1993

Accurate Computation of Field Reject Ratio Based on Fault Latency

Dharamvir Das
Cadence Design Systems, Noida, India

Sharad C. Seth
University of Nebraska-Lincoln, seth@cse.unl.edu

Vishwani D. Agrawal
AT&T Bell Laboratories. Murray Hill, NJ

Accurate Computation of Field Reject Ratio Based on Fault Latency

Dharamvir Das, Sharad C. Seth, *Senior Member, IEEE*, and Vishwani D. Agrawal, *Fellow, IEEE*

Abstract—**The field reject ratio, the fraction of defective devices that pass the acceptance test, is a measure of the quality of the tested product. Although the assessment of quality is important, an accurate measurement of the field reject ratio of tested VLSI chips is often not feasible. We show that the known methods of field reject ratio prediction are not accurate since they fail to realistically model the process of testing. We model the detection of a fault by an input test vector as a random event. However, we recognize that the detection of a fault may be delayed for various reasons: the fault may be detectable only by application of a sequence of vectors or it may not have been targeted until later. In our statistical model, a fault is characterized by two parameters: a per-vector detection probability and an integer-valued latency. Irrespective of the detection probability, the fault cannot be detected by a vector sequence shorter than its latency. The circuit is characterized by the joint distribution of latency and detection probability over all faults. This distribution, obtained by applying the Bayes' rule to the actual test data, enables us to compute the field reject ratio. The sensitivity of this approach to variations in the measured parameters is also investigated.**

I. Introduction

FOR VLSI devices, the *field reject ratio* (or *reject ratio* for short) is defined as the ratio of faulty chips among the chips passed by the tests. Thus, a reject ratio of 0.001 means that the average number of faulty chips after testing is one in a thousand. *Average outgoing quality* (AOQ) and *defect level* are equivalent terms used in the industry for the field reject ratio, represented in parts per million (ppm). A large number of chips must be in use in the field before an adequate amount of field return data can be obtained to estimate the reject ratio. Direct measurement of reject ratio, therefore, is difficult and can be quite expensive. Researchers have proposed several indirect methods.

It is clear that if the chip fabrication process is perfect, there will be no defective parts, and hence the reject ratio will be zero. Or, if the testing process is perfect, no defective parts will escape tests, and again, the reject ratio will be zero. These considerations suggest that any model for reject ratio computation must take into account parameters characterizing the processing line and chip testability.

Manuscript received December 4, 1992; revised May 24, 1993.

D. Das is with Cadence Design Systems, Noida, 201305, India.

S. C. Seth is with the Department of Computer Science and Engineering, University of Nebraska, Lincoln, NE 68588.

V. D. Agrawal is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 9213097.

Previous attempts [1]–[6] have derived relations between fault coverage and product quality. For an acceptable reject ratio, e.g., 1 in 10 000, the models in the cited works compute fault coverage requirements close to 100%, which is difficult to obtain in large highly sequential circuits. Difficulties are caused by the presence of redundant faults and the ways a fault simulator models faults, circuit initialization, race conditions, etc.

A method that computes reject ratio in the absence of fault simulator data was developed by Seth and Agrawal [6]. They define a per-vector probability of failure detection for the circuit. Implicit in this model are the probability of occurrence of a fault and the probability of detection given the fault has occurred. Note that the detection probability of a fault is a conditional probability. That is, it is the probability of detecting the fault given that the fault is present in the circuit. Hence, the product of the two probabilities is the *failure detection probability* of the chip by a vector. In other words, a realistic coverage requirement can be obtained by weighting the detection probabilities of faults with their occurrence probabilities. In this method, only wafer test data are needed to determine the reject ratio. The entire process of testing is characterized by a detection probability density function which is determined using the measured fraction of failing chips versus the number of vectors.

II. Motivation

We derive the motivation for the present work from the shortcomings of the available methods of estimating reject ratio, observed while evaluating them on experimental data [7]. These methods are reviewed in the next two subsections, followed by their experimental evaluation.

A. Fault Coverage Based Methods

Methods to compute reject ratio based on fault coverage data often require that the yield be known at least approximately. Consider M single stuck faults that can occur on a chip. Of these, m are covered by the given test vectors. The fault coverage f is then m/M. Suppose K faults are present on the chip being tested. Let $q_k(K)$ be the probability of detecting exactly k faults given that K faults have occurred. Then, $q_k(K)$ has a hypergeometric

1063-8210/93$03.00 © 1993 IEEE

doi: 0.1109/92.250201

density function and is given by [8]

$$q_k(K) = \frac{\binom{K}{k}\binom{M-K}{m-k}}{\binom{M}{m}}. \quad (1)$$

The probability of passing the chip having K faults as good is given by (1) when $k = 0$, i.e.,

$$q_0(K) = \frac{\binom{M-K}{m}}{\binom{M}{m}} \simeq (1-f)^K. \quad (2)$$

This approximation is quite accurate for $K << \sqrt{M(1-f)/f}$. Since K is a random variable, given a distribution for K, the expectation of $q_0(K)$ gives the measured yield, i.e., the sum of true yield y and the fraction of defective chips tested as good. If the true yield is known, the reject ratio is computed by the following formula:

$$r = \frac{E(q_0(K)) - y}{E(q_0(K))} \quad (3)$$

where $E(q_0(K))$ is the expectation of $q_0(K)$ and is given by

$$E(q_0(K)) = \sum_{K=0}^{M} (1-f)^K p(K). \quad (4)$$

In the above expression, $p(K)$ is the density function of K. Results for different probability density function $p(K)$ of K are summarized in Table I. The geometric density model was introduced by Wadsack [1], and the binomial density was used by Williams and Brown [3]. The generalization of these two models based on the gamma density function was also given by Wadsack [2]. In Table I, we have shown the geometric and binomial density cases because these are most frequently quoted in the literature. The key feature of the remaining two models [4], [5] in Table I is that they assume the faults to be clustered, as is generally believed to be the case for VLSI chips. Notice that except for compound density, all others require the true yield as a parameter. In the case of compound density, parameters A, a, b, and c together determine the yield.

B. Chip Failure Probability (CFP) Method

The methods tabulated above use the fault coverage f as a parameter. Another method that does not depend on the fault coverage of test vectors was proposed by Seth and Agrawal [6]. Their method relies on the fact that detection probability of a fault is really a conditional probability. It is the probability of detection by an input vector given a fault is present. Associated with each fault is its probability of occurrence on the chip. Just as all faults are not equally detectable, they also do not occur with equal probability. The product of these two probabilities is the absolute detection probability of a chip by a test vector. This is equivalent to obtaining fault coverage where faults are weighted by their occurrence probabilities. For example, a fault that never occurs will have zero weight and is not required to be covered. We include the following details of the CFP method since they are relevant to the new method given in Section III. The analysis presented in this and the remaining sections uses the following notation:

- C total number of chips tested
- N total number of test vectors applied
- C_i number of chips that fail at vector i
- y true yield, i.e., fraction of good chips
- y_n estimated yield of chips after application of n vectors.

Each failing chip has associated with it a random variable x which is the probability of a fault occurring on the chip and being detected by a test vector. If $x = 0$, then the chip has no defects. Since x is a probability, its value lies in the range 0–1. Let $F(x)$ represent the density of the chips. Then, $F(x)\Delta x$ is the fraction of chips in which faults have occurred, and are detected with detection probability between x and $x + \Delta x$. Chips having faults with detection probability 0 are essentially good chips. Therefore, $F(0) = y$. It is easy to verify that

$$\int_0^1 F(x)\,dx = 1. \quad (5)$$

Let $\pi(x)$ represent the distribution of defective chips. Then,

$$F(x) = y\delta(x) + \pi(x) \quad (6)$$

where $\delta(x)$ is the Kronecker delta function. Suppose a fault has occurred on a chip, and that n test vectors have been applied. Since x is the probability of the fault being detected by a test vector, the probability that the chip has not failed after the application of n test vectors is $(1-x)^n$. The expectation of this probability with the distribution $F(x)$ gives the yield of chips after the nth vector, i.e.,

$$y_n = \int_0^1 (1-x)^n F(x)\,dx$$

$$= y + \int_0^1 (1-x)^n \pi(x)\,dx. \quad (7)$$

The density function $F(x)$ is estimated from chip failure detection data obtained by testing a sample of C chips with a test sequence of N vectors. As these vectors are applied, the number of chips that fail for the first time at each vector is recorded. For vector number i, let C_i denote the number of such chips. The probability that a chip fails at the ith vector is $x(1-x)^{i-1}$. Thus, C_i/C is used as a weight for this probability in determining the distribution function $F(x)$. If a uniform *a priori* distribution is as-

TABLE I
REJECT RATIOS FOR VARIOUS DISTRIBUTIONS OF K

Density Function $p(K)$ = Prob.(number of faults = K)	Reject Ratio
Geometric [1]: $p(K) = y(1-y)^n$ where y = true yield and f = fault coverage.	$(1-f)(1-y), \quad K = 0, 1, 2 \cdots$
Binomial [3]: $p(K) = \binom{M}{K}(1-P_M)^{M-K}P_M^K, \quad K = 0, 1, 2 \cdots$ $P_M = 1 - y^{1/M}$ where M = total number of faults that can occur.	$1 - y^{(1-f)}$
Modified Poisson [4]: $p(K) = (1-y)\dfrac{(K_0-1)^{K-1}}{(K-1)!}e^{-(K_0-1)} \quad K = 1, 2, \cdots$ $p(0) = y$ where K_0 = average number of faults on a faulty chip.	$\dfrac{(1-f)(1-y)e^{-(K_0-1)f}}{y + (1+f)(1-y)e^{-(K_0-1)f}}$
Compound [5]: $p(K) = \sum_{x=0}^{\infty} p_2(K\|x)p_1(x)$ where $p_1(x)$ = Prob.(number of defects = x) $= \binom{x+a-1}{x}(Ab)^x(1+Ab)^{-x-a}$ A = chip area b = defect density a = clustering parameter $p_2(K\|x)$ = Prob.(number of faults = $K\|x$ defects) $= \dfrac{(cx)^K}{K!}e^{-cx}, \quad K = 0, 1, 2 \cdots$ c = average number of faults per defect	$1 - \left\lvert \dfrac{1+Ab(1-e^{-c})}{1+Ab(1-e^{-cf})} \right\rvert^{-a}$

sumed for $F(x)$, then the weight is modified to reflect the Bayes' estimation [9]. Also, another component of $F(x)$ is the probability that a chip does not fail after the application of N vectors and is given by $(1-x)^N$. The weight for this is $(1 - y - (1/C)\sum_{i=1}^{N} C_i)(N+1)$. Hence, the distribution of chips over the random variable x is determined to be

$$F(x) = y\delta(x) + \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)(N+1)(1-x)^N + \frac{1}{C}\sum_{i=1}^{N} C_i i(i+1)x(1-x)^{i-1}. \quad (8)$$

Substituting for $F(x)$ in (7), we get

$$y_n = y + \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)\left(\frac{N+1}{N+n+1}\right) + \frac{1}{C}\sum_{i=1}^{N} C_i\left(\frac{i(i+1)}{(n+i)(n+i+1)}\right). \quad (9)$$

In the above expression, the true yield y is still an unknown parameter. It is evaluated by equating the estimated yield y_N after N vectors to the measured yield. The measured yield is $1 - (\sum_{i=1}^{N} C_i)/C$ which, when equated to y_N obtained from (9), gives the following solution for y:

$$y = 1 - \frac{1}{C}\sum_{i=1}^{N} C_i - \left(\frac{2N+1}{N}\right)\left(\frac{1}{C}\right)\sum_{i=1}^{N} C_i \cdot \frac{i(i+1)}{(N+i)(N+i+1)}. \quad (10)$$

The reject ratio is now computed from (9) and (10) as

$$r = \frac{y_N - y}{y_N}. \quad (11)$$

Next, we will illustrate the application of the above analysis to experimental data.

C. Experimental Data

The CFP method also allows us to estimate the true yield. Wafer test data for a CMOS chip obtained from Delco Electronics were used to compare the reject ratio

TABLE II
ESTIMATED REJECT RATIOS FOR A CMOS DEVICE

Geometric	0.00087
Binomial	0.00103
Modified Poisson	0.00048
Compound	0.00064
CFP	0.00532

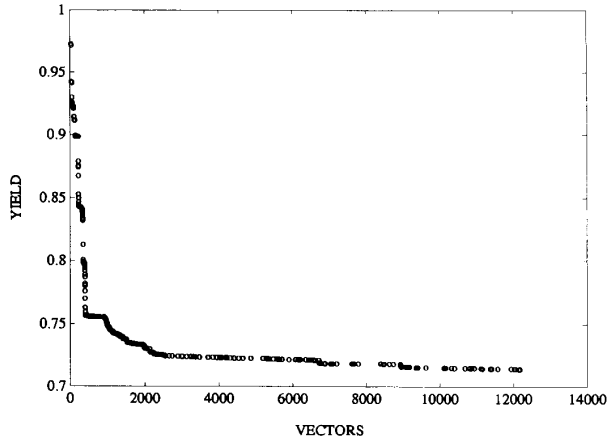

Fig. 1. Experimental data on yield for a chip.

values computed from the formulas given in the previous sections [7]. The test data obtained were for 79 912 devices. Of these, 847 failed the parametric test and 7699 failed the continuity test. Thus, functional testing was done on 64 366 devices. The test consisted of 12 188 test vectors, and had a stuck fault coverage of 99.7% as measured by a fault simulator. True yield computed by the CFP method (from (10), which involves only the chip failure data) is 0.7092. The reject ratios computed by different methods are given in Table II.

The estimated reject ratios lie between 480 and 5320 parts per million. This wide variation is undesirable, and leads us to suspect the assumptions made in deriving the results in Sections II-A and II-B. The methods of Section II-A rely on fault coverage and the density function of the number of faults. The method of Seth and Agrawal [6] does not rely on the fault coverage information.

Our recent investigation suggests that the accuracy of the Seth and Agrawal method could be improved by considering possible latencies in the detection of faults on a chip by the chip test [10]. The latencies can arise due to a variety of causes (e.g., sequential nature of the circuit or functionally partitioned nature of long test sequences), but here we are concerned primarily with their effect. Fig. 1 shows a plot of the yield of chips as a function of test length (measured in number of vectors). A careful look at the plot reveals discontinuities or sudden *jumps* in the yield at several points (a similar phenomenon is also seen in the fault coverage versus vectors graphs). We believe that such discontinuities are an essential part of any chip test data, and that they occur because of clustering of fault latencies. In the following section, we propose a latency model capable of providing an accurate fit to the experimental data.

III. CURVE FITTING TO EXPERIMENTAL DATA

As mentioned above, there may be many plausible explanations for the observed phenomenon. For example, in a sequential circuit, the fault activation may require the control of several flip-flops. Depending on the levels (sequential depth) of the flip-flops in the circuit, a sequence of test vectors will be needed. Further, the fault effect may have to be propagated through several levels of flip-flops, again requiring another sequence of vectors. Jumps can also occur in combinational circuits if the vectors are specifically generated to test different parts of the circuits.

The idea of fitting a model to experimental data is to eliminate random variations. However, the delayed detection of faults due to the sequential nature of the circuit is not entirely a random phenomenon; the stepped increase in fault coverage is real and not random. Any attempt at fitting a smoothly rising curve will therefore lead to erroneous results. The method described in the following sections is such that it tracks the experimental data.

Let us associate with each fault an integer called *latency* or the number of test vectors that must be applied before the fault is considered detectable by the subsequent vectors. The latency is zero for all detectable stuck type faults in a combinational circuit. The latency of some faults in a sequential circuit may also be zero; such faults may be called *combinational*, while the faults with nonzero latency will be called *sequential*. It is the discrete nature of the integer-valued latency random variable that will give rise to jumps in the observed value of yield as a function of vector number.

A. Chip Failure Analysis

We will assume that chip failure detection on an applied test vector is a random event. For a chip with a fault, we can speak of the following random variables.

1) d: A random variable representing the latency of a fault. It takes values in the set $\{0, 1, 2, \cdots, \infty\}$. A fault with latency d can only be detected by vectors $d + 1$, $d + 2$, etc.

2) x: A random variable representing the detection probability of a fault of latency d. This is the probability that a fault with latency d has occurred and is detected by the vector, $0 \leq x \leq 1$.

3) $g_n(x, d)$: A function of two random variables. This represents the probability that a chip fails at the nth vector.

$$g_n(x, d) = x(1 - x)^{n-d-1} I_{\{d+1, \cdots, \infty\}}(n) \qquad (12)$$

where $I_{\{d+1, \cdots, \infty\}}(n)$ is the indicator function[1] [9]. The

[1]Indicator function: Let Ω be any space with points ω and A any subset of Ω. The indicator function of A is defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

expression indicates that a chip with a fault of latency d cannot fail for the first d vectors. Thereafter, its failure detection probability is determined by the per-vector detection probability x.

Yield Model: Let us define a density function $F(x, d)$, such that $F(x, d)\Delta x$ is the fraction of chips that have latency d and per-vector detection probability between x and $x + \Delta x$. If y is the yield, then only a fraction $1 - y$ of the total chips can fail. Hence, we can write

$$F(x, d) = y\delta(x, d) + \pi(x, d) \tag{13}$$

where $\delta(x, d)$ is the Kronecker delta function. The partial density function $\pi(x, d)$ corresponds to only the faulty chips. Since $F(x, d)$ is a density function, we have

$$\sum_{d=0}^{\infty} \int_0^1 F(x, d)\, dx = y + \sum_{d=0}^{\infty} \int_0^1 \pi(x, d)\, dx = 1.$$

Suppose, after application of n test vectors, that a certain fraction of chips has not failed. Then, the expected value of this fraction is the yield of chips after n vectors and is denoted by y_n.

Prob.(a chip does not fail after n vectors)

$$= \begin{cases} 1 & n \le d \\ (1 - x)^{n-d} & n > d \end{cases}$$

$$= I_{\{0, \cdots, d\}}(n) + (1 - x)^{n-d} I_{\{d+1, \cdots, \infty\}}(n).$$

Therefore,

$$y_n = \sum_{d=0}^{\infty} \int_0^1 [I_{\{0, \cdots, d\}}(n) + (1 - x)^{n-d} \cdot I_{\{d+1, \cdots, \infty\}}(n)]F(x, d)\, dx. \tag{14}$$

We assume that the random variables x and d are independent. This assumption is justified since the detection probability and latency of a fault depend on rather independent circuit characteristics. Detection probability is strongly influenced by the functionality of the circuit, while latency depends on the location of the fault site relative to the flip-flops in the circuit. Under the assumption, $\pi(x, d) = \pi_x(x)\pi_d(d)$, where $\pi_x(x)$ and $\pi_d(d)$ are the probability density functions of random variables x and d.

Probability of Chip Failure Detection at ith Vector: From (12), the probability that a chip fails at vector number i is $x(1 - x)^{i-d-1}$, $i > d$. Let N be the test length. Therefore, i takes a value between 1 and N. Since x and d are random variables, we use Bayes' theorem to write the probability that a chip fails at the ith vector as

$$\pi_i(x, d) = \frac{x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d) q_x(x)\, q_d(d)}{\sum_{d=0}^{\infty} \int_0^1 x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d) q_x(x) q_d(d)\, dx} \tag{15}$$

where $q_x(x)$ and $q_d(d)$ are the *a priori* density functions of detection probability and latency. For simplicity, we may assume uniform distribution for $q_x(x)$ and $q_d(d)$, with d taking integer values from 0 to N. Here, we assume that the test length is sufficiently long so as to include all the latencies.

$$q_x(x) = 1 \qquad 0 \le x \le 1$$

$$q_d(d) = \frac{1}{N + 1} \qquad 0 \le d \le N.$$

Hence,

$$\pi_i(x, d) = \frac{x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d)}{\sum_{d=0}^{N-1} \int_0^1 x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d)\, dx}$$

$$= \frac{i + 1}{i} x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d)$$

$$= k_i x(1 - x)^{i-d-1} I_{\{0, \cdots, i-1\}}(d) \tag{16}$$

where

$$k_i = (i + 1)/i. \tag{17}$$

It is easily verified that

$$\sum_{d=0}^{\infty} \int_0^1 \pi_i(x, d)\, dx = 1. \tag{18}$$

Probability of Chip Not Failing: The probability that a chip does not fail on the application of a test sequence is

$$(1 - x)^{N-d} \qquad 0 \le d \le N.$$

The corresponding Bayesian probability distribution is

$$\pi_0(x, d) = \frac{(1 - x)^{N-d} I_{\{0, \cdots, N\}}(d)}{\sum_{d=0}^{N} \int_0^1 (1 - x)^{N-d}\, dx}$$

$$= \frac{(1 - x)^{N-d} I_{\{0, \cdots, N\}}(d)}{\sum_{i=1}^{N+1} \frac{1}{i}}$$

$$= k_0(1 - x)^{N-d} I_{\{0, \cdots, N\}}(d) \tag{19}$$

where $k_0 = 1/\Sigma_{i=1}^{N+1} (1/i)$. Also,

$$\sum_{d=0}^{\infty} \int_0^1 \pi_0(x, d)\, dx = 1. \tag{20}$$

Estimation of $\pi(x, d)$: Having obtained the analytical expressions for the probability of a chip failing at the ith vector and for the chip not failing over the entire test sequence, we can now determine $\pi(x, d)$ from experimental data. Let a sample of C chips be tested by a sequence of N vectors. As these vectors are applied, we record the number of chips that fail for the first time on each vector. Let C_i denote the number of such chips for vector number

i. If y is the true yield, then $(1 - y - (1/C)\sum_{i=1}^{N} C_i)$ is the fraction of chips that are bad but did not fail on any of the vectors from 1 through N. To determine the complete chip failure detection probability distribution, $\pi_i(x, d)$ is weighted with C_i/C and $\pi_0(x, d)$ is weighted with $(1 - y - (1/C)\sum_{i=0}^{N} C_i)$. Thus,

$$\pi(x, d) = \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)\pi_0(x, d) + \frac{1}{C}\sum_{i=1}^{N} C_i\pi_i(x, d). \quad (21)$$

Using (18) and (20), we get

$$\sum_{d=0}^{\infty}\int_0^1 \pi(x, d)\,dx = 1 - y \quad (22)$$

which verifies that $\pi(x, d)$ is indeed a density distribution of failed chips. Substituting this in (13), after substitution for $\pi_i(x, d)$ and $\pi_0(x, d)$ from (16) and (19), we get

$$F(x, d) = y\delta(x, d) + \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)k_0(1 - x)^{N-d} \cdot I_{\{0, \cdots, N\}}(d) + \frac{1}{C}\sum_{i=d+1}^{N} C_i k_i x(1 - x)^{i-d-1}. \quad (23)$$

The second term in the above equation is the faulty chips not rejected by any of the N vectors. The third term groups the chips according to the vector number at which they failed. From the wafer test data, we have obtained the probability density distribution of chips in terms of the detection probability and fault latency.

Substituting for $F(x, d)$ in (14) gives us the following yield equation:

$$y_n = \sum_{d=0}^{\infty}\int_0^1 [I_{\{0, \cdots, d\}}(n) + (1 - x)^{n-d} I_{\{d+1, \cdots, \infty\}}(n)] \times \left[y\delta(x, d) + \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)k_0(1 - x)^{N-d} \cdot I_{\{0, \cdots, N\}}(d) + \frac{1}{C}\sum_{i=d+1}^{N} C_i k_i x(1 - x)^{i-d-1}\right] dx.$$

Multiplying out the terms in the square brackets, we get

$$y_n = \sum_{d=0}^{\infty}\int_0^1 y\delta(x, d) I_{\{0, \cdots, d\}}(n)\,dx + \sum_{d=0}^{\infty}\int_0^1 y\delta(x, d)\,(1 - x)^{n-d} I_{\{d+1, \cdots, \infty\}}(n)\,dx + \sum_{d=0}^{\infty}\int_0^1 \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)k_0(1 - x)^{N-d} I_{\{0, \cdots, d\}}(n) I_{\{0, \cdots, N\}}(d)\,dx + \sum_{d=0}^{\infty}\int_0^1 \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)k_0(1 - x)^{N+n-2d} I_{\{d+1, \cdots, \infty\}}(n) I_{\{0, \cdots, N\}}(d)\,dx + \sum_{d=0}^{\infty}\int_0^1 \frac{1}{C}\sum_{i=d+1}^{N} C_i k_i x(1 - x)^{i-d-1} I_{\{0, \cdots, d\}}(n)\,dx + \sum_{d=0}^{\infty}\int_0^1 \frac{1}{C}\sum_{i=d+1}^{N} C_i k_i x(1 - x)^{n+i-2d-1} I_{\{d+1, \cdots, \infty\}}(n)\,dx.$$

The first term in the above equation evaluates to zero since $I_{\{0, \cdots, d\}}(n) = 0$ for all values of $n > 0$. Using the property of Kronecker delta, the second term evaluates to y. To solve the integrals in the remaining terms, we use the following definition of beta function [9]:

$$\beta(m, n) = \int_0^1 x^{m-1}(1 - x)^{n-1}\,dx = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m + n)} \quad (24)$$

and

$$\Gamma(n) = (n - 1)!. \quad (25)$$

Therefore,

$$y_n = y + \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)k_0 \times \left[\sum_{d=0}^{n-1}\frac{1}{N + n - 2d + 1} + \sum_{d=n}^{N}\frac{1}{N - d + 1}\right] + \frac{1}{C}\sum_{d=n}^{N-1}\sum_{i=d+1}^{N}\frac{C_i k_i}{(i - d + 1)(i - d)} + \frac{1}{C}\sum_{d=0}^{n-1}\sum_{i=d+1}^{N}\frac{C_i k_i}{(n + i - 2d)(n + i - 2d + 1)}. \quad (26)$$

Also, the measured yield for N vectors is given by

$$y_N = \frac{\text{number of chips that pass the tests}}{C} = 1 - \frac{1}{C}\sum_{i=1}^{N} C_i. \quad (27)$$

Evaluating Reject Ratio: We have now obtained an analytical expression for y_n in terms of tester data. We need to determine the value of y so that the computed yield tracks the measured yield. We will therefore make an assumption that the analytical yield tracks the measured yield and the two are equal at the last test vector, i.e., at $n = N$ $y_n = y_N$. Making these substitutions in (26), we get

$$y_N = y + (y_N - y)k_0\left[\sum_{d=0}^{N-1}\frac{1}{2N - 2d + 1} + 1\right] + \frac{1}{C}\sum_{d=0}^{N-1}\sum_{i=d+1}^{N}\frac{C_i k_i}{(N + i - 2d)(N + i - 2d + 1)}. \quad (28)$$

Solving for y, we have

$$y = y_N - \frac{e_2}{1 - e_1} \tag{29}$$

where

$$e_1 = k_0 \left[\sum_{d=0}^{N-1} \frac{1}{2N - 2d + 1} + 1 \right]$$

and

$$e_2 = \frac{1}{C} \sum_{d=0}^{N-1} \sum_{i=d+1}^{N} \frac{C_i k_i}{(N + i - 2d)(N + i - 2d + 1)}.$$

The above expressions can be used for estimating the true yield, the apparent yield after n vectors, and the reject ratio from the chip failure data. The reject ratio is then computed as

$$r_N = \frac{y_N - y}{y_N} = \frac{e_2}{(1 - e_1)y_N}. \tag{30}$$

The application of this analysis is illustrated in the next section.

IV. Experiment

The wafer test data used earlier in Section II-C are used again to compute the reject ratio by the latency based method. The observed functional yield of the chip after the 12 188 clock steps is 0.712954, that is, the fraction of chips that passed the full test. According to our assumption in the preceding section, for $N = 12\ 188$, $y_N = 0.712954$. The estimated true yield y computed using (29) is 0.712923. The resulting reject ratio, computed from (30), is 43 ppm. The resolution in this measurement is 0.000015, which corresponds to one chip out of the total of 64 366 chips tested in this experiment. Fig. 2 shows the fit obtained for the experimental data. From the figure, we see that the computed yield closely tracks the measured yield. Fig. 3 gives the same data between 1–500 vectors at an enlarged scale to show how well our model can fit the jumps.

The experimental data also enable us to obtain the density function as described by (21). The derivation of this equation assumed *a priori* uniform distributions for the chip failure detection probability and the latency. Based on this assumption, Bayes' rule was used to estimate the actual distribution of failed chips. This distribution characterized by the experimental data is shown in Fig. 4. Although the chip failure detection density is concentrated at latency values that correspond to the vector numbers at which the actual chips failed, it has a nonzero value for all latencies. This can be explained as follows. If we rewrite (21) as

$$\pi(x, d) = \pi_b(x, d) + \pi_f(x, d) \tag{31}$$

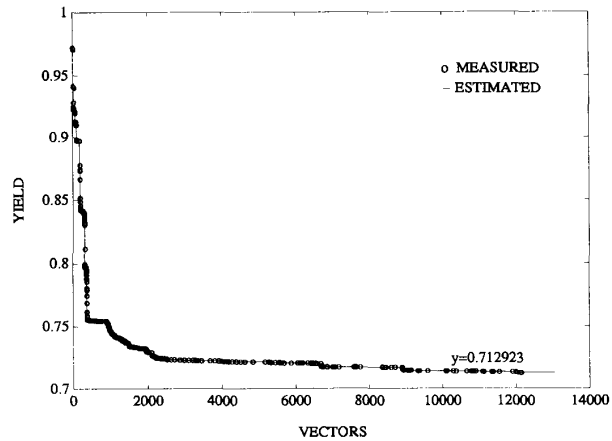

Fig. 2. Fitting the experimental data.

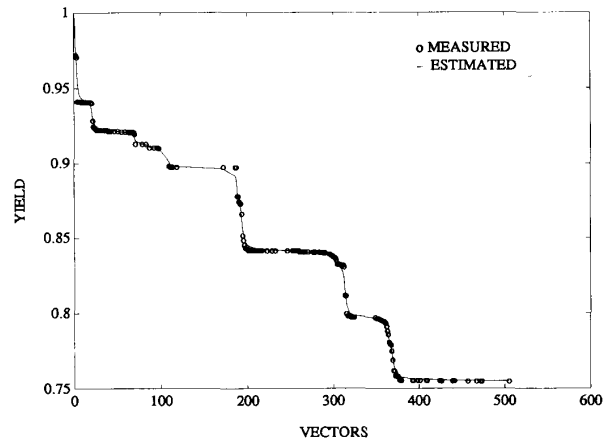

Fig. 3. Efficacy of the model in fitting the jumps.

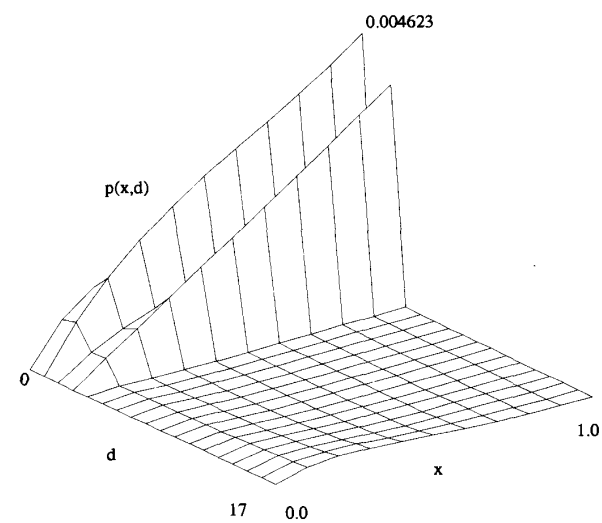

Fig. 4. Density distribution of failed chips.

where

$$\pi_b(x, d) = \left(1 - y - \frac{1}{C}\sum_{i=1}^{N} C_i\right)\pi_0(x, d)$$

$$\pi_f(x, d) = \frac{1}{C}\sum_{i=1}^{N} C_i\pi_i(x, d)$$

we see that the coefficient of $\pi_0(x)$, $(1 - y - (1/C)\ \Sigma_{i=1}^{N} C_i)$ is proportional to the number of bad chips tested as good. These chips can have a latency value between 0 and N and a detection probability in the interval [0, 1]. For a given value of d, the density function of these chips varies as $(1 - x)^{N-d}$. When $x = 0$, then $\pi_b(x, d) = (1 - y - (1/C)\ \Sigma_{i=1}^{N} C_i)$ for all values of d, and for $x = 1$, the value is 0. This component of the density function has the same shape for all latency values and is not included in the surface plot of Fig. 4.

Fig. 4 is actually the surface plot for the second component, i.e., $\pi_f(x, d)$. For clarity of illustration, only a limited range of latency d is shown. For any given value of d, the contribution to the density function is made by the fraction of chips that have failed during the testing process. If C_i chips failed at the ith vector, then these chips have a latency value between 0 and $i - 1$, and therefore make a contribution to the density function whose variation due to x is given as $x(1 - x)^{i-d-1}$. This expression has the value 0 at $x = 0$ and $x = 1$. For $d = i - 1$, the density function is proportional to x.

V. Robustness Analysis

In this section, we will analyze the variation in reject ratio as computed from (30) due to statistical variations in chip failure data. For a given N (the test length), e_1 in (30) is independent of the C_i's. Hence, r_N depends on C_i's only through e_2. We rewrite the expression for e_2 as a summation of C_i's, and expanding the outer summation, we get

$$e_2 = \frac{1}{C}\sum_{i=1}^{N}\frac{C_i k_i}{(N + i)(N + i + 1)} + \frac{1}{C}\sum_{i=1}^{N}\frac{C_i k_i}{(N + i - 2)(N + i - 1)} + \frac{1}{C}\sum_{i=3}^{N}\frac{C_i k_i}{(N + i - 4)(N + i - 3)} + \cdots + \frac{1}{C}\sum_{i=N}^{N}\frac{C_i k_i}{(N + i - 2N + 2)(N + i - 2N + 3)}$$

$$= \frac{1}{C}\left[C_1 k_1\left\{\frac{1}{(N + 1)(N + 2)}\right\} + C_2 k_2\left\{\frac{1}{(N + 2)(N + 3)} + \frac{1}{N(N + 1)}\right\} + \cdots + C_N k_N\left\{\frac{1}{2N(2N + 1)} + \frac{1}{(2N - 2)(2N - 1)} + \cdots + \frac{1}{2.3}\right\}\right]$$

$$= \frac{1}{C}[C_1 w_1 + C_2 w_2 + \cdots + C_N w_N].$$

In the above equation, the coefficients of C_i's have been labeled as w_i's. We can view e_2 as a weighted sum of C_i's.

Let V denote variance, Cov the covariance, and E the expectation of a random variable. To estimate the variance of r_N, it suffices to estimate the variance of e_2 since r_N depends on C_i's through e_2. Thus,

$$V(e_2) = E[V(e_2 \mid N)] + V[E(e_2 \mid N)]$$

$$V(e_2 \mid N) = \frac{1}{C^2}\sum_{i=1}^{N} w_i^2 V(C_i) + \frac{1}{C^2} \times \sum_{\substack{1 \le i,j \le N \\ i \ne j}} w_i\, w_j \cdot \mathrm{Cov}\,(C_i, C_j)$$

$$E(e_2 \mid N) = \frac{1}{C}\sum_{i=1}^{N} w_i E(e_i).$$

The number C_i of chips failed at vector number i is a statistical quantity. To evaluate the variance of e_2, we need to know the variance of C_i for all i. The distribution for any given C_i depends on the parameters of the process line, information about which is difficult to analyze. Therefore, we will use a Monte Carlo experiment to establish the robustness of the predicted value of reject ratio.

Monte Carlo Experiment: We assume a distribution for the random variable C_i and use a random number generator to obtain samples of C_i according to the distribution [11]. Thus, we can modify the chip failure data to compute the spread in the values of the reject ratio using (30). In the above analysis, we have made an assumption that test length is sufficiently long so as to cover all latencies. Therefore, we will assume that the number of test vectors N is fixed. We choose a binomial distribution with mean equal to the observed value of C_i. Now, if the data are modified a specified number of times, we can establish the confidence interval for reject ratio. The Monte Carlo procedure can be algorithmically stated as follows:

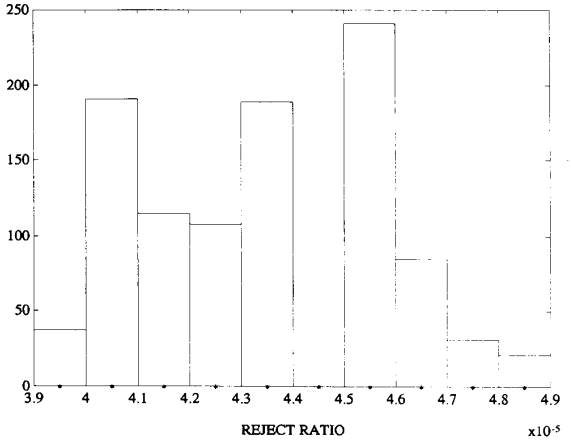

Fig. 5. Spread of reject ratio values.

```
while iterations < 1000 {
    for (i=1, i<=datapoints, i++){
        Change C[i] according to the binomial distribution;
    }
    Compute the observed yield with this new data set;
    if observed yield is within established guard band;
    then compute the new reject ratio;
    else reject the data set;
}
```

For the chip under consideration, the failure data are obtained in lots and then merged as a single data set. Thus, we do have the knowledge of percentage variation for each datapoint. Assuming a uniform distribution centered around the fraction of chips failing at any vector number and the width of this distribution as the maximum of the percentage variation measured in the observed data, we perform the Monte Carlo experiment. The experiment is repeated 1000 times. Fig. 5 is the histogram of reject ratio values. The height of a vertical bar represents the number of times the corresponding reject ratio value was observed. A 95% confidence interval of (41, 47) ppm for reject ratio is established by this method. The spread in the values of the reject ratio is from 35 to 50 ppm, as the yield varied from 0.69 to 0.73.

VI. Conclusion

The phenomenon of latency of faults in sequential circuits has been observed for a long time [12]. However, this is the first time a yield model for this phenomenon is presented. The model with two parameters, namely, per-vector detection probability and latency, has the necessary degrees of freedom to provide a close fit to experimental test data. As a result, reject ratio predictions will be more realistic. When latency is neglected, the fit to data becomes crude, and the analysis predicts a much higher reject ratio. At an abstract level, the attempt in this work is on extracting information from the yield versus vector number graph, which is related to the structural properties of the circuit. This information is used to predict the reject ratio.

References

[1] R. L. Wadsack, "Fault coverage in digital integrated circuits," *Bell Syst. Tech. J.*, vol. 57, pp. 1475–1488, May–June 1978.
[2] ——, "VLSI: How much fault coverage is enough," in *Proc. Int. Test Conf.*, Oct. 1981, pp. 547–554.
[3] T. W. Williams and N. C. Brown, "Defect level as a function of fault coverage," *IEEE Trans. Comput.*, vol. C-30, pp. 987–988, Dec. 1981.
[4] V. D. Agrawal, S. C. Seth, and P. Agrawal, "Fault coverage requirements in production testing of LSI circuits," *IEEE J. Solid State Circuits*, vol. SC-17, pp. 57–61, Feb. 1982.
[5] S. C. Seth and V. D. Agrawal, "Characterizing the LSI yield equation from wafer test data," *IEEE Trans. Comput.-Aided Design*, vol. CAD-3, pp. 123–126, Apr. 1984.
[6] ——, "On the probability of fault occurrence," in *Defect and Fault Tolerance in VLSI Systems*, I. Koren, Ed. New York: Plenum, 1989, pp. 47–52.
[7] D. V. Das, S. C. Seth, P. T. Wagner, J. C. Anderson, and V. D. Agrawal, "An experimental study on reject ratio prediction for VLSI circuits: Kokomo revisited," in *Proc. Int. Test Conf.*, Sept. 1990, pp. 712–720.
[8] V. D. Agrawal, S. C. Seth, and P. Agrawal, "LSI product quality and fault coverage," in *Proc. Design Automation Conf.*, June 1981, pp. 196–203.
[9] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill, 1974.
[10] D. V. Das, S. C. Seth, and V. D. Agrawal, "Estimating the quality of manufactured digital sequential circuits," in *Proc. Int. Test Conf.*, Oct. 1991, pp. 210–217.
[11] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. New York: Cambridge Univ. Press, 1988.
[12] J. J. Shedletsky and E. J. McCluskey, "The error latency of a fault in a sequential digital circuit," *IEEE Trans. Comput.*, vol. C-25, pp. 655–659, June 1976.

Dharamvir Das received the B.Tech. (EE) and M.Tech. (EE) degrees in 1973 and 1975 from the Indian Institute of Technology, Kanpur, and the Ph.D. degree in 1992 from the University of Nebraska, Lincoln. He is now with Cadence Design Systems, Noida, India.

He has worked for over 12 years at Tata Consultancy Services in India in the Systems Engineering Department. While at TCS, he traveled to the U.S. to work on various IC design projects at Hewlett Packard, Palo Alto, and Texas Instruments, Dallas.

Sharad C. Seth (S'66-M'70-SM'82) received the B.Eng. degree from the University of Jabalpur, India, the M.Tech. degree from the Indian Institute of Technology, Kanpur, and the Ph.D. degree from the University of Illinois.

He is a Professor with the Computer Science and Engineering Department at the University of Nebraska, Lincoln. He has held visiting positions at Indian Institute of Technology, Kanpur, India and at AT&T Bell Laboratories, Murray Hill, NJ. His current interests are VLSI testing, parallel CAD algorithm, and document analysis.

He has published extensively in the areas of testing and the analysis of optically scanned documents, and has coauthored with V. D. Agrawal the tutorial, *Test Generation for VLSI Chips* (IEEE-CS Press).

Dr. Seth serves on the editorial board of the *Journal of Electronic Testing: Theory and Applications* and is a member of ACM.

Vishwani D. Agrawal (S'68-M'70-SM'80-F'86), for a photograph and biography, see this issue, p. 461.