Optimal Margin Computation for At-Speed Test

Jinjun Xiong, Vladimir Zolotov, Chandu Visweswariah, Peter A. Habitz

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 called *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 on test margin and motivation behind this work. Section III discusses modeling and formulates the problem to be solved. Section IV presents both conservative and optimal approaches for computing uniform test margins. Section V treats the case of optimal per-chip margins. Section VI describes

Dr. Xiong, Dr. Zolotov, and Dr. Visweswariah are with IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598. Dr. Habitz is with IBM Systems and Technology Group, Essex Junction, VT 05452. The authors can be contacted via e-mail at {*jinjun, zolotov, chandu, habitz*}@*us.ibm.com*

numerical experiments and presents a comparison of different test margin methods. Section VII 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.

B. Concept of Test Margin

For chips coming off the manufacturing line, whether or not their performance meets customer-specified clock frequency is not known. Instead, some testing paths from the chip are selected and tested at a higher testing frequency in the fab. The difference between the testing frequency and customer specification is called *test margin*.

Any chips that pass this more stringent testing frequency are assumed to have high probability of meeting customer specifications, hence are shipped to customers. On the contrary, any chips that fail to pass this stringent testing frequency are assumed to have low probability of meeting customer specification, hence are discarded at the fab.

As the selection of testing paths is not "perfect," it is apparent that four types of chips can be produced through this testing procedure.

1) *Good chips* are chips whose testing paths pass the testing frequency and their chip performance indeed meets the customer specification.
2) *Bad chips* are chips whose testing paths fail to pass the testing frequency and their chip performance does not meet the customer specification.
3) *Yield loss chips* are chips whose testing paths fail to pass the testing frequency but their chip performance in fact meets the customer specification.
4) *SPQL (shipped-product quality loss) chips* are chips whose testing paths pass the testing frequency but their chip performance in fact does not meet the customer specification.

In a typical ASIC business model, a foundry charges money for every chip it ships to customers, and this includes both good chips and SPQL chips. The ratio of shipped chips to all manufactured chips is called *yield*, i.e.,

$$Yield = 100 \times \frac{Good_chips + SPQL_chips}{Total_chips}\%. \quad (1)$$

where $Total_chips$ represents all manufactured chips, i.e., the sum of all four types of chips as discussed above.

The flip side of the coin is that the foundry is also penalized by shipping too many SPQL chips to customers. The foundry is contracted to guarantee a certain level of SPQL requirement, i.e.,

$$SPQL = 100 \times \frac{SPQL_chips}{Good_chips + SPQL_chips}\%. \quad (2)$$

In other words, the required SPQL level is the ratio of failing chips to all chips shipped to customers. For example, when the required SPQL is set as 0.5%, it means that every 1000 chips that pass testing (and hence are shipped to customers) contains no more than five failing chips. If this required SPQL were not met, the foundry would be penalized "severely" to compensate the customer's financial loss.

Intuitively, it is clear that the more stringent the testing, i.e., the higher the test margin, the lower the yield, but the better the SPQL, and *vice versa*. Therefore, a foundry has incentive to find an optimal test margin such that it can ship as many chips to customers as possible while staying at the required SPQL.

But how to find such an optimal test margin has not been addressed in the literature. In practice, test margin is most likely determined from previous experience. In this work, we mathematically formulate this test margin selection problem, and provide a rigorously proven optimal solution to this problem.

C. 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, flip-flop setup time, and so on. Thus timing slack can be expressed as $S = T_R - T_A = T_{clk} - \tau - T_A$.

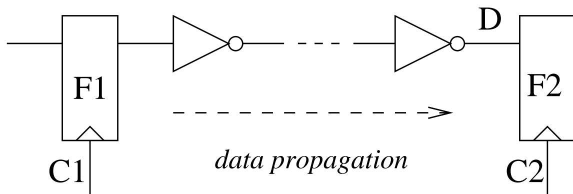

Fig. 1. Flop-to-flop data propagation.

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, chip slack never exceeds test 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*.

III. Modeling and Problem Formulation

A. Statistical modeling of timing slacks

Because of process variation, chip and test slacks are no longer fixed numbers among all chips manufactured. Instead, different chips may exhibit different chip slacks and test slacks, depending on each chip's manufacturing conditions. This effect can be captured by modeling both chip and test slacks as random variables.

Recently, statistical static timing analysis (SSTA) has been an active research area in combating the impact of process variation on chip timing. Based on SSTA, the distribution of both chip slack and test slack can easily be obtained. Moreover, the correlation between chip and test slacks can also be computed. As an example, the SSTA tools as developed in [5], [6] approximate timing slack as a linear form as

$$S = S_0 + \sum_{i=1}^{n} a_i \Delta X_i + a_R \Delta R_a, \tag{3}$$

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. This model is able to capture the impacts of chip-to-chip variations, within-chip variations, and local independently random variations.

Any slack in the form (3) is a linear combination of Gaussians and hence has a Gaussian probability distribution. The variance σ^2 of this distribution is given by

$$\sigma^2 = \sum_{i=1}^{n} a_i^2 + a_R^2. \tag{4}$$

Similarly, the joint distribution of chip and test slacks is a multivariate Gaussian distribution, whose covariance $cov(j,k)$ is computed from the sensitivities of the corresponding linear forms as

$$cov(j,k) = \sum_{i=1}^{n} a_{j,i} a_{k,i}, \tag{5}$$

where $a_{j,i}$ and $a_{k,i}$ are sensitivities to globally correlated variations of the two distributions.

Note that the techniques developed in this work is, however, not limited to any particular SSTA tools used or how we obtain the distributions.

B. Problem Formulation

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 therfore 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, the four types of chips as discussed above are evident.

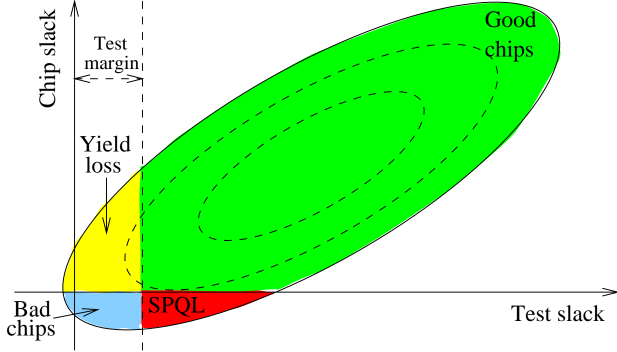

Fig. 2. Joint distribution of chip and test slack.

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 slack and chip slack helps to reduce both SPQL and yield loss.

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. This can be formally stated as follows:

Formulation 1: Given chip slack S_C and test slack S_T modeled as a joint multivariate Gaussian distribution, a required SPQL q as defined by (2), find the test margin S_M such that the yield as defined by (1) is maximized.

One testing policy is that all chips are tested with the same test margin S_M, i.e., *uniform test margin* for all chips. Another testing policy is that each chip is tested with a chip-specific test margin based on some *a priori* knowledge about each individual chip. One example of such chip-specific *a priori* knowledge is PSRO measurements. Using this additional information about each chip, we can tighten the JPDF shown in Fig. 2. The resulting conditional probability distribution has better correlation between chip and test slacks and allows us to improve testing quality. Without loss of generality, we assume that the additional chip-specific information is also represented as a slack random variable with Gaussian distribution, denoted as S_P. For example, if PSRO slack is used to provide such information, the distribution of PSRO slack can be obtain either by statistical timing of its open loop circuit or by Monte Carlo SPICE simulation.

The advantage of uniform test margin is its simple test setup and short test time, as all chips are tested the same way without need to adjust the tester. On the contrary, by leveraging knowledge about each individual chip's characteristics, a chip-specific test margin can improve yield without sacrificing SPQL. And the downside of this testing policy is the relatively longer test setup and test time. In practice, it is up to the foundry to determine which test policy works the best.

For the uniform test margin policy, the problem can be mathematically formalized as the solution of the optimization problem

$$(\mathbf{P1}) \quad \max_{S_M} P(S_T \geq S_M) \tag{6}$$

$$\text{s.t.} \quad P(S_C \leq 0 | S_T \geq S_M) \leq q, \tag{7}$$

where $P(S_T \geq S_M)$ is the probability of shipping a chip, and it is the same as yield in (1); and $P(S_C \leq 0 | S_T \geq S_M)$ is the conditional probability that a shipped chip is deficient, i.e., SPQL as defined in (2).

For the chip-specific test margin policy, we can find the optimal test margin by solving the optimization problem

$$(\mathbf{P2}) \quad \max_{S_M(S_P)} P\left(S_T \geq S_M\left(S_P\right)\right) \tag{8}$$

$$\text{s.t.} \quad P\left(S_C \leq 0 | S_T \geq S_M\left(S_P\right)\right) \leq q, \tag{9}$$

where the test margin $S_M(S_P)$ is a function of the chip-specific measurement S_P. Although the objective function and constraint of this problem look similar to those of (6) and (7), the difference is the dependence of test margin on measurement S_P, which 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. And both the objective function and constraint are functionals in **(P2)**.

IV. Uniform Test Margin

A. *Conservative uniform test margin*

Before we solve (**P1**) optimally, we present a fast approach to solve (**P1**) conservatively. It is well known that any feasible solution to a maximization problem is a lower bound of the optimal solution. As (**P1**) has only one constraint, we provide an analytic solution to satisfy this constraint, thus providing a lower bound solution.

We first decompose chip slack into a linear combination of a part that is correlated to test slack, and a part that is uncorrelated, i.e.,

$$S_C = \alpha S_T + S_u, \tag{10}$$

where S_u is uncorrelated with test slack S_T. As chip slack S_C and test slack S_T are positively correlated in practice, it makes sense to assume that α is a positive constant. The decomposition as shown in (10) is always

possible given S_C and S_T are Gaussian with distributions $N(\mu_C, \sigma_C^2)$ and $N(\mu_T, \sigma_T^2)$, respectively, and their correlation coefficient ρ. We can show that

$$\begin{aligned}\alpha &= \rho\frac{\sigma_C}{\sigma_T},\\ \sigma_u^2 &= \sigma_C^2 - \alpha^2\sigma_T^2,\\ \mu_u &= \mu_C - \alpha\mu_T.\end{aligned}$$

Then SPQL can then be expressed as

$$\begin{aligned}&P(S_C \leq 0|S_T \geq S_M)\\ =&\frac{P(S_C \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)}\\ =&\frac{P(\alpha S_T + S_u \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)}\\ \leq&\frac{P(\alpha S_M + S_u \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)}\\ =&\frac{P(\alpha S_M + S_u \leq 0)P(S_T \geq S_M)}{P(S_T \geq S_M)}\\ =&P(\alpha S_M + S_u \leq 0).\end{aligned} \tag{11}$$

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

$$S_M = \frac{-1}{\alpha}\left(\sigma_u \Phi^{-1}(q) + \mu_u\right), \tag{12}$$

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

B. Exact uniform test margin

In order to solve the uniform test margin problem (**P1**) optimally, we first prove two monotocity properties on yield and SPQL as a function of test margin S_M.

Theorem 1: The yield $P(S_T \geq S_M)$, i.e., probability of shipping a chip, is a monotonically decreasing function of test margin S_M, therefore it reaches its maximum at the minimum allowed value S_M.

Proof: According to the definition of yield, it can be written as

$$P(S_T \geq S_M) = \int_{S_M}^{\infty} p_t(S_T)dS_T, \tag{13}$$

where $p_t(S_T)$ is the PDF of S_T. As $p_t(S_T)$ is positive and does not depend on S_M, it is trivial to see that this probability is a monotonically decreasing function of test margin S_M. $\square$

Theorem 2: The SPQL $P(S_C \leq 0|S_T \geq S_M)$, i.e., the conditional probability of a defective shipped chip, is a monotonic function of test margin S_M, provided that the correlation coefficient ρ between chip and test slacks is not 0. When $\rho > 0$, SPQL is a monotonically decreasing function of test margin S_M; and When $\rho < 0$, SPQL is a monotonically increasing function of test margin S_M.

Proof: We denote $Q = P(S_C \leq 0|S_T \geq S_M)$ as SPQL, and the proof of SPQL monotocity is equivalent to proving the derivative of Q greater or smaller than 0, i.e., $Q' = \frac{dQ}{dS_M}$. According to the definition of SPQL, we have

$$\begin{aligned}Q &= P(S_C \leq 0|S_T \geq S_M)\\ &= \frac{P(S_C \leq 0, S_T \geq S_M)}{P(S_T \geq S_M)}.\end{aligned}$$

For simplicity, we denote

$$\begin{aligned}A &= P(S_C \leq 0, S_T \geq S_M),\\ B &= P(S_T \geq S_M).\end{aligned}$$

Hence the derivative of SPQL with respect to S_M is

$$Q' = \frac{dQ}{dS_M} = \frac{A'B - AB'}{B^2} = \frac{A'}{B^2}\left(B - \frac{AB'}{A'}\right). \tag{14}$$

It is easy to show that

$$\begin{aligned}A &= \int_{S_M}^{\infty}\int_{-\infty}^{0} p_c(S_C, S_T)dS_C dS_T\\ &= \int_{S_M}^{\infty}\left(p_t(S_T)\int_{-\infty}^{0} p_c(S_C|S_T)dS_C\right)dS_T\end{aligned} \tag{15}$$

$$B = \int_{S_M}^{\infty} p_t(S_T)dS_T \tag{16}$$

$$A' = \frac{dA}{dS_M} = -p_t(S_M)\int_{-\infty}^{0} p_c(S_C|S_M)dS_C \tag{17}$$

$$B' = \frac{dB}{dS_M} = -p_t(S_M), \tag{18}$$

where $p_c(S_C, S_T)$ is the JPDF of S_C and S_T as shown in Fig. 2; $p_c(S_C|S_T)$ 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 $p_c(S_C|S_T = S_M)$.

It is clear from (17) that $A' < 0$. From (14) with the consideration of $B^2 > 0$, and $A' < 0$, we can see that proving $Q' > 0$ is equivalent to proving $B - \frac{AB'}{A'} < 0$, and proving $Q' < 0$ is equivalent to proving $B - \frac{AB'}{A'} > 0$.

Using the formula for a conditional PDF, we get

$$\frac{AB'}{A'} = \int_{S_M}^{\infty} p_t(S_T)\frac{\int_{-\infty}^{0} p_c(S_C|S_T)dS_C}{\int_{-\infty}^{0} p_c(S_C|S_M)dS_C}dS_T. \tag{19}$$

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

$$\int_{-\infty}^{0} p_c(S_C|S_T)dS_C = \Phi\left(-\frac{\mu_C + \rho\frac{\sigma_C}{\sigma_T}(S_T - \mu_T)}{\sigma_C\sqrt{(1-\rho^2)}}\right) \tag{20}$$

$$\int_{-\infty}^{0} p_c(S_C|S_M)dS_C = \Phi\left(-\frac{\mu_C + \rho\frac{\sigma_C}{\sigma_T}(S_M - \mu_T)}{\sigma_C\sqrt{(1-\rho^2)}}\right) \tag{21}$$

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

The two Gaussian integrals (20) and (21) 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 correlation coefficient ρ between test and chip slacks as follows:

$$0 < \frac{\int_{-\infty}^{0} p_c(S_C|S_T)dS_C}{\int_{-\infty}^{0} p_c(S_C|S_M)dS_C} \begin{cases} < 1 & \text{if } \rho > 0 \\ = 1 & \text{if } \rho = 0 \\ > 1 & \text{if } \rho < 0 \end{cases} \tag{22}$$

Since the integrand of (19) is the integrand of B in (16) multiplied by the ratio (22), we have

$$\frac{AB'}{A'} \begin{cases} < B & \text{if } \rho > 0 \\ = B & \text{if } \rho = 0 \\ > B & \text{if } \rho < 0 \end{cases} \tag{23}$$

In other words, $B - \frac{AB'}{A'}$ in (14) is positive when ρ is positive; and $B - \frac{AB'}{A'}$ is negative when ρ is negative.

Thus, the SPQL derivative in (14) is negative when ρ is positive; and the SPQL derivative is positive when ρ is negative. SPQL is therefore a monotonic function of the test margin, which completes our proof. □

Based on Theorem 1 and Theorem 2, we can make the following conclusion about the optimal uniform test margin

Theorem 3: The optimal uniform test margin to the problem (**P1**) can be computed from the constraint

$$P(S_C \leq 0|S_T \geq S_M) = q. \tag{24}$$

Proof: The proof is trivial as both the objective function (6) and constraint (7) are monotonic functions of S_M based on Theorem 1 and Theorem 2, respectively. □

According to Theorem 3, the optimal uniform test margin is obtained by solving (24). Rewriting (24), we have

$$\int_{S_M}^{\infty}\int_{-\infty}^{0} p_c(S_C, S_T)dS_C dS_T = q\int_{S_M}^{\infty} p_t(S_T)dS_T. \tag{25}$$

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 efficiently through any one-dimensional root finding algorithm.

V. Per-Chip Test Margin

In this section, we take advantage of the *a priori* knowledge about some measurements of each chip in order to determine chip-specific optimal test margins. Without loss of generality, we assume in this section that these *a priori* measurements are given as PSRO measurement. The margin computation is specific to each chip based on its PSRO measurements. As discussed before, we assume that the measured PSRO frequency is converted to a timing slack modeled as a Gaussian random variable S_P, either by SSTA or linear regression of Monte Carlo results.

As shown in section III, the chip-specific test margin $S_M(S_P)$ as formulated in (**P2**) is a functional. By rewriting the problem in integrals, we have

$$\max_{S_M(S_P)} \int_{-\infty}^{\infty}\int_{S_M(S_P)}^{\infty} p_t(S_T, S_P)dS_T dS_P \tag{26}$$

$$\text{s.t.} \frac{\int_{-\infty}^{\infty}\int_{S_M(S_P)}^{\infty}\int_{-\infty}^{0} p_c(S_C, S_T, S_P)dS_C dS_T dS_P}{\int_{-\infty}^{\infty}\int_{S_M(S_P)}^{\infty} p_t(S_T, S_P)dS_T dS_P} \leq q. \tag{27}$$

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 fact that (in non-degenerate cases) the optimal solution is achieved when SPQL is exactly as required (similar to Theorem 3). For brevity, we do not consider here the trivial case when the probability of manufacturing a bad chip is less than the required SPQL. Therefore, the constraint (27) is transformed to

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

Then problem (**P2**) becomes a *conditional variational optimization problem* [9] in the form of

$$\max_{y(x)} \int_{-\infty}^{\infty} F(x, y(x))dx \tag{28}$$

$$\text{s.t.} \int_{-\infty}^{\infty} G(x, y(x))dx = 0. \tag{29}$$

From variational calculus, it is known that the optimal solution to the above problem can be obtained through the Lagrangian method as

$$\begin{aligned} &\max_{y(x)} \int_{-\infty}^{\infty} F(x, y(x))dx - \lambda\int_{-\infty}^{\infty} G(x, y(x))dx \\ &= \int_{-\infty}^{\infty} \left(F(x, y(x)) - \lambda G(x, y(x))\right)dx \\ &= \int_{-\infty}^{\infty} H(x, y(x))dx, \end{aligned} \tag{30}$$

where λ is the Lagrange multiplier.

According to [9], for the functional (30) to reach its optimum, it requires $y(x)$ to satisfy the condition as

$$\frac{\partial H(x,y)}{\partial y} = 0. \quad (31)$$

By using the same notation as in (**P2**) and through inspection, we can easily write the Lagrangian of (**P2**) as

$$\int_{-\infty}^{\infty}\int_{S_M(S_P)}^{\infty}\Big((1+\lambda q)p_t(S_T,S_P) - \lambda\int_{-\infty}^{0} p_c(S_C,S_T,S_P)dS_C\Big)dS_T dS_P. \quad (32)$$

Through the optimality condition (31), we obtain the following equation for $S_M(S_P,\lambda)$:

$$(1+\lambda q)p_t(S_M,S_P) - \lambda\int_{-\infty}^{0} p_c(S_C,S_M,S_P)dS_C = 0. \quad (33)$$

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

$$\int_{-\infty}^{0} p_c(S_C|S_M,S_P)dS_C = q + \frac{1}{\lambda}. \quad (34)$$

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

$$S = \begin{pmatrix} S_C \\ S_T \\ S_P \end{pmatrix} = \begin{pmatrix} S_C \\ S_{TP} \end{pmatrix} \quad (35)$$

$$\mu = \begin{pmatrix} \mu_C \\ \mu_T \\ \mu_P \end{pmatrix} = \begin{pmatrix} \mu_C \\ \mu_{TP} \end{pmatrix} \quad (36)$$

$$\Sigma = \begin{pmatrix} \sigma_C^2 & \rho_{C,T} & \rho_{C,P} \\ \rho_{C,T} & \sigma_T^2 & \rho_{T,P} \\ \rho_{C,P} & \rho_{T,P} & \sigma_P^2 \end{pmatrix} = \begin{pmatrix} \sigma_C^2 & \rho_{C,TP} \\ \rho_{C,TP}^T & \Sigma_{TP} \end{pmatrix} \quad (37)$$

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

$$\hat{\mu}_c = \mu_C + \rho_{C,TP}\Sigma_{TP}^{-1}(S_{MP} - \mu_{TP}) \quad (38)$$

$$\hat{\sigma}_c^2 = \sigma_C^2 - \rho_{C,TP}\Sigma_{TP}^{-1}\rho_{C,TP}^T \quad (39)$$

where according to equation (35)

$$S_{MP} = \begin{pmatrix} S_M(S_P) \\ S_P \end{pmatrix}. \quad (40)$$

Performing integration of the Gaussian PDF in (34) and solving, we get

$$\hat{\mu}_c = -\hat{\sigma}_c\Phi^{-1}(q + 1/\lambda), \quad (41)$$

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

Substituting expressions for $\rho_{C,TP}$, Σ_{TP}, S_{MP} and μ_{TP} into (38) and performing matrix-vector multiplication we can show that

$$\hat{\mu}_c = \mu_C + \alpha S_M(S_P) + \beta S_P - \alpha\mu_T - \beta\mu_P \quad (42)$$

where α and β are expressed through variances and covariances of the test, chip and PSRO slacks. Excluding $\hat{\mu}_c$ from (41) and (42), 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

$$S_M(S_P) = \gamma S_P + \eta. \quad (43)$$

The Lagrange multiplier λ can easily be found by computing η. By changing the order of integration in the numerator of (27) 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

$$S_P = \frac{u-\gamma v}{\sqrt{1+\gamma^2}} \qquad S_T = \frac{\gamma u + v}{\sqrt{1+\gamma^2}} \quad (44)$$

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

$$\frac{\int_{-\infty}^{0}\int_{\frac{\eta}{\sqrt{1+\gamma^2}}}^{\infty}\int_{-\infty}^{\infty} p_c(S_C, \frac{\gamma u+v}{\sqrt{1+\gamma^2}}, \frac{u-\gamma v}{\sqrt{1+\gamma^2}})du dv dS_C}{\int_{\frac{\eta}{\sqrt{1+\gamma^2}}}^{\infty}\int_{-\infty}^{\infty} p_t(\frac{\gamma u+v}{\sqrt{1+\gamma^2}}, \frac{u-\gamma v}{\sqrt{1+\gamma^2}})du dv} = q. \quad (45)$$

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 (45) 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 (43), we get the optimal value of the test margin. Thus, by combining analytical and numerical methods we can determine the optimal test policy.

Equation (43) 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 (26). This integral is exactly the denominator of (27) 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.

VI. Numerical results

We validate our proposed techniques by using two industrial 90 nm ASICs, each with over a million placeable objects. The distributions 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 uniform margin, optimal uniform margin, and optimal per-chip margin.

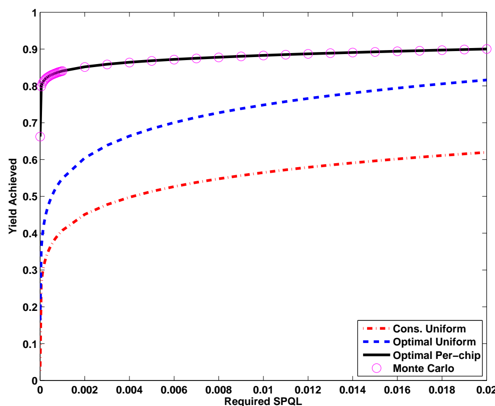

Fig. 3. Achieved yield versus required SPQL.

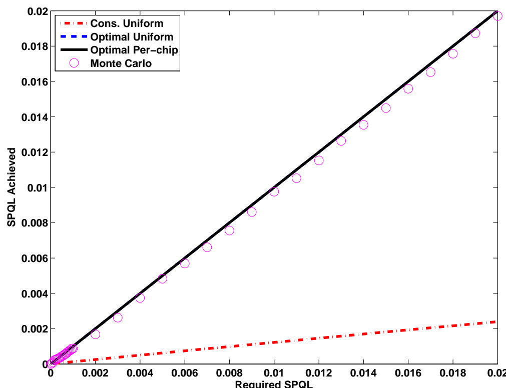

Fig. 4. Achieved SPQL versus required SPQL.

Fig. 3 shows the comparison of yield achieved under different test margin polices (and Monte Carlo simulation results that will be explained later) for different values of required SPQL shown on the x-axis. From the figure, 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. Particularly for small SPQL values, per-chip margins result in a significant yield improvement. This is expected because the former two policies are optimal in leveraging SPQL to the fullest to maximize yield, and it explains why both polices meet the required SPQL almost exactly as shown in Fig. 4, in which we compare the achieved SPQL and the required SPQL for these three methods. From Fig. 4, we also observe that, 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.

Fig. 5 and Fig. 6 compare test margins under different required SPQL among the three policies. Fig. 5 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 Fig. 6, where policies corresponding to four different required SPQL values are shown with the PSRO measurements on the x-axis. 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.

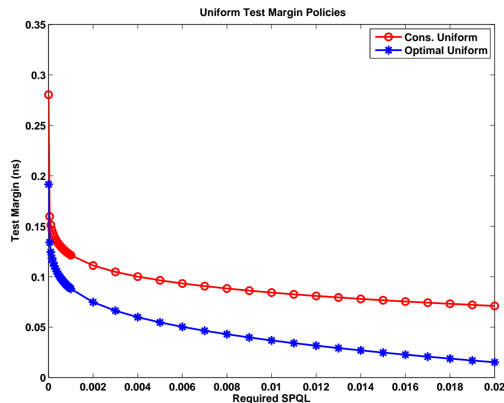

Fig. 5. Test margins versus required SPQL for uniform margins.

To further verify our theoretical derivation, we apply

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%)

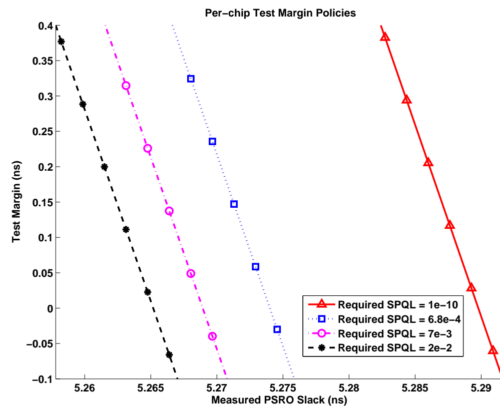

Fig. 6. Test margins versus required SPQL for per-chip margins.

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 and Fig. 4, 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. For example, for design one, under the required SPQL as 1.0e-6 (i.e., 0.0001%), by moving from the conservative to the optimal uniform policy, we see that the achieved yield increases from 14% to 20%; while by moving from the optimal uniform policy to a per-chip policy, we see further yield increases by as much as 55%. For different designs, the most significant gains in yield is achieved at the highest quality levels (lowest SPQL requirement). Monte Carlo simulation matches our theoretical results very closely across the board, thus demonstrating the value and correctness of the proposed techniques.

VII. 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 is 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. Degregorio, 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.