

Optimal Margin Computation for At-Speed Test

Jinjun Xiong[†], Vladimir Zolotov[†], Chandu Visweswariah[†], Peter A. Habitz^{*}

[†]IBM Thomas J. Watson Research Center ^{*}IBM Systems and Technology Group
Yorktown Heights, NY 10598 Essex Junction, VT 05452

{jinjun, zolotov, habitz, chandu}@us.ibm.com

Abstract—In the face of increased process variations, at-speed manufacturing test is necessary to detect subtle delay defects. This procedure necessarily tests chips at a slightly higher speed than the target frequency required in the field. The additional performance required on the tester is called *test margin*. There are many good reasons for margin including voltage and temperature requirements, incomplete test coverage, aging effects, coupling effects and accounting for modeling inaccuracies. By taking advantage of statistical timing, this paper proposes an optimal method of test margin determination to maximize yield while staying within a prescribed Shipped Product Quality Loss (SPQL) limit. If process information is available from wafer testing of scribe line structures or on-chip process monitoring circuitry, this information can be leveraged to determine a *per-chip test margin* which can further improve yield.

I. INTRODUCTION

Increased process variations make worst-case design too pessimistic. Manufactured chips exhibit wide performance distributions with a large fraction of fast chips and a long tail of slower ones. At-speed test has become an important addition to traditional testing methodology to screen out the tail and thereby target higher frequencies [1], [2]. It is typical to test a chip at a higher frequency than the performance requirement [3]. The difference between the test frequency and the chip’s operational frequency is *test margin*. There are many good reasons for margining: voltage and temperature in the field are different from the test chamber; test coverage is incomplete, so we need to provide margin for the untested portions; aging and coupling effects are not seen on the tester; and margining helps to cover modeling inaccuracies. The tester determines which chips are shipped to the customer, and the rest are discarded. Some of the shipped chips may actually be deficient, leading to Shipped Product Quality Loss (SPQL), which is constrained to be under a certain fraction (e.g., 0.1%). On the other hand, due to conservative test margins, some of the rejected chips may actually be good, resulting in yield loss. A conservative test margin improves SPQL but worsens yield loss, while an aggressive margin may cause unacceptably high SPQL.

Much of the delay test literature is devoted to faults due to local defects affecting individual transistors and interconnects [2]. Increased process variation creates delay faults of a different kind [4]. The effects of process variation are not localized, but felt by all components of the chip. Process variations cause subtle delay changes everywhere which can accumulate along signal propagation paths and adversely impact chip performance. Path delays become random variables correlated with each other. Many factors like circuit topology,

cell placement, global and spatial correlation contribute to the overall correlation and complicate the analysis. Statistical static timing analysis (SSTA) [5], [6] was introduced to predict probability distributions of circuit timing characteristics.

In this paper we focus on degradation of chip performance due to process variations. We use a statistical approach to compute an optimal test margin that maximizes yield while staying within an SPQL requirement. We leverage the capabilities of statistical timing to calculate probability distributions of path delays. We analyze two scenarios. In the first, the same or *uniform test margin* is applied to every chip. In the second, we assume that for each chip we measure performance-sensitive ring oscillators (PSROs) during wafer test, prior to at-speed testing. This information helps to improve yield by applying a *per-chip test margin* computed individually for each manufactured chip. Individual test margins can be applied by the tester to adjust the clock period or apply a derated voltage specific to each chip.

We will show that an “intuitive” margin computation method produces sub-optimal results, and demonstrate via Monte Carlo analysis that an optimal margin improves yield. A functional calculus approach enables the computation of optimal per-chip margins which further improve yield.

The rest of this paper is organized as follows. Section II provides the necessary background for test margin computation and formulates the problem to be solved. Section III presents intuitive and optimal approaches for computing uniform test margins. Section IV treats the case of optimal per-chip margins. Section V describes numerical experiments and presents a comparison of different test margin methods. Section VI draws conclusions and discusses future work.

II. BACKGROUND AND MOTIVATION

A. Chip performance testing

Chip disposition methods include PSRO measurements and at-speed testing. Measuring PSRO frequencies is fast and inexpensive. In some methodologies, disposition of ASIC chips was based exclusively on PSRO measurements. The correlation between PSRO performance and chip performance is imperfect, and hence at-speed test is an important addition.

In the test chamber, it is difficult to recreate the chip’s operational environment and often impossible to test the final intended function of the chip. It is easier to construct a special set of test patterns that are targeted at measuring delays of specially chosen critical paths, so-called *structural testing*. LSSD (Level-Sensitive Scan Design) [7] allows us to test

internal circuits of the chip by controlling and observing external pins only. At-speed structural test (ASST) [7], [1] exploits scan techniques to provide a powerful and low-cost test capability. Patterns are scanned in at a relatively low tester frequency, and then the functional clock of the chip is used to operate the chosen paths at-speed before the results are scanned out. Thus, high-frequency parts can be tested with a low-frequency tester. By means of an on-chip Test Waveform Generator (TWG) and deskewer, test frequency can be gradually increased till a path fails. Thus the maximum frequency of a part can be determined. While useful for diagnostics and model-to-hardware correlation, this procedure is slow and typically not applied during mass production. ASST is used to make a Boolean decision on whether each chip passes or fails at a single frequency. In this paper we compute the optimal test frequency for such a procedure.

B. Clock frequency and timing slack

Fig. 1 shows a fragment of a sequential circuit. Clock signal C1 launches data from flip-flop F1. The data signal D propagates through combinational logic and is captured at flop F2 by clock signal C2. Signal D can be latched by F2 only if its arrival time T_A is less than the required time T_R . The difference $S = T_R - T_A$ between the required and actual arrival times is timing slack. Zero slack is the minimum value at which the circuit can operate correctly. The required time T_R can be expressed in terms of cycle time T_{clk} as $T_R = T_{clk} - \tau$, where τ accounts for such effects as clock skew, latch setup time, and so on. Thus timing slack can be expressed as $S = T_R - T_A = T_{clk} - \tau - T_A$.

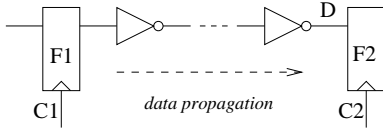


Fig. 1. Flop-to-flop data propagation.

The timing slack of a collection of paths is defined as the minimum of their individual slacks, in other words, the slack of the most critical path. With this background, we make the following definitions: *chip slack* is the timing slack of all paths of the chip; and *test slack* is the timing slack of only those paths that are tested. For convenience we assume that both chip and test slacks are computed relative to the chip’s operational frequency. Thus, test slack never exceeds chip slack since a subset of the chip’s paths is tested.

If at the operational frequency, test slack has a positive value S_T , it means that the clock cycle can be decreased by S_T and the chip will still pass the test. Testing with a frequency corresponding to a clock cycle reduced by S_T is equivalent to demanding a positive slack of at least S_T . Therefore, instead of computing an optimal test frequency, we compute the optimal value of required test slack and then convert it into a corresponding test frequency. In the context of timing analysis, slack is more convenient than clock frequency. Since slack and frequency can trivially be derived from each other,

in the rest of this paper, *test margin* refers to the *additional slack required during testing*.

C. Joint distribution of chip and test slacks

Due to process variation, chip and test slacks are correlated. Statistical timing [5], [6] approximates them as linear forms

$$S = S_0 + \sum_{i=1}^n a_i \Delta X_i + a_R \Delta R_a \quad (1)$$

where ΔX_i and ΔR_a are zero-mean unit Gaussians. Variables ΔX_i model globally correlated variations of process parameters and ΔR_a models uncorrelated variation. S_0 is the mean or nominal value of the slack. Coefficients a_i and a_R are sensitivities to the corresponding variations. The benefit of this representation is that the correlation between two canonical forms can be immediately judged based on sensitivities to common process variables.

PSRO slack can also be computed in the form (1) either by statistical timing of its open loop circuit or by linear regression of Monte Carlo SPICE simulation results. Any slack in the form (1) is a linear combination of Gaussians and hence has a Gaussian probability distribution. Similarly, the joint distribution of chip and test slacks is a multivariate Gaussian distribution. The variances σ_j^2 and covariances $cov(j, k)$ of these distributions are computed from the sensitivities of the corresponding canonical forms as

$$\sigma_j^2 = \sum_{i=1}^n a_{j,i}^2 + a_{j,R}^2 \quad cov(j, k) = \sum_{i=1}^n a_{j,i} a_{k,i}. \quad (2)$$

where $a_{j,i}$ and $a_{j,R}$ are sensitivities to globally correlated and uncorrelated variations.

Fig. 2 shows the Joint Probability Density Function (JPDF) of test and chip slacks. The ellipses represent contours of equal probability. Chips above the horizontal axis have positive chip slack, and therefore satisfy performance requirements. Chips below the horizontal axis are bad because their slack is negative. The dotted vertical line represents the test margin. Chips to the right of this line pass the test and are shipped to a customer, while chips to the left of this line are discarded. From Fig. 2, four types of chips are evident. The region marked “Good chips” represents good chips that pass the test and are shipped. The region marked “Bad chips” comprises chips that fail the test and are discarded. The region marked “Yield loss” represents chips that fail the test, but are actually good chips. The region marked “SPQL” contains chips that pass the test but do not satisfy performance requirements.

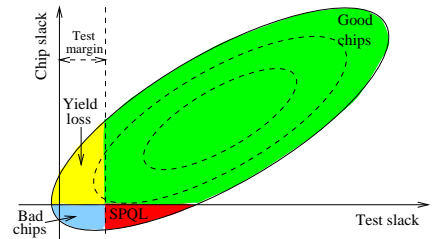


Fig. 2. Joint distribution of chip and test slack.

From Fig. 2 we see that as the margin increases, SPQL improves but yield loss worsens, and *vice versa*. We also see that better correlation between test and chip slack helps to reduce both SPQL and yield loss. Using additional information such as PSRO measurements, we can tighten the JPFD shown in Fig. 2. The resulting conditional probability distribution has better correlation between chip and test slacks and allows us to improve testing quality.

Our goal here is to compute test margin so as to maximize yield without exceeding a given SPQL value. This problem is equivalent to maximizing the fraction of chips shipped subject to an SPQL constraint.

III. UNIFORM TEST MARGIN

The problem of uniform test margin is formulated as follows: Formulation 1. Given chip and test slacks S_C, S_T in the form of (1) and a maximum allowed SPQL q , compute test margin S_M as the solution of the optimization problem

$$\max_{S_M} P(S_T \geq S_M) \quad (3)$$

$$\text{s.t. } P(S_C \leq 0 | S_T \geq S_M) \leq q, \quad (4)$$

where $P(S_T \geq S_M)$ is the probability of shipping a chip and $P(S_C \leq 0 | S_T \geq S_M)$ is the conditional probability that a shipped chip is deficient (i.e., SPQL).

A. Intuitive approach to uniform test margin

We consider the statistical difference $\Delta S = S_C - S_T$ between the chip and test slacks. For passing chips, we know that $S_T \geq S_M$, so S_M is a lower bound for S_T . Thus,

$$\begin{aligned} P(S_C < 0 | S_T \geq S_M) &= P(S_T + \Delta S < 0 | S_T \geq S_M) \\ &\leq P(S_M + \Delta S < 0 | S_T \geq S_M). \end{aligned} \quad (5)$$

A warning to the naïve reader: this is not a straightforward way to compute S_M since ΔS is correlated to S_T .

We instead decompose chip slack into a linear combination of a part that is correlated to test slack, and a part that is uncorrelated, i.e., $S_C = \alpha S_T + \Delta S_u$ where ΔS_u is uncorrelated with test slack, and we assume that α is a positive constant. Then SPQL can be expressed as

$$\begin{aligned} P(S_C \leq 0 | S_T \geq S_M) &= \frac{P(S_C \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)} \\ &= \frac{P(\alpha S_T + \Delta S_u \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)} \\ &\leq \frac{P(\alpha S_M + \Delta S_u \leq 0) P(S_T \geq S_M)}{P(S_T \geq S_M)} \\ &= P(\alpha S_M + \Delta S_u \leq 0). \end{aligned} \quad (6)$$

Thus to satisfy $\text{SPQL} \leq q$, it is sufficient to have $P(\alpha S_M + \Delta S_u \leq 0) \leq q$, and therefore a conservative estimate for the test margin is

$$S_M = \frac{-1}{\alpha} (\sigma_u \Phi^{-1}(q) + \mu_u), \quad (7)$$

where Φ^{-1} represents the inverse of the CDF of a unit Gaussian, and μ_u, σ_u are the mean and standard deviation of

ΔS_u . The resulting test margin will guarantee the required SPQL level, but is sub-optimal.

B. Exact uniform test margin

The probability of shipping a chip is

$$P(S_T \geq S_M) = \int_{S_M}^{\infty} p_t(S_T) dS_T, \quad (8)$$

where $p_t(S_T)$ is the PDF of S_T . This probability is a monotone function of test margin S_M , and therefore it reaches its maximum at the minimum allowed value S_M .

The conditional probability of a defective shipped chip, i.e., SPQL is

$$\begin{aligned} Q = P(S_C \leq 0 | S_T \geq S_M) &= \frac{P(S_C \leq 0 \& S_T \geq S_M)}{P(S_T \geq S_M)} \\ &= \frac{\int_{S_M}^{\infty} \int_{-\infty}^0 p_c(S_C, S_T) dS_C dS_T}{\int_{S_M}^{\infty} p_t(S_T) dS_T}, \end{aligned} \quad (9)$$

where $p_c(S_C, S_T)$ is the JPFD shown in Fig. 2. Provided that the correlation coefficient between chip and test slacks is not 0, SPQL is a monotone function of test margin S_M , too. The proof is outlined below.

The derivative of SPQL with respect to S_M is

$$\frac{dQ}{dS_M} = \frac{A'B - AB'}{B^2} = \frac{A'}{B^2} \left(B - \frac{AB'}{A'} \right) \quad (10)$$

where

$$\begin{aligned} A &= \int_{S_M}^{\infty} \int_{-\infty}^0 p_c(S_C, S_T) dS_C dS_T \\ &= \int_{S_M}^{\infty} \left(p_t(S_T) \int_{-\infty}^0 p_c(S_C | S_T) dS_C \right) dS_T \end{aligned} \quad (11)$$

$$B = \int_{S_M}^{\infty} p_t(S_T) dS_T \quad (12)$$

$$A' = \frac{dA}{dS_M} = -p_t(S_M) \int_{-\infty}^0 p_c(S_C | S_M) dS_C \quad (13)$$

$$B' = \frac{dB}{dS_M} = -p_t(S_M), \quad (14)$$

where $p_c(S_C | S_T)$ is shorthand notation for the conditional probability density of S_C at a given value of S_T and $p_c(S_C | S_M)$ implies $p_c(S_C | S_T = S_M)$. Using the formula for a conditional PDF, we get

$$\frac{AB'}{A'} = \int_{S_M}^{\infty} p_t(S_T) \frac{\int_{-\infty}^0 p_c(S_C | S_T) dS_C}{\int_{-\infty}^0 p_c(S_C | S_M) dS_C} dS_T. \quad (15)$$

Using the formula for a conditional Gaussian [8], we get

$$\int_{-\infty}^0 p_c(S_C | S_T) dS_C = \Phi \left(-\frac{\mu_C + \rho \frac{\sigma_C}{\sigma_T} (S_T - \mu_T)}{\sigma_C \sqrt{(1 - \rho^2)}} \right) \quad (16)$$

$$\int_{-\infty}^0 p_c(S_C | S_M) dS_C = \Phi \left(-\frac{\mu_C + \rho \frac{\sigma_C}{\sigma_T} (S_M - \mu_T)}{\sigma_C \sqrt{(1 - \rho^2)}} \right) \quad (17)$$

where $\Phi(x)$ is a standard Gaussian CDF.

The two Gaussian integrals (16) and (17) differ from each other only in the appearance of S_T and S_M . For passing chips, $s_t \geq S_M$, so the ratio of these integrals depends on the covariance ρ between test and chip slacks as follows:

$$0 < \frac{\int_{-\infty}^0 p_c(S_C|S_T)dS_C}{\int_{-\infty}^0 p_c(S_C|S_M)dS_C} \begin{cases} < 1 & \text{if } \rho > 0 \\ = 1 & \text{if } \rho = 0 \\ > 1 & \text{if } \rho < 0 \end{cases} \quad (18)$$

Since the integrand of (15) is the integrand of (12) multiplied by the ratio (18), $(B - AB'/A')$ in (10) is positive (negative) when ρ is positive (negative). Thus, the SPQL derivative (10) is negative (positive) when ρ is positive (negative). SPQL is therefore a monotone function of the test margin, which confirms our observation made in Section II-C.

Since both the objective function (3) and constraint (4) are monotone functions of S_M , the optimal test margin can be computed from the constraint

$$P(S_C \leq 0 | S_T \geq S_M) = q. \quad (19)$$

Rewriting (9), SPQL is

$$\int_{S_M}^{\infty} \int_{-\infty}^0 p_c(S_C, S_T) dS_C dS_T = q \int_{S_M}^{\infty} p_t(S_T) dS_T. \quad (20)$$

Taking into consideration that $p_c(S_C, S_T)$ and $p_t(S_T)$ are Gaussian PDFs, this equation can be simplified. The double integral of the left-hand side can be reduced to a one-dimensional integral by analytic integration over S_T and the right-hand side can also be integrated analytically. The resulting equation has only a single-dimensional integral and can be solved numerically.

IV. PER-CHIP TEST MARGIN

In this section, we take advantage of PRSO measurements to determine optimal test margins. The margin computation is specific to each chip based on its PSRO measurement. As before, we assume that the measured PSRO frequency is converted to a timing slack in linear canonical form (1), either by SSTA or linear regression of Monte Carlo results. Under these assumptions we formulate the problem as follows:

Formulation 2. Given chip, test and PSRO slacks S_C , S_T , S_P in the form of (1), for each value of PSRO slack S_P compute the test margin $S_M(S_P)$ to maximize the fraction of shipped chips without exceeding the required SPQL q , i.e.,

$$\begin{aligned} & \max_{S_M(S_P)} P(S_T \geq S_M(S_P)) & (21) \\ \text{s.t. } & P(S_C \leq 0 | S_T \geq S_M(S_P)) \leq q. & (22) \end{aligned}$$

The objective function and constraint of this problem look similar to those of (3) and (4). The only difference is the dependence of test margin on measured PSRO slack. However, this difference dramatically changes the optimization problem. Instead of computing a single optimal value of S_M , we need to compute the function $S_M(S_P)$ that delivers the optimal solution for each chip. Now both the objective function and

constraint are functionals. This difference is more obvious if we reformulate the probabilities as integrals.

$$\begin{aligned} & \max_{S_M(S_P)} \int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} p_t(S_T, S_P) dS_T dS_P & (23) \\ \text{s.t. } & \frac{\int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} \int_{-\infty}^0 p_c(S_C, S_T, S_P) dS_C dS_T dS_P}{\int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} p_t(S_T, S_P) dS_T dS_P} \leq q. & (24) \end{aligned}$$

We see that our optimization problem belongs to the domain of variational calculus [9]. We formulate the constraint in the form of an equality using the obvious fact that (in non-degenerate cases) the optimal solution is achieved when SPQL is exactly as required and no less. For brevity, we do not consider here the trivial case when the probability of manufacturing a bad chip is less than the required SPQL. For convenience we transform the constraint in linear form as

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} \int_{-\infty}^0 p_c(S_C, S_T, S_P) dS_C dS_T dS_P \\ & - q \int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} p_t(S_T, S_P) dS_T dS_P = 0. \end{aligned} \quad (25)$$

The Lagrangian can be written as

$$\begin{aligned} & \int_{-\infty}^{\infty} \int_{S_M(S_P)}^{\infty} \left((1 + \lambda q) p_t(S_T, S_P) \right. \\ & \left. - \lambda \int_{-\infty}^0 p_c(S_C, S_T, S_P) dS_C \right) dS_T dS_P \end{aligned} \quad (26)$$

where λ is the Lagrange multiplier.

From variational calculus [9], it is known that when the functional $\int_{-\infty}^{\infty} H(x, y(x)) dx$ reaches its optimum, $y(x)$ satisfies the equation $\frac{\partial H}{\partial y} = 0$. Therefore, we get the following equation for $S_M(S_P, \lambda)$:

$$(1 + \lambda q) p_t(S_M, S_P) - \lambda \int_{-\infty}^0 p_c(S_C, S_M, S_P) dS_C = 0. \quad (27)$$

Dividing by $\lambda p_t(S_M, S_P)$ and using the formula for conditional probability we get

$$\int_{-\infty}^0 p_c(S_C | S_M, S_P) dS_C = q + \frac{1}{\lambda}. \quad (28)$$

Assume that the vector of slacks S , vector of mean values μ and correlation matrix Σ of the JPDP $p_c(S_C, S_T, S_P)$ are partitioned as follows

$$S = \begin{pmatrix} S_C \\ S_T \\ S_P \end{pmatrix} = \begin{pmatrix} S_C \\ S_{TP} \end{pmatrix} \quad (29)$$

$$\mu = \begin{pmatrix} \mu_C \\ \mu_T \\ \mu_P \end{pmatrix} = \begin{pmatrix} \mu_C \\ \mu_{TP} \end{pmatrix} \quad (30)$$

$$\Sigma = \begin{pmatrix} \sigma_C^2 & \rho_{C,T} & \rho_{C,P} \\ \rho_{C,T} & \sigma_T^2 & \rho_{T,P} \\ \rho_{C,P} & \rho_{T,P} & \sigma_P^2 \end{pmatrix} = \begin{pmatrix} \sigma_C^2 & \rho_{C,TP} \\ \rho_{C,TP} & \Sigma_{TP} \end{pmatrix} \quad (31)$$

Then the conditional PDF $p_c(S_C|S_M, S_P)$ is a Gaussian distribution [8] with mean $\hat{\mu}_c$ and variance $\hat{\sigma}_c$ given by

$$\hat{\mu}_c = \mu_C + \rho_{C,TP} \Sigma_{TP}^{-1} (S_{MP} - \mu_{TP}) \quad (32)$$

$$\hat{\sigma}_c^2 = \sigma_C^2 - \rho_{C,TP} \Sigma_{TP}^{-1} \rho_{C,TP}^T \quad (33)$$

where according to equation (29)

$$S_{MP} = \begin{pmatrix} S_M(S_P) \\ S_P \end{pmatrix}. \quad (34)$$

Performing integration of the Gaussian PDF in (28) and solving, we get

$$\hat{\mu}_c = -\hat{\sigma}_c \Phi^{-1}(q + 1/\lambda), \quad (35)$$

where $\Phi(x)$ is the standard normal CDF.

Substituting expressions for $\rho_{C,TP}$, Σ_{TP} , S_{MP} and μ_{TP} into (32) and performing matrix-vector multiplication we can show that

$$\hat{\mu}_c = \mu_C + \alpha S_M(S_P) + \beta S_P - \alpha \mu_T - \beta \mu_P \quad (36)$$

where α and β are expressed through variances and covariances of the test, chip and PSRO slacks. Excluding $\hat{\mu}_c$ from (35) and (36), and solving for S_M we get

$$S_M(S_P) = -\frac{\beta}{\alpha} S_P + \mu_T - \frac{\mu_C}{\alpha} + \frac{\beta}{\alpha} \mu_P - \frac{\hat{\sigma}_c}{\alpha} \Phi^{-1}\left(q + \frac{1}{\lambda}\right).$$

We see that test margin is a linear function of PSRO slack. For brevity we rewrite it as

$$S_M(S_P) = \gamma S_P + \eta. \quad (37)$$

The Lagrange multiplier λ can easily be found by computing η . By changing the order of integration in the numerator of (24) and transforming nested integrals into an integral over the area $S_T \geq S_M(S_P) = \gamma S_P + \eta$,

$$\frac{\int_{-\infty}^0 \left(\iint_{S_T \geq \gamma S_P + \eta} p_c(S_C, S_T, S_P) dS_T dS_P \right) dS_C}{\iint_{S_T \geq \gamma S_P + \eta} p_t(S_T, S_P) dS_T dS_P} = q.$$

Rotating the coordinate system by variable transformations

$$S_P = \frac{u - \gamma v}{\sqrt{1 + \gamma^2}} \quad S_T = \frac{\gamma u + v}{\sqrt{1 + \gamma^2}} \quad (38)$$

and converting the integrals over the area back into nested integrals, we get

$$\frac{\int_{-\infty}^0 \int_{\frac{\eta}{\sqrt{1+\gamma^2}}}^{\infty} \int_{-\infty}^{\infty} p_c\left(S_C, \frac{\gamma u + v}{\sqrt{1+\gamma^2}}, \frac{u - \gamma v}{\sqrt{1+\gamma^2}}\right) du dv dS_C}{\int_{\frac{\eta}{\sqrt{1+\gamma^2}}}^{\infty} \int_{-\infty}^{\infty} p_t\left(\frac{\gamma u + v}{\sqrt{1+\gamma^2}}, \frac{u - \gamma v}{\sqrt{1+\gamma^2}}\right) du dv} = q. \quad (39)$$

The region of integration of the two inner integrals of the numerator and both the integrals of the denominator is a half plane, and these integrals can be expressed analytically in terms of the standard Gaussian CDF function $\Phi(x)$. This transforms (39) into a single integral. Applying numerical integration, we can efficiently solve this equation for η by any root-finding technique. The infinite upper limit does not create any problem for numerical integration because the relevant

part of the function is located near the mean value of the distribution.

Substituting the computed value of η into (37), we get the optimal value of the test margin. Thus, by combining analytical and numerical methods we can determine the optimal test policy. Equation (37) shows that the optimal test policy is a linear function of measured PSRO slack. We also can predict yield, i.e., the fraction of manufactured chips that passes the optimally determined test and is shipped to the customer. It is equal to the probability of shipping chips expressed by the integral (23). This integral is exactly the denominator of (24) which, as we showed, is expressed analytically in terms of a Gaussian CDF. Substituting the value of η into this expression, we obtain the yield corresponding to our optimal test policy.

V. NUMERICAL RESULTS

We validate our proposed techniques by using two industrial 90 nm ASICs, each with over a million placeable objects. The linear canonical forms of chip slack, test slack and PSRO slack are obtained from a statistical timing analysis engine [5]. The sources and amounts of process variation are determined according to foundry rules for this technology. For a given user-specified SPQL requirement, we compute three test margins: conservative (“intuitive”) uniform margin, optimal uniform margin, and optimal per-chip margin.

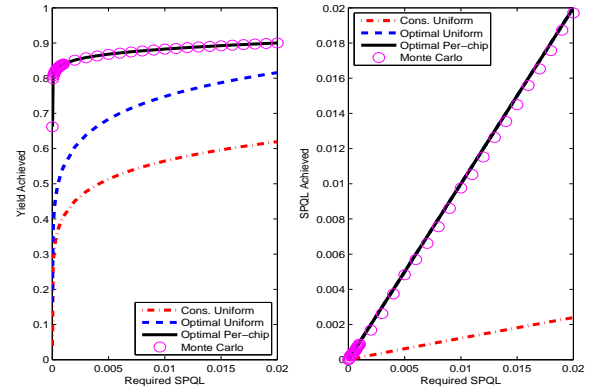


Fig. 3. Achieved yield and SPQL for different margin policies.

Fig. 3 shows the comparison of these three methods for different values of required SPQL shown on the x-axis. The left plot shows the comparison of yield achieved under different test margin policies; the right plot shows the achieved SPQL for the same policies. From the left plot, it is evident that for a given SPQL requirement, the optimal per-chip margin achieves the highest yield, the optimal uniform margin achieves the second highest, and the conservative uniform margin is the lowest. This is expected because the former two policies are optimal in leveraging SPQL to the fullest to maximize yield, and it explains why both policies meet the required SPQL almost exactly as shown in the right plot. In contrast, the conservative uniform margin over-achieves on SPQL at the cost of yield. The optimal per-chip policy exploits chip-specific information, and hence achieves the best yield.

TABLE I
COMPARISON OF DIFFERENT TEST MARGIN POLICIES.

Required SPQL		1.0e-6	1.0e-5	1.0e-4	1.0e-3	1.0e-2	
Design One	Achieved SPQL	Conservative uniform	0.0 (0.0)	1.0e-6 (0.0)	1.3e-5 (3.5e-5)	1.3e-4 (2.2e-4)	1.2e-3 (1.5e-3)
		Optimal uniform	1.0e-6 (0.0)	1.0e-5 (0.0)	1.0e-4 (1.8e-4)	1.0e-3 (1.3e-3)	1.0e-2 (1.0e-2)
		Optimal per-chip	1.0e-6 (0.0)	1.0e-5 (0.0)	1.0e-4 (9.9e-5)	1.0e-3 (8.8e-4)	1.0e-2 (9.8e-3)
	Achieved yield	Conservative uniform	14% (14%)	20% (20%)	29% (29%)	41% (41%)	56% (57%)
		Optimal uniform	20% (20%)	28% (28%)	39% (39%)	55% (55%)	75% (75%)
		Optimal per-chip	75% (75%)	78% (78%)	81% (81%)	84% (84%)	88% (88%)
Design Two	Achieved SPQL	Conservative uniform	1.5e-7 (0.0)	9.7e-7 (0.0)	8.9e-6 (2.0e-5)	8.3e-5 (9.6e-5)	7.4e-4 (7.7e-4)
		Optimal uniform	1.0e-6 (0.0)	1.0e-5 (1.9e-5)	1.0e-4 (9.4e-5)	1.0e-3 (9.4e-4)	1.0e-2 (9.9e-3)
		Optimal per-chip	1.0e-6 (0.0)	1.0e-5 (0.0)	1.0e-4 (1.0e-4)	1.0e-3 (1.1e-3)	1.0e-2 (9.6e-3)
	Achieved yield	Conservative uniform	31% (32%)	40% (40%)	51% (51%)	63% (63%)	76% (76%)
		Optimal uniform	40% (40%)	51% (52%)	64% (64%)	78% (77%)	91% (91%)
		Optimal per-chip	62% (62%)	69% (69%)	77% (77%)	85% (85%)	94% (94%)

Fig. 4 compares test margins among the three policies. The left plot shows the comparison between conservative and optimal uniform margins as a function of required SPQL. It clearly shows that the conservative policy produces higher margins than necessary, resulting in yield loss. As we have shown in the previous section, optimal per-chip margin is a linear function of PSRO slack. This is illustrated in the right plot of Fig. 4, where policies corresponding to four different required SPQL values are shown. It can be seen from the figure that as required SPQL becomes stricter (i.e., smaller), the margin policy becomes tighter and a higher margin is demanded of working chips. As PSRO slack increases (faster hardware), a lower margin is sufficient for a given SPQL value.

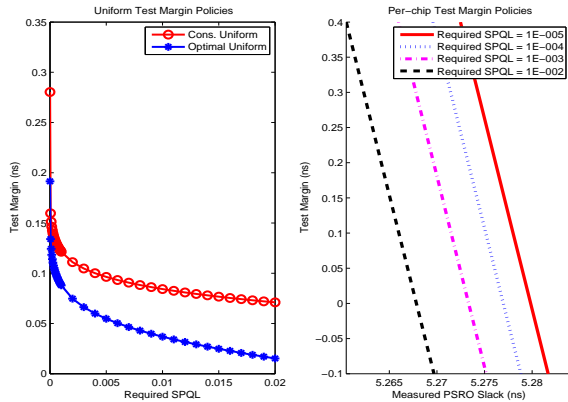


Fig. 4. Test margins comparison for the required SPQL.

To further verify our theoretical derivation, we apply Monte Carlo simulation to obtain 100,000 manufactured chips and their corresponding chip, test, and PSRO slacks. We then apply the three proposed test margin policies to all chips, and compute the respective SPQLs and yields. We compare Monte Carlo results with our theoretical results. In Fig. 3, Monte Carlo results are shown as circles and closely follow our theoretical results for all three policies, confirming the validity of our derivation. So as not to clutter the plot, Monte Carlo results are only shown for the optimal per-chip margin policy.

We present numerical comparisons in Table I for five required SPQL values. For each policy, both theoretical results and Monte Carlo results in parentheses are shown. Again, we see that the conservative uniform margin achieves unnecessarily low SPQL at the cost of yield. Both optimal margin policies

achieve exactly the required SPQL in order to maximize yield. Moving from the conservative to the optimal uniform policy increases yield from 6% to 19%. Moving from the optimal uniform policy to a per-chip policy further increases yield by as much as 55%, with the most significant gains at the highest quality levels! Monte Carlo simulation matches our theoretical results very closely across the board, thus demonstrating the value and correctness of the proposed techniques.

VI. FUTURE WORK AND CONCLUSIONS

This paper presented a method for optimal determination of test margins for at-speed testing. Yield loss can be minimized for a given Shipped Product Quality Loss (SPQL) limit. By exploiting statistical timing of the chip and the subset of the chip that is tested, the joint probability density function of the chip slack and test slack are used to determine the optimal test margin. In addition, partial process information can be exploited to further optimize test margin on a per-lot or per-chip basis. All the computations in this paper assume perfect knowledge and modeling of process variation distributions and delay sensitivities. A topic of future work is to extend this framework to handle unknown or erroneous models – i.e., determination of test margins in the presence of bounded modeling errors and testing errors.

REFERENCES

- [1] V. Iyengar, T. Yokota, K. Yamada, T. Anemikos, R. Bassett, M. De-gregorio, R. Farmer, G. Grise, M. Johnson, D. Milton, M. Taylor, and F. Woytowich. At-speed structural test for high-performance ASICs. *ITC*, pages 2.4:1–10, October 2006. Santa Clara, CA.
- [2] M. L. Bushnell and V. D. Agrawal. *Essentials of electronic testing for digital, memory and mixed-signal VLSI circuits*. Kluwer Academic Publishers, 2000.
- [3] M. Amodeo and B. Cory. Defining faster-than-at-speed delay test. *Nanometer Test Article*, April 2005. Cadence Inc.
- [4] P. S. Zuchowski, P. A. Habitz, J. D. Hayes, and J. H. Oppold. Process and environmental variation impacts on ASIC timing. *ICCAD*, pages 336–342, November 2004. San Jose, CA.
- [5] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan. First-order incremental block-based statistical timing analysis. *DAC*, pages 331–336, June 2004. San Diego, CA.
- [6] H. Chang and S. S. Sapatnekar. Statistical timing analysis considering spatial correlations using a single PERT-like traversal. *ICCAD*, pages 621–625, November 2003. San Jose, CA.
- [7] J. Saxena, K. M. Butler, J. Gatt, R. Raghuraman, S. P. Kumar, S. Basu, D. J. Campbell, and J. Berech. Scan-based transition fault testing – implementation and low-cost test challenges. *ITC*, pages 1120–1129, October 2002. Baltimore, MD.
- [8] M. J. Press. *Applied multivariate analysis*. Dover Publications, 2005.
- [9] R. Weinstock. *Calculus of variations*. Dover Publications, 1974.