culprit. Assuming the potentially relevant steps are numbered 151 through 170, then substituting the aforementioned odds into Equation (3) yields

$$\begin{aligned} &H_{\text{POST}\|\text{PRE-TEST}} \\ &\quad= \sum_{i=1}^{150} 0 \cdot \log_2 \tfrac{0}{0.002} + \sum_{i=151}^{170} 0.05 \cdot \log_2 \tfrac{0.05}{0.002} \\ &\qquad+ \sum_{i=171}^{500} 0 \cdot \log_2 \tfrac{0}{0.002} \\ &\quad= 4.64 \text{ bits/process step.} \end{aligned} \tag{4}$$

With prior knowledge of the history of the process, the engineer could assess that some of the 20 candidate steps are more likely to be the culprits than others are. The final entropy of the 20 candidate steps would then be lower than what is inferred by equiprobability, increasing the relative entropy of the information extraction in equation (4) and the apparent value of electrical parametric testing.

The value of the analyses provided by functional testers can be assessed in the same manner. If the functional tester indicates that parameters pertaining to steps X_{151} through X_{170} are out of spec, then the root cause of the problem has been localized to these steps, and the Kullback–Leibler formula yields the same value as equation (4) does.

As mentioned previously, the information extraction rate is the more meaningful metric for the value of yield analysis technologies. An engineer who had no access to a parametric tester could, for example, have reduced the possible number of culprit process steps from 500 to 20 by stripping back the wafer to the to the process layers that hosted the problem. However, this

would have taken more than 100 times as long as localizing the problem by a parametric tester, which could evaluate a wafer in about an hour. (In addition, a whole wafer, possibly worth a few thousand dollars, would have to be sacrificed for stripback.) A parametric tester can thus be assigned an information extraction rate of about 5 bits/process step/hour, whereas stripback most likely does not reduce data at a rate faster than 0.05 bits/process step/hour. Most engineers therefore value access to a parametric tester very highly.

Parametric testing is an extremely valuable tool for localizing a problem, but by itself it can rarely be used to pinpoint the source of an electrical fault to the actual process step. However, randomizing and recording the wafer order prior to executing every process step may yield enough information to identify the culprit step precisely and rapidly [28], [29]. Data from parametric testing is correlated to the wafer order at each process step. Any correlation between an electrical parameter and wafer order can potentially infer causality.

Revisiting the case of the out-of-spec threshold voltage of the NMOS transistor with access to wafer position data allows an engineer to reduce the entropy of the data source much further than equation (4) would suggest. The wafer position data would identify a single process step, say threshold implantation, as the culprit of the electrical fault, effectively reducing the data from equiprobability over 500 steps to virtual certainty. Following the same line of reasoning pursued in deriving equation (4) results in an information extraction of 8.96 bits/process step for the combined approach parametric tester/wafer position data approach, a significant improvement over what a parametric tester can do by itself. Analysis of wafer position data also adds little time to the information extraction process, which pegs the information extraction rate of the combined approach at about 8 bits/process step/hour.

Fig. 1 summarizes the effect of information extraction. Given no initial information, all process steps have the same chance of being the source of the fault. Electrical parametric testing concentrates the probability into 20 steps, increasing the information content and reducing the information entropy in the process. Wafer position analysis pinpoints the process step and reduces the information entropy to very low levels.

B. Shortening the Data Cycle

The data analysis activity described in the previous section is only one in a series of steps in the experimentation cycle [14]. The other steps have to be included to get a meaningful benchmark of yield management strategies. Thus the sum of the design time, fabrication time and analysis time is the appropriate denominator for information extraction, which slows down the learning rate over a full VLSI process cycle of about 50 days to (8.96 bits/process step)/(50 days) = 0.18 bits/process step/day. At that learning rate, hundreds of millions of dollars could be lost by the time the source of a problem has been identified. Semiconductor manufacturers have thus resorted to fabricating fractions of the process in parallel on test wafers in the manner suggested in the introduction and in reference [30].

The limitations and the value of short-cycle methods have to be assessed according to how they extract information from a whole gambit of variables (information sources). The expected entropy of multiple variables or sources serves as a metric for the assessment of the limitations. It is given by

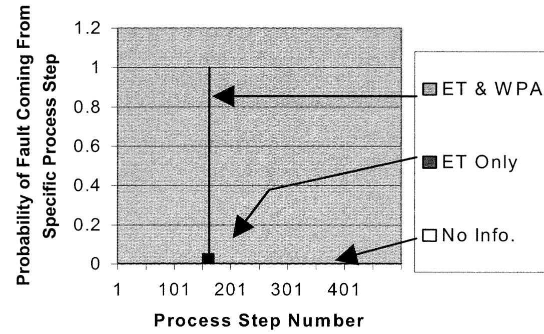


Fig. 1. Probability mass functions of fault sources. Three distributions are shown.

$$\langle H_{P(X_i)}\rangle_j = \cdot - \sum_{j=1}^{n}\sum_{i=1}^{m} P(X_{ij}) \cdot \log_2 P(X_{ij}) \text{ bits/state} \quad (5)$$

where n represents the number of electrical parameters characterize a process. The average information extraction, which is useful in the assessment of the value of short cycle experiments, is given by

$$\langle H_{P(X_i)\|P(Y_i)}\rangle_j = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} P(X_{ij}) \cdot \log_2 \frac{P(X_{ij})}{P(Y_{ij})} \text{ bits/state.} \quad (6)$$

For the purpose of analyzing the limitations and value of short cycle experiments, it is useful to assume that a full VLSI process is broken up into five modules with a fabrication cycle of 10 days each; and that these modules are fabricated in parallel. Let us also assume that 100 electrical parameters (100 sources) characterize this process ($n = 100$). However, short cycle methods cannot capture faults such as plasma damage where a problem near the end of the process affects a structure fabricated near the beginning of the process. We, therefore, assume that 10 out of 100 parameters would remain in a state of equiprobability over 500 process steps. Substituting these assumptions into equation (5) yields

$$\begin{aligned}\langle H_{P(Xi)}\rangle_j &= -10(0.002)\log_2(0.002) - 90(0)\\ &= 0.18 \text{ bits/process step}\end{aligned} \quad (7)$$

which represents the lower limit of entropy that short cycle experiments can achieve by themselves. Experiments that cover the full VLSI process must thus be conducted, in order to guarantee problem localization.

The value of short-cycle experiments can be estimated by substituting the above conditions into equation (6). If 10 out of 100 electrical parameters remain in a state of equiprobability and the other ninety parameters experience the previously calculated information extraction of 8.96 bits/process step, then equation (6) yields

$$H_{\text{POST}\|\text{PRE-SHORT CYCLE}} = \frac{10 \times 0 + 90 \times 8.96}{100} = 8.06 \text{ bits/process step.} \quad (8)$$

Given an experimentation cycle of 10 days this translates into a learning rate of about 0.8 bits/process step/day, a marked improvement over the 0.18 bits/process step/day for the full VLSI process. However, electrically testing wafers that have been realized by the full process compensates for the limitations of short-cycle experiments by capturing faults that short-cycle experiments do not detect. The true information extraction for full-process experiments therefore equals the sum of the output of equation (4) plus the output of equation (7), which, when WPA is included, amounts to

$$\begin{aligned} H_{\text{POST}\|\text{PRE-TEST}} &= 8.96 + 0.18 \\ &= 9.14 \text{ bits per process step.} \end{aligned} \quad (9)$$

Given an experimentation cycle of 50 days for a full VLSI process, this quantity converts to a learning rate of slightly more than 0.18 bits per process step per day, which still compares unfavorably to the learning rate for short-cycle experiments. Short-cycle experiments are therefore considered valuable in spite of their limitations.

C. Automatic Defect Classification, Trainability and False Alarms

The potentially dire consequences of not detecting an electrical fault early during the process have motivated technology managers in the semiconductor industry to introduce an inspection step after about every ten process steps. During these inspections optical imaging and light scattering tools find defects that could cause faults. Most of these tools have the capability to segment their imaging data to separate defects that have been added to the wafers from defects that were detected at previous inspections. The inspection tools also transfer the coordinates of the defects to defect review tools, which enable an engineer to classify the defects and identify their source.

ADC as applied in the semiconductor industry is the process of automatically categorizing wafer defects into one of multiple classes using data captured by wafer analysis instruments. The type of data that is used by the ADC algorithms varies with the application. It could be optical microscope image data, scanning electron microscope (SEM) image data, material composition information (e.g., from SEM energy dispersive spectroscopy), or confocal microscope image data. ADC compares the defect image to a set of images of known defect types and attempts to classify them into previously established categories. These categories are typically associated with process steps through historical data. ADC therefore has an excellent chance of identifying the source of a fault-causing defect [31].

The phrase "excellent chance" implies finite odds of misclassification or classification into a category called unknown. These phenomena complicate the entropy picture by perceivably adding to the entropy of the source during the information extraction process. Fig. 2 illustrates this effect in a hypothetical case where ADC identifies the culprit process step in 73% of all attempts. The other 27% of all attempts either result in misclassification or classification as an unknown. In absence of a well-documented history of the process, misclassified or unclassified defects have an equal chance of being generated in any of the 10 possible process steps. Let us also assume that for the specific case in Fig. 2 ADC points to only three defect sources out of the 10 under consideration. The classifier assigns 40% of all defects to step X_{104}, 20% to step X_{106} and 13% to step X_{107}. The remaining probabilities of 0.027 per process step come from misclassification or classification as unknown. Therefore, $P(X_{104}) = 0.4 + 0.027 = 0.427$; $P(X_{106}) = 0.2 + 0.027 = 0.227$; $P(X_{107}) = 0.13 + 0.027 = 0.157$; and $P(X_{101}) = P(X_{102}) = P(X_{103}) = P(X_{105}) = P(X_{108}) = P(X_{109}) = P(X_{110}) = 0.027$. $P(Y_i) = 0.1$ for all ten process steps in question because initially the odds of the defects being caused by any of the 10 sources are equal. Substituting these data into equation (3) gives the entropy reduction provided by this ADC classifier

$$\begin{aligned} &H_{\text{POST}\|\text{PRE-ADC}} \\ &\quad = 0.427 \cdot \log_2 \tfrac{0.427}{0.1} + 0.227 \cdot \log_2 \tfrac{0.227}{0.1} \\ &\qquad + 0.157 \cdot \log_2 \tfrac{0.157}{0.1} + 7 \times 0.027 \cdot \log_2 \tfrac{0.027}{0.1} \\ &\quad = 0.91 \text{ bits per process step.} \end{aligned} \quad (10)$$

The experimentation cycle of for this process sequence, which consists of 10 process steps, equals one day. Therefore, the learning rate provided by this ADC system is 0.91 bits per process step per day.

Automatic Defect Classification is a "trainable" diagnostic technology: its classification accuracy increases with the cumulative number of observations of a specific defect type. For example, if the ADC classifier made the categorization in (10) after viewing 1000 wafers of a certain type, then it is likely to make more accurate classifications after having viewed 2000 wafers. The percentage of defects that have been misclassified or classified as unknown may shrink from 27% to 7%. Under these circumstances the probability of defect localization may concentrate to $P(X_{104}) = 0.5 + 0.007 = 0.507$; $P(X_{106}) = 0.25 + 0.007 = 0.257$; $P(X_{107}) = 0.18 + 0.007 = 0.187$; and $P(X_{101}) = P(X_{102}) = P(X_{103}) = P(X_{105}) = P(X_{108}) = P(X_{109}) = P(X_{110}) = 0.007$. The relative entropy of this concentration is given by

$$\begin{aligned} &H_{2000\cdot\|\cdot 1000\cdot\cdot\text{Wafers}} \\ &\quad = 0.507 \cdot \log_2 \tfrac{0.507}{0.427} + 0.257 \cdot \log_2 \tfrac{0.257}{0.227} \\ &\qquad + 0.187 \cdot \log_2 \tfrac{0.187}{0.157} + 7 \times 0.007 \cdot \log_2 \tfrac{0.007}{0.027} \\ &\quad = 0.123 \text{ bits per process step.} \end{aligned} \quad (11)$$

From this quantity we can infer a training rate of 0.123 bits per process step per 1000 wafers, which is likely to decrease as classifications become increasingly accurate.

ADC can generate false alarms by spotting defects that do not cause electrical faults. If the semiconductor manufacturer deems it desirable to identify the source of these false or "cosmetic" defects then the analysis expressed in equation (10) holds. However, if the manufacturer only wants to identify the source of "killer" defects, then an additional bucket for false defects needs to be created. If we assume that 25% of the defects in the above analysis are false in all categories, then $P(X_{104}) = 0.75*(0.4+0.027) = 0.320$; $P(X_{106}) = 0.75*(0.2+0.027) = 0.170$; $P(X_{107}) = 0.75*(0.13+0.027) = 0.118$; and $P(X_{101}) = P(X_{102}) = P(X_{103}) = P(X_{105}) = P(X_{108}) = P(X_{109}) = P(X_{110}) = 0.75*0.027 = 0.020$; $P(\text{false}) = 0.25$. Under

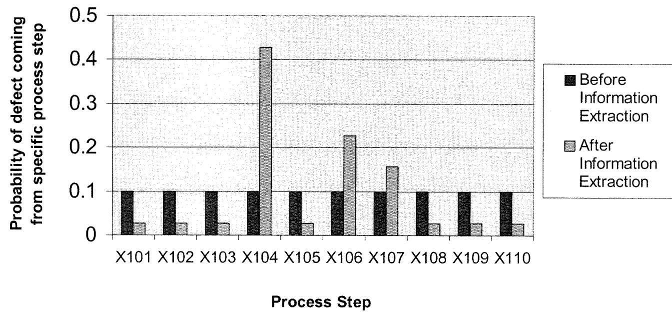


Fig. 2. Probability mass functions of an automatic defect classification.

these circumstances, the Kullback–Leibler formula yields an information extraction of

$$\begin{aligned} H_{\text{POST}\|\text{PRE-ADC}} &= 0.320 \cdot \log_2 \tfrac{0.320}{0.075} + 0.170 \cdot \log_2 \tfrac{0.170}{0.075} + 0.118 \cdot \log_2 \tfrac{0.118}{0.075} \\ &\quad + 7 \times 0.020 \cdot \log_2 \tfrac{0.020}{0.075} + 0.250 \cdot \log_2 \tfrac{0.250}{0.250} \\ &= 0.68 \text{ bits per process step} \end{aligned} \tag{12}$$

which translates into a learning rate of 0.68 bits per process step per day. The information extraction and its associated learning rate can be increased to the value given by (10) by reducing the proportion of false counts to zero.

D. Spatial Signature Analysis

A spatial signature is defined as a population of defects that originates from a single manufacturing problem. Spatial signature analysis is an artificial intelligence method that relies on capturing operator experience through a teaching method to emulate the human response to various manufacturing situations. This has been successfully accomplished through the development and application of an image processing-based, fuzzy classifier system. The technique uses data collected from current in-line inspection tools to interpret and rapidly identify characteristic patterns, or "signatures," that are uniquely associated with the manufacturing process. The SSA system then alerts fabrication engineers to probable yield-limiting conditions that require attention, and uniquely assigns a signature to a single process step even when multiple signatures overlap on a wafer map [16], [32]–[35]. (See Fig. 3.)

Currently SSA does not by itself routinely assign the process step attributable to a defect, but a yield engineer with knowledge of machine history and prior defect patterns can localize the defect with sufficient accuracy. Given a sufficiently large database library of historical SSA images and a database link between the stored images and the process step that generated them, there is in principle no reason why SSA could not make the attribution to the process step automatically. Since SSA, like ADC, is a "trainable" diagnostic technology, the accuracy of the attribution would increases with the number of images in the library. The value of SSA is thus likely to increase over time.

We can extend the argument of information entropy reduction to the SSA approach as follows. We assume 100 defects have been detected and that there are 10 possible sources of these defects: Steps X_{201} through X_{210}. Before applying SSA, the probability $P(Y_i)$ of determining the source of any defect is 0.1. Let

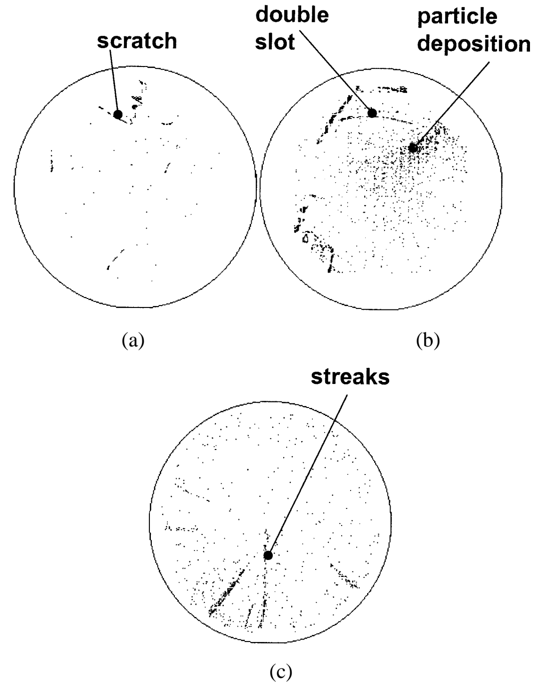


Fig. 3. Various types of signatures found by SSA. Fig. 3(a)–(c) serve as examples of systematic clusters on a series of wafer maps. Each cluster is made up of many individual defects that are correlated to each other based on the manufacturing source [33].

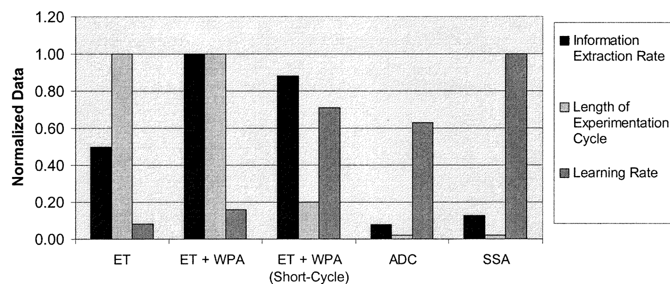
us also assume that SSA separates the 92 of these defects into 3 large clusters such as the ones shown in Fig. 3, and only leaves behind 8 isolated defects to classify. Thirty-five of the one-hundred defects are assigned to step X_{203}; fifteen to X_{205}; and forty-two to step X_{208}. The remaining eight unclassified defects could come from any step between and including X_{201} through X_{210}, adding a 0.008 to the odds of a defect coming from each of the 10 states. Thus $P(X_{201}) = P(X_{202}) = P(X_{204}) = P(X_{206}) = P(X_{207}) = P(X_{209}) = P(X_{210}) = 0.008$; $P(X_{203}) = (0.35 + 0.008) = 0.358$; $P(X_{205}) = (0.15 + 0.008) = 0.158$; and $P(X_{208}) = (0.42 + 0.008) = 0.428$. Then the relative entropy of this information extraction is given by

$$\begin{aligned} H_{\text{POST}\|\text{PRE-SSA}} &= 0.158 \cdot \log_2 \tfrac{0.158}{0.1} + 0.358 \cdot \log_2 \tfrac{0.358}{0.1} \\ &\quad + 0.428 \cdot \log_2 \tfrac{0.428}{0.1} + 7 \times 0.008 \cdot \log_2 \tfrac{0.008}{0.1} \\ &= 1.45 \text{ bits per process step} \end{aligned} \tag{13}$$

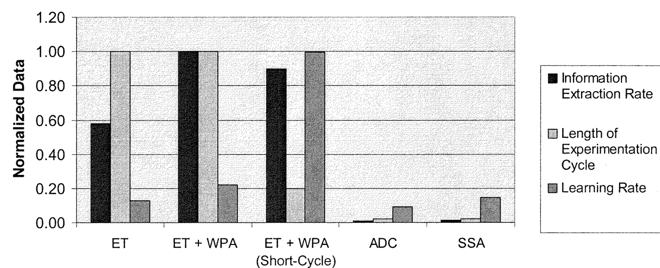
which corresponds to a learning rate of 1.45 bits per process step per day.

V. Summary and Discussion

Technology managers in the semiconductor industry need strategies for yield management and fault reduction. Accelerated yield learning gives competitive advantage in R&D environment, which is characterized by radical experience curves. Capital productivity generates competitive advantage in the production environment, where an undetected source of electrical faults can cause enormous losses. The authors have identified the yield-learning rate as the key success metric for both environments, and recognized it as a benchmark for the



(a)



(b)

Fig. 4. Comparative information extraction rates, lengths of experimentation cycles and yield learning rates. Data are normalized. (a) Models a mature manufacturing environment. (b) Models an early manufacturing environment.

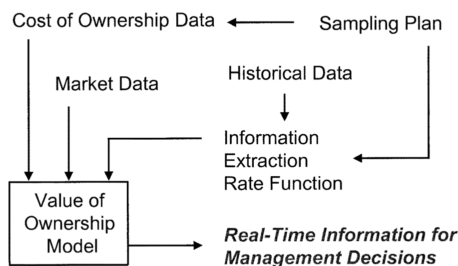


Fig. 5. Inputs and output of a real-time value of ownership model.

valuation of (sometimes unrelated) technologies. The authors defined the yield-learning rate as the ratio between the information extraction rate and the length of the experimentation cycle. The information extraction has also been quantified by using the Kullback–Leibler formula [equation (3)] for relative information entropy, which enables estimation of the relative value of four technologies—electrical testing, wafer position data, automatic defect classification, and spatial signature analysis.

Technology managers in the semiconductor industry are not confronted with an either/or decision between these diagnostic technologies, because the aforementioned limitations of short-cycle analysis practices, which extend to ADC and SSA, mandate diagnoses of wafers that have been exposed to the full ULSI process. Electrical testing on some level is a must, because it is the only technology that identifies sophisticated faults that have roots in more than one process module. However, utilizing diagnostic technologies that can identify sources of faults on short notice is also extremely important, because it may help semiconductor manufacturers avoid extremely high losses. Managers would thus like to know the optimal amount of resources to invest in each of the aforementioned technologies. They need a Value-of-Ownership model.

Fig. 4 shows the results of a valuation effort for the aforementioned yield analysis technologies, which is based on their relative learning rates and includes their respective limitations. Fig. 4(a) models a mature production environment, for which we assume that short cycle experiments capture 90% of all faults, whereas ADC and SSA only capture 80% of all faults. Analysis of the valuations indicates that electrical testing exhibits a very high information extraction rate, especially when it is combined with wafer position analysis. The value of electrical testing plus WPA as expressed by the learning rate is increased dramatically when these methods are used in conjunction to analyze wafers that have been fabricated on short cycle, even though a complete characterization of a full semiconductor process is no longer possible under these conditions. Inspection-related technologies like ADC and SSA have an even higher value due to their association with extremely short experimentation cycles. Fig. 4(b) shows that this picture changes dramatically if we model an early manufacturing environment, for which we assume that ADC and SSA localize only 10% of all faults because hitherto unobserved defect types have not been incorporated into defect libraries. In this case, the amount of entropy reduction for ADC and SSA is limited, the residual entropy is very high, and the expected credit for problem localization has to be given to techniques such as electrical testing and WPA, which do not depend on libraries of previously observed faults for their analysis.

Fig. 5 shows that data on cost of ownership must feed into any operationalizable Value-of-Ownership model. Cost-of-Ownership models give managers a very good idea of the true cost of yield analysis technologies, which unfortunately varies directly with value. Estimates place the cost of ownership of defect inspection and review with automatic defect classification at around $10 per wafer per inspection [36]. The cost of ownership of electrical parametric testing including the associated data analysis varies significantly with sampling plan, but most experts believe it is nearly an order of magnitude less than that of defect inspection. The high cost test wafers limits the utility of short-cycle experiments. They are primarily conducted in R&D, but yield analysis experts have been known to utilize short-cycle experiments on an *ad hoc* basis to localize the root cause of yield problems in manufacturing [30].

The information extraction rate also depends strongly on the (user-specific) sampling plan, because a larger sample size generates lower uncertainty, higher information content and lower information entropy. In addition, managers typically have ample historical data on the relative frequency of fault types that have plagued their factories. Matching this knowledge to the value metrics established in this paper will allow technology mangers to significantly improve upon the current level of cost/value propositions for yield management and fault reduction, and the technologies described in this paper may enable them to achieve a value assessment in real time. For example, it is quite conceivable that an ADC system can estimate the damage done by a distribution of defects once it has localized the source of the defect type. The shop floor control system knows when the defect distribution was generated and how many wafers have passed through the culpable step since the appearance the defect distribution. The shop floor control system can also absorb economic data like the spot market price of the product chips. It is

thus quite conceivable that one day the shop floor control system will send a technology manager the following message.

"The inspection at step 256 has identified a defect distribution that results from step 248. Machine #83 is the culprit. You have already lost \$130k $\pm$ \$20k. You are continuing to lose money at a rate of \$25k $\pm$ \$4.2k per hour. If you shut Machine #83 down, you will only lose \$11k $\pm$ \$3.3k per hour until the machine is back up. Your track record indicates that you would have lost an additional \$840k $\pm$ \$90k, had you attempted to solve this problem without me. Don't you think my value significantly exceeds my cost?"

REFERENCES

[1] R. Leckie, "A model for analyzing test capacity, cost and productivity," in *IEEE Int. Test Conf.*, 1986, pp. 213–218.

[2] R. Martinez, V. Czitrom, N. Pierce, and S. Srodes, "A methodology for optimizing cost of ownership," in *Proc. SPIE*, vol. 1803, 1992, pp. 363–387.

[3] J. Secrest and P. Burggraaf, "The reasoning behind cost of ownership," *Semiconduct. Int.*, May 1993.

[4] D. Dance and D. Jimenez, "Applications of cost of ownership," *Semiconduct. Int.*, pp. 6–7, Sept. 1994.

[5] L. Chao, D. Dance, and T. DiFloria, "Get a handle on your cost of test," *Test and Measurement World*, pp. 45–50, Apr. 1995.

[6] E. Wang, M. Holtan, R. Akella, I. Emami, M. McIntyre, D. Jensen, and D. Fletcher, "Valuation of yield management investments," in *Proc. IEEE/SEMI/ASMC*, Cambridge, MA, 1997, pp. 1–7.

[7] P. Silverman, "Capital productivity: Major challenge for the semiconductor industry," *Solid State Technol.*, vol. 37, no. 3, p. 104, March 1994.

[8] C. Weber, D. Jensen, and E. D. Hirleman, "What drives defect detection technology?," *Micro*, pp. 51–72, June 1998.

[9] C. Stapper and R. Rosner, "Integrated circuit yield management and yield analysis: Development and implementation," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, pp. 95–101, May 1995.

[10] C. Weber, B. Moslehi, and M. Dutta, "An integrated framework for yield management and defect/fault reduction," *IEEE Trans. Semiconduct. Manufact.*, vol. 8, no. 2, pp. 110–120, May 1995.

[11] M. Iansiti and J. West, "Technology integration," *Harvard Bus. Rev.*, pp. 69–79, May–June 1997.

[12] D. Jensen, C. Gross, and D. Mehta, "New industry document explores defect reduction technology challenges," *Micro*, pp. 35–44, Jan. 1998.

[13] C. Weber, "Accelerating three-dimensional experience curves in integrated circuit process development," Management of Technology Master's thesis, MIT, May 1996.

[14] ——, "Metrology-based control and profitability in the semiconductor industry," in *Proc. SPIE Conf. Metrology-Based Control in Micro-Manufacturing*, San Jose, CA, Jan. 25, 2001, pp. 8–20.

[15] S. H. Thomke, "Managing experimentation in the design of new products," *Manage. Sci.*, vol. 44, no. 6, pp. 743–762, June 1998.

[16] V. Sankaran, C. Weber, K. W. Tobin, and F. Lakhani, "Inspection in semiconductor manufacturing," in *Webster's Encyclopedia of Electrical and Electronics Engineering*. New York: Wiley, 1999, vol. 10, Pattern Analysis and Machine Intelligence, pp. 242–262.

[17] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. of Illinois Press, 1949.

[18] N. Abramson, *Information Theory and Coding*. New York: McGraw-Hill, 1963.

[19] P. Beckmann, *Probability in Communication Engineering*. New York: Harcourt, Brace & World, Inc., 1967.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[21] R. V. L. Hartley, "Transmission of Information," *Bell Syst. Tech. J.*, vol. 31, pp. 751–763, 1928.

[22] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968.

[23] C. Weber, "Knowledge transfer and the limits to profitability: An empirical study of problem-solving practices in semiconductor manufacturing and process development," *IEEE Trans. Semiconduct. Manufact.*, vol. 15, pp. 420–426, Nov. 2002.

[24] Semiconductor Industry Association, *Nat. Technol. Roadmap for Semiconductors*, 1997.

[25] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, "Expert and novice performance in solving physics problems," *Sci.*, vol. 208, pp. 1335–1342, 1980.

[26] E. Von Hippel and M. Tyre, "How leaning by doing is done: Problem identification in novel process equipment," *Res. Policy*, vol. 24, pp. 1–12, 1995.

[27] C. Alcorn, D. Dworak, N. Haddad, W. Henley, and P. Nixon, "Kerf test structure designs for process and device characterization," *Solid State Technol.*, pp. 229–235, May 1985.

[28] G. Scher, D. Eaton, J. Sorensen, B. Fernelius, and J. Akers, "In-line statistical process control and feedback for VLSI integrated circuit manufacturing," *IEEE Trans. Comp., Hybrids Manufact. Technol.*, vol. 13, no. 3, pp. 484–489, Sept. 1990.

[29] G. Scher, "Wafer tracking comes of age," *Semiconduct. Int.*, pp. 126–131, May 1991.

[30] C. Weber, "A standardized method for CMOS unit process development," *IEEE Trans. Semiconduct. Manufact.*, vol. 5, pp. 94–100, May 1992.

[31] M. H. Bennett, K. W. Tobin, and S. S. Gleason, "Automatic defect classification: Status and industry trends," in *Proc. SPIE Metrology, Inspection, and Process Control for Microlithography IX*, vol. 2439, San Jose, CA, March 1995, p. 210.

[32] K. W. Tobin, S. S. Gleason, F. Lakhani, and M. H. Bennett, "Automated analysis for rapid defect sourcing and yield learning," in *Future Fab International*. London, U.K.: Technology Publishing Ltd., 1997, vol. 1, p. 313.

[33] S. S. Gleason, K. W. Tobin, and T. P. Karnowski, "An integrated spatial signature analysis and automatic defect classification system," in 191st Meeting Electrochemical Society, Inc., May 1997.

[34] K. W. Tobin, S. S. Gleason, T. P. Karnowski, and M. H. Bennett, "An image paradigm for semiconductor defect data reduction," in SPIE's 1996 Int. Symp. Microlithography, Santa Clara Convention Center, Santa Clara, CA, March 10–15, 1996.

[35] K. W. Tobin, T. P. Karnowski, and F. Lakhani, "Integrated applications of inspection data in the semiconductor manufacturing environment," in *Proc. SPIE Conf. Metrology-Based Control in Micro-Manufacturing*, San Jose, CA, Jan. 25, 2001, pp. 31–40.

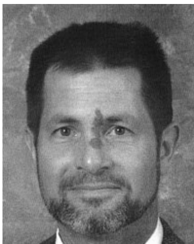
[36] A. Shapiro, T. James, and B. Trafas, "Advanced inspection for 0.25-μm-generation semiconductor manufacturing," in *Proc. SPIE's 22nd Annu. Int. Symp. Microlithography*, Santa Clara, CA, 1997, pp. 445–451.

Charles Weber received the A.A. degree in physical science from the American College of Switzerland in 1975; the B.S. degree in engineering physics from the University of Colorado, Boulder, in 1978; the M.S. degree in electrical engineering from the University of California, Davis, in 1981; and the S.M. degree in Management of Technology from the Massachusetts Institute of Technology, Cambridge, in 1996.

He joined Hewlett-Packard Company as a process engineer in an IC manufacturing facility. He subsequently transferred to HP's IC process development center, working in electron beam lithography, parametric testing, microelectronic test structures, clean room layout and yield management. From 1996 to 1998, he managed the defect detection project at SEMATECH, as an HP assignee. He is currently a doctoral candidate in Management of Technology at MIT's Sloan School of Management, and he has accepted a position as Assistant Professor of Engineering and Technology Management at Portland State University, Portland, OR.

Vijay Sankaran received the M.S. degree in computer-aided design from Southern Illinois University, Edwardsville, IL, and the Ph.D. degree in intelligent inspection systems for semiconductor manufacturing from Rensselaer Polytechnic Institute, Troy, NY, in 1996.

From 1996 to 1999, he was the technical lead for advanced inspection technologies at SEMATECH. In 1999, he co-chaired the international technology roadmap development for defect reduction and served as Chair of the Consortium for Metrology of Semiconductor Nanodefects, which is based at Arizona State University. Since 2000, he has founded two consecutive software startups is in architecture for collaborative process management, for which he acted as technical executive.



Kenneth W. Tobin, Jr. (M'99) received the B.S. degree in physics, the M.S. in nuclear engineering, both from Virginia Tech, Blacksburg, VA, in 1983 and 1984, respectively, and the Ph.D. degree in nuclear engineering from the University of Virginia, Charlottesville, VA, in 1987.

He leads the Image Science and Machine Vision Group at the Oak Ridge National Laboratory. He has authored and co-authored over ninety publications and he currently holds four U.S. Patents in the areas of computer vision and photonics.

Dr. Tobin is a Fellow of the International Society for Optical Engineering (SPIE) where he is currently Chairman of the Conference on Process Characterization and Diagnostics in IC Manufacturing. He was the first invited U.S. organizer of the International Conference on Quality Control by Artificial Vision. He was the Tennessee Academy of Science's Industrial Scientist of the Year in 2001.

Gary Scher received the S.B. and S.M. degrees in physics from the Massachusetts Institute of Technology, Cambridge, both in 1977, and the M.A. degree in Russian studies and the Ph.D. degree in physics, both from Harvard University, Cambridge, MA, in 1980 and 1983, respectively.

He is President of Sleuthworks, Inc., a company he co-founded in 1992 to provide implementation of automated wafer-level tracking and data analysis in the semiconductor industry. The work that led to the formation of Sleuthworks was begun in 1983 as a process and yield improvement engineer at Hewlett-Packard Company and continued as manager of the Wafer Sleuth project at SEMATECH from 1989–1991. Since then Sleuthworks has completed installation of Wafer Sleuth in 25 fabs worldwide. He is the author of several papers on electronic behavior in disordered materials and articles related to yield improvement in semiconductor manufacturing.