Statistical Timing Yield Optimization by Gate Sizing

Debjit Sinha, *Member, IEEE*, Narendra V. Shenoy, *Senior Member, IEEE*, and Hai Zhou, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a statistical gate sizing approach to maximize the timing yield of a given circuit, under area constraints. Our approach involves statistical gate delay modeling, statistical static timing analysis, and gate sizing. Experiments performed in an industrial framework on combinational International Symposium on Circuits and Systems (ISCAS'85) and Microelectronics Center of North Carolina (MCNC) benchmarks show absolute timing yield gains of 30% on the average, over deterministic timing optimization for at most 10% area penalty. It is further shown that circuits optimized using our metric have larger timing yields than the same optimized using a worst case metric, for iso-area solutions. Finally, we present an insight into statistical properties of gate delays for a commercial 0.13-μm technology library which intuitively provides one reason why statistical timing driven optimization does better than deterministic timing driven optimization.**

***Index Terms*—Gate sizing, optimization, statistical gate delay modeling, statistical timing analysis, timing yield, variability, VLSI.**

I. Introduction

An increasing significance of variability in modern deep submicrometer integrated circuits necessitates statistical approaches to timing analysis and optimization. Researchers have proposed multiple approaches to statistical static timing analysis [2]–[6] in the past few years. A majority of these approaches consider circuit component delays as Gaussian random variables since it facilitates fast analytical evaluation. Timing analysis involves add and max operations. A max operation on Gaussian random variables is nontrivial. Chang *et al.* [3] and Visweswariah *et al.* [5] propose to approximate the maximum of multiple Gaussians with a Gaussian using Clark's approach [7] to obtaining the max of two Gaussians. Pairwise max operations are, thus, employed in the computation of the maximum of multiple Gaussians, each of which involve approximations. However, none of the above approaches describe the impact of the ordering of pairwise max operations on the resulting inaccuracy in the final solution.

Multiple approaches to statistical timing optimization have emerged recently. Agarwal *et al.* propose a sensitivity-based gate sizing algorithm, and faster approaches that perform sensitivity calculation based on slack computation [8], to minimize the 99-percentile point of a circuit's delay distribution. Intra-die variability is considered, and gate delay variations are assumed to be 10% of their nominals. A robust gate sizing methodology based on geometric programming is proposed by Singh *et al.* [9]. They incorporate an uncertainty ellipsoid to model variations and attain to optimize circuit area under worst case timing constraints. Guthaus *et al.* [10] propose a gate sizing algorithm to optimize circuit area while satisfying a given timing yield target. They employ a sensitivity metric to select gates for resizing. Our experiments conclude that node and edge criticalities evaluated in their approach can only be estimated in closed form to be within 20% of those obtained from Monte Carlo simulations. This is due to the assumption of independence between the criticalities of any two paths while evaluating a node or an edge criticality. As a result, they may be inadequate for guiding timing optimization.

In this paper, we present an approach to area constrained statistical timing yield optimization that involves statistical modeling, statistical timing analysis, and gate sizing. We do not focus on improving a given percentile point of a circuit's delay distribution, but attain to maximize the probability that given timing constraints are met, under variations. Statistical gate delay modeling is performed for a commercial 0.13-μm technology library from a foundry. We employ Visweswariah's approach [5] for statistical static timing analysis, and present a formal proof that validates their variance matching methodology used in the computation of the maximum of two Gaussians. We also consider a smart ordering for pairwise max operations on Gaussians during the computation of the maximum of multiple Gaussians. It is observed that the ordering achieves accuracy improvements in the final solution. Gate sizing is performed using a statistical global sizing algorithm. We prove that maximizing the timing yield of a circuit is equivalent to maximizing a simple expression involving the mean and the standard deviation of the circuit's slack distribution. Experiments performed in an industrial framework show absolute timing yield gains of 30% on the average in comparison to a commercial synthesis tool for an area overhead of at most 10%. We observe that for iso-area solutions, our metric obtains larger timing yields than optimization for the worst case slack. Finally, we present insight into statistical properties of gate delays from a commercial technology library which intuitively provides one reason why statistical timing driven optimization does better than deterministic timing driven optimization.

The rest of this chapter is organized as follows. Sections II and III present our approaches to statistical modeling and statistical static timing analysis, respectively. In Section IV, we propose our statistical gate-sizing algorithm for timing yield optimization, and present experimental results in Section V. We provide insight into statistical properties of gate delays in Section VI, and draw conclusions in Section VII.

Manuscript received November 17, 2005; revised March 28, 2006. This work was supported in part by the National Science Foundation under Grant CCR-0238484. This paper is an extended version of the authors' work in the Proceedings International Conference of Computer-Aided Design, pp. 1037–1041, 2005.

D. Sinha and H. Zhou are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: haizhou@eecs.northwestern.edu).

N. V. Shenoy is with the Advanced Technology Group, Synopsys Inc., Mountain View, CA 94043 USA.

Digital Object Identifier 10.1109/TVLSI.2006.884166

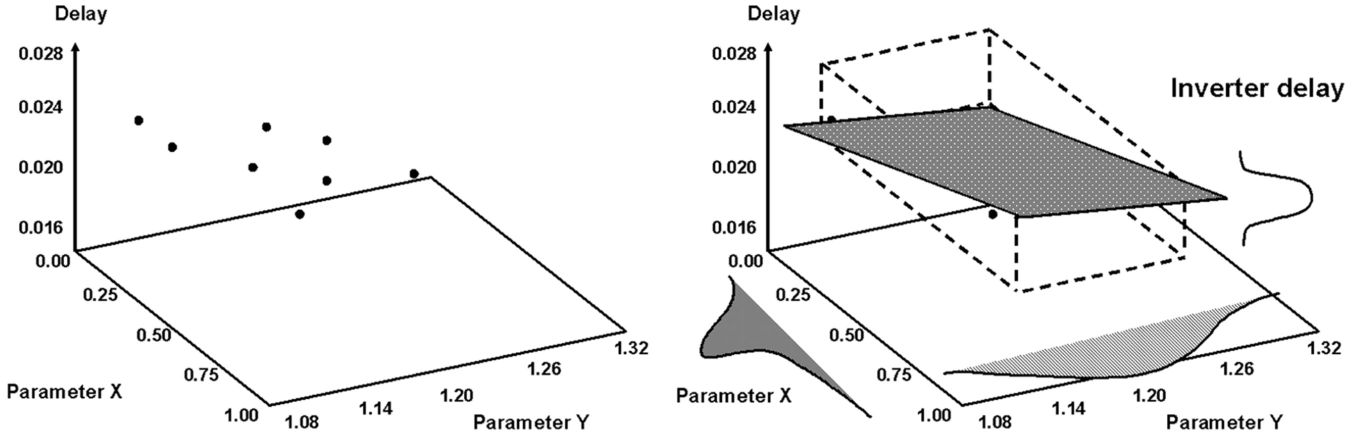

Fig. 1. Statistical inverter delay modeling example.

II. Statistical Modeling

Statistical delay modeling involves expressing circuit component delays as functions of the parameters of variation, which we model as Gaussian random variables. Based on the work in [3] and [5], we assume that gate delays are approximated by a linear function of the parameters. We also assume that these parameters are independent, since a dependent set of Gaussian parameters can be transformed into an equivalent set of independent Gaussian parameters using principal component analysis [3]. Circuit component delays are, therefore, expressed as

$$a_0 + \sum_{i=1}^{n} a_i \Delta X_i + a_{n+1} \Delta R_a. \tag{1}$$

In the previous expression, a_0 denotes the mean or nominal value of the delay, ΔX_i's $(i = 1, 2, \ldots, n)$ represent the variations of n global parameters X_i's $(i = 1, 2, \ldots, n)$ from their nominal values, and a_i's and $(i = 1, 2, \ldots, n)$ denote the delay sensitivities to their corresponding sources of variation. ΔR_a represents the variation from the nominal of an independent random variable R_a that is associated with each component, and a_{n+1} denotes the delay sensitivity to R_a.

To compute the delay sensitivities for any gate in the circuit, we obtain precharacterized gate delay values as functions of their loading capacitance and input slews (based on deterministic timing analysis at nominal corner) at multiple corners in the parameter space. The parameters are normalized by subtracting their nominal values followed by a division by their standard deviations. A least-squares fit is finally employed to obtain the desired delay sensitivities that express the gate delay as a linear function of normal random variables, as expressed in (1). This procedure is repeated for each gate in the circuit. Fig. 1 shows precharacterized delay values for some inverter in a circuit at multiple corners in a 2-D parameter space. A least square fit of the obtained points results in a plane, the slope of which in the two coordinate directions give the sensitivities of the inverter delay to the parameters, respectively. The inverter delay is, thus, obtained as a weighted linear sum of Gaussian random variables.

III. Statistical Static Timing Analysis

Statistical static timing analysis requires propagation of delay distributions through the circuit. This involves add and max operations on the delay random variables. Since we express circuit component delays as a linear combination of Gaussian random variables, the add operation is performed in a straight forward manner and yields another Gaussian. In this work, we employ Visweswariah's approach [5] to computing the maximum of two Gaussian delay random variables A and B, which are expressed as a weighted linear sum of normal random variables as in (1). We denote the (mean, variance) of A and B as $(a_0, \sigma_A^2 = \sum_{i=1}^{n+1} a_i^2)$ and $(b_0, \sigma_B^2 = \sum_{i=1}^{n+1} b_i^2)$, respectively, where the a_i's and b_i's represent delay sensitivities. We use $\rho = (\sum_{i=1}^{n} a_i b_i)/(\sigma_A \sigma_B)$ to denote the correlation coefficient between A and B, and define the following:

$$\phi(x) \triangleq \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \tag{2}$$

$$\Phi(y) \triangleq \int_{-\infty}^{y} \phi(x) dx \tag{3}$$

$$\theta \triangleq (\sigma_A^2 + \sigma_B^2 - 2\rho\sigma_A\sigma_B)^{1/2} \tag{4}$$

$$\alpha \triangleq \frac{a_0 - b_0}{\theta}. \tag{5}$$

The mean $\mu_{\max}$ and variance $\sigma_{\max}^2$ of $\max(A, B)$ are computed as follows (Clark's approach [7]):

$$\mu_{\max} = a_0\Phi(\alpha) + b_0\Phi(-\alpha) + \theta\phi(\alpha) \tag{6}$$

$$\begin{aligned}\sigma_{\max}^2 = {} & \left(\sigma_A^2 + a_0^2\right)\Phi(\alpha) + \left(\sigma_B^2 + b_0^2\right)\Phi(-\alpha) \\ & + (a_0 + b_0)\theta\phi(\alpha) - \mu_{\max}^2.\end{aligned} \tag{7}$$

Approximation of $\max(A, B)$ with a Gaussian C having a canonical form $(c_0 + \sum_{i=1}^{n} c_i \Delta X_i + c_{n+1} \Delta R_c)$ is performed as follows (Visweswariah's approach [5]):

$$c_0 = \mu_{\max} \tag{8}$$

$$c_i = a_i\Phi(\alpha) + b_i[1 - \Phi(\alpha)] \qquad \forall i \in 1, 2, \ldots, n \tag{9}$$

$$c_{n+1} = \left[\sigma_{\max}^2 - \sum_{i=1}^{n} c_i^2\right]^{1/2}. \tag{10}$$

$\Phi(\alpha)$, in the previous expression, denotes the tightness probability of A over B, that is, the probability that A dominates B. Our first contribution to this approach is that we formally validate the variance matching approach in (10). We prove in the appendix that $(\sigma_{\max}^2 - \sum_{i=1}^{n} c_i^2)$ is always nonnegative. This implies that the variance matching approach never involves the computation of the square root of a negative quantity. Required time estimation in statistical timing analysis is performed by a backward propagation of delay distributions and involves the subtract and min operations. These operations are similar to the add and max operations.

When a gate has more than two fan-ins (fan-outs), the max (min) operation for the arrival (required) time distribution calculation is done one pair at a time, each step of which involves approximations. We observe that an arbitrary order of these pairwise operations may accumulate errors and can significantly affect the accuracy of the final solution. We employ a greedy approach for smart pairwise max (min) operations based on the approximation error computations [11]. Slack estimation during timing analysis involves subtract operations which can be performed on the canonical forms of the timing distributions. A min operation on the slack distributions at the primary outputs gives the circuit slack.

IV. Statistical Gate Sizing

We formally define the timing yield of a circuit to be the probability that the circuit slack is nonnegative. This probability can be computed by integrating the slack probability density function (pdf) from 0 to ∞. Given the circuit slack (after statistical timing analysis) as a Gaussian random variable S with mean μ_S and standard deviation σ_S, the timing yield P of the circuit is given by

$$P \triangleq \frac{1}{\sqrt{2\pi}\sigma_S}\int_0^{\infty} \exp\left[-\frac{(x-\mu_S)^2}{2\sigma_S^2}\right] dx. \tag{11}$$

In this work, we attain to maximize the timing yield of a circuit using gate sizing, under given area constraints. We next prove that maximizing the timing yield is equivalent to maximizing the ratio of the mean to the standard deviation of the circuit slack distribution.

Theorem 1:

$$\max \frac{1}{\sqrt{2\pi}\sigma_S}\int_0^{\infty} \exp\left[-\frac{(x-\mu_S)^2}{2\sigma_S^2}\right] dx \equiv \max \frac{\mu_S}{\sigma_S}.$$

Proof: We define $y \triangleq (\mu_S - x)/(\sigma_S)$. Under variable transformation

$$\frac{1}{\sqrt{2\pi}\sigma_S}\int_0^{\infty} \exp\left[-\frac{(x-\mu_S)^2}{2\sigma_S^2}\right] dx = \frac{1}{\sqrt{2\pi}} - \int_{\mu_S/\sigma_S}^{-\infty} \exp\left[-\frac{y^2}{2}\right] dy$$
$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\mu_S/\sigma_S} \exp\left[-\frac{y^2}{2}\right] dy$$

which is strictly increasing with $\frac{\mu_S}{\sigma_S}$. This proves our claim. ■

Our statistical gate sizing approach, thus, attains to maximize the metric (μ_S/σ_S), under area constraints. For sake of comparison, we also consider maximizing the metric $(\mu_S - 3\sigma_S)$, under identical area constraints; such an objective function attains to maximize the worst case slack.

We design a statistical global gate sizing (SGGS) algorithm for timing yield optimization as an extension to the global gate sizing algorithm [12]. Our choice of the global sizing algorithm is motivated by results obtained by Coudert *et al.* [12], which show that this algorithm is superior to common greedy or genetic approaches to circuit optimization in terms of performance and power/delay curves. The proposed algorithm considers the circuit as a network N of nodes with a global cost function *Cost* that is to be maximized under given area constraints $Area_{\max}$. The global cost function used in our approach is the metric (μ_S/σ_S), where μ_S and σ_S denote the mean and the standard deviation, respectively, of the circuit slack distribution S. Each node in the network is implemented using some gate from the given technology library. Multiple gates, each belonging to the same gate class as the node, can be mapped to a given node. We refer to this process as resizing a node. The variation in the global cost due to resizing a node is denoted its gradient for a particular resize operation. We define the local cost of a node as the ratio of the mean to the standard deviation of the slack distribution at its output. Variations in the local cost of a node due to various resizing operations are termed as corresponding local gradients.

We describe the algorithmic flow next. A set *update* maintains a list of nodes whose gradients are to be computed. This set is initialized with all nodes in N. Another set *moves* maintains a list of nodes that can potentially be resized. This set is initially kept empty. For any node n, the gradient computation for each possible resize involves a run of statistical timing analysis on the entire circuit. This makes the gradient evaluation computationally very expensive. In practice, we observe that the impact of a node resize on the local gradients decrease quickly (approximately geometrically [12]) with increasing fan-in and fan-out level. We, therefore, extract a subnetwork N_n for each node n in *update*, which is made out of two transitive levels of fan-in and fan-out around n. The inputs and outputs of the subnetwork N_n are annotated with the corresponding arrival and required time distributions, respectively, from the original network N. Statistical timing analysis is now performed on N_n and the local gradient at the output of this subnetwork is used as the metric for evaluation. Unless no possible resize operation on n improves this metric, the new gate involved in a possible resize that maximizes this metric is termed as the *best-gate* for n. The node n and its *best-gate* are now added as a possible resize operation to the set *moves*. However, the resize is not actually performed at this stage.

Following the above procedure for each node in the set *update*, a *MultiMove* routine picks a subset of possible resize operations from the set *moves* that provide maximum cumulative gain in the global cost. These resize operations are then performed and the resized nodes are returned in a new set *moved*. The *MultiMove* routine determines the subset for the move based on the descent direction or by a conjugation of directions of the cost gradients [12]. In our experiments, we employ a greedy heuristic that chooses the best two nodes for resize in terms of yield improvement in each *MultiMove* operation. A new set of nodes whose gradients need to be recomputed

```
• Algorithm: SGGS($N$, $Cost$, $Area_{max}$)
• Output: Circuit with improved timing yield
• begin
    $update$ = $N$;
    $moves$ = $\emptyset$;
    do {
      $old\_cost = Cost(N)$;
      foreach $n \in update$ {
        extract sub-network $N_n$ around $n$;
        find best-gate $g$ for $n$ wrt $Cost(N_n)$;
        if $g \neq gate(n)$ {
          $n.move = g$;
          $moves = moves \cup \{n\}$;
        }
      }
      $moved = MultiMove(N, Cost, moves)$;
      $update = PerturbedNodes(moved)$;
    } until ($Converge(old\_cost, Cost(N), moved)$
      $\vee$ ($Area \geq Area_{max}$ ))
• end
```

Fig. 2. SGGS algorithm.

are now derived from *moved* in the function *PerturbedNodes*. In our approach, we choose a node for gradient recomputation only if it is sufficiently perturbed, that is, if one of its close neighbors (within one or two transitive fan-in or fan-out levels) has been resized. The entire process is repeated till convergence, wherein future iterations do not improve the global cost (timing yield of the circuit) further or till the runtime/area constraints of the design are violated. For comparison, this procedure is repeated starting with the original design, using $(\mu_S - 3\sigma_S)$ as both the global cost function and the local cost function. The complexity of this algorithm using the best-fit polynomial is shown to be $k^{1.2}$, where k denotes the number of internal nodes [13]. The pseudocode of the SGGS algorithm is presented in Fig. 2.

V. Implementation and Experimental Results

The proposed statistical modeling, statistical timing analysis, and gate sizing routines are implemented in an industrial framework, as an addition to a commercial synthesis and optimization tool. Experiments are performed on combinational International Symposium on Circuits and Systems (ISCAS'85) and Microelectronics Center of North Carolina (MCNC) benchmarks mapped to a 0.13-μm commercial technology library from a foundry.

For our experiments, we choose V_{dd} and temperature as the parameters of variation. We acknowledge that these parameters may have a nonlinear impact on delays. However, precharacterized gate delay values were available for a commercial 0.13-μm library that we intended to use in our experiments. It was not immediately possible to recharacterize these gates for other parametric variations, and we did not use artificial values for the same as done in a majority of other mentioned approaches to statistical optimization. In any case, our approach is not limited to the use of any particular parameters of variation.

TABLE I
Statistical Timing Analysis Results

Circuit	AT μ (ns)		AT σ (ns)		Run time (s)	
	SSTA	MC	SSTA	MC	SSTA	MC
C432	2.194	2.197	0.165	0.172	0.1	6.5
C499	1.316	1.311	0.095	0.095	1.2	14.6
C880	1.973	1.968	0.143	0.143	0.4	14.0
C1355	1.829	1.831	0.141	0.139	0.8	20.4
C1908	2.208	2.214	0.161	0.160	0.7	14.3
C2670	1.950	1.957	0.177	0.174	1.5	24.7
C3540	3.242	3.234	0.261	0.260	0.8	37.7
C5315	3.029	3.024	0.246	0.242	7.3	63.0
C6288	9.996	9.968	0.779	0.789	0.7	85.1
C7552	3.313	3.305	0.261	0.254	5.1	71.9
sct	0.485	0.484	0.030	0.030	0.1	2.8
alu2	2.584	2.590	0.211	0.213	0.2	12.9
too_large	1.048	1.047	0.071	0.073	0.1	13.6
frg2	1.486	1.490	0.097	0.098	1.3	29.3

We consider V_{dd} variations in the range of 1.08 to 1.32 V. The nominal value V_0 is set to 1.2 V and the standard deviation is set as the following:

$$3\sigma_V = 1.32\ \mathrm{V} - 1.20\ \mathrm{V}.$$

Similarly, we consider temperature variations from 0 to 125 °C, with nominal temperature T_0 as 25 °C and standard deviation σ_T set to 8.33 °C. For any characterization point X, the delay equation is set up as the following:

$$D_X = D_0 + D_1 \frac{T_X - T_0}{\sigma_T} + D_2 \frac{V_X - V_0}{\sigma_V}.$$

D_0 represents the typical delay obtained from gate characterization at T_0 and V_0. This formulation is scalable to any number of parameters. A least squares fit procedure is employed to obtain the coefficients D_is. The accuracy of this approach is dependent on the number of characterization points that are available in the library.

Statistical timing analysis is next performed to obtain the global circuit slack distribution S, with mean μ_S and variance σ_S^2. Timing yield of the circuit is obtained from (11). Table I shows obtained statistical timing analysis results. We present obtained arrival time (AT) mean (μ) and standard deviation (σ) values, and those obtained from Monte Carlo simulations for comparison. Figures reported are for 10 000 random samples of Monte Carlo simulations. Benchmark sizes ranged from about 100 to 2000 gates. From Table I, we observe that the average and maximum error in the estimation of the mean and standard deviation of the circuit delay distribution is under 1% and 4.1%, respectively. SSTA is found to be faster than Monte Carlo simulations by 42.2× on the average.

For timing yield improvement estimation, we perform deterministic timing optimization on a given circuit using a commercial synthesis tool, which attains to improve the circuit slack under area constraints. Statistical timing analysis is then performed to obtain the slack distribution at the primary output of the circuit, the mean of which we denote as μ_{Static}. We next perform statistical timing optimization using our proposed gate sizing approach to obtain a new circuit slack distribution. To

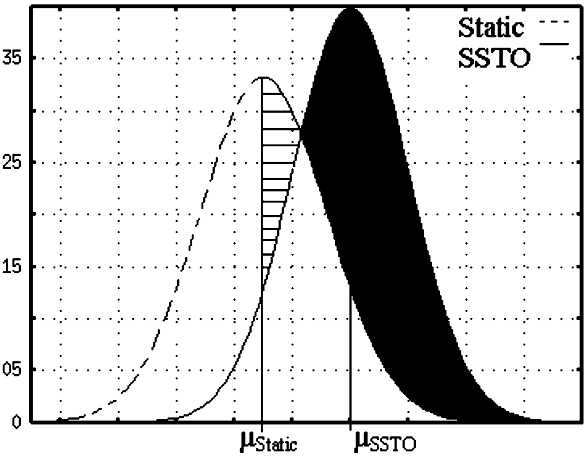

Fig. 3. Timing yield improvement denoted by area in black—area in stripes.

TABLE II
RELATIVE TIMING YIELD IMPROVEMENT RESULTS

Circuit	Statistical optimization metric	
	Maximize $\mu_S - 3\sigma_S$	Maximize μ_S/σ_S
C432	0.0105	0.0122
C499	0.3032	0.3158
C880	0.0037	0.0037
C1355	0.4939	0.4963
C1908	0.1319	0.1584
C2670	0.1730	0.2153
C3540	0.4949	0.4977
C5315	0.4923	0.4925
C7552	0.4998	0.4995
cm85a	0.0569	0.2037
sct	0.1821	0.4342
alu2	-0.0570	0.1240
too_large	0.4269	0.4580
frg2	0.4516	0.3774

estimate the relative gain in timing yield, we compute the relative timing yield of the circuit after the deterministic and statistical optimization passes as the area under their respective circuit slack PDFs from μ_{Static} to ∞. Fig. 3 shows this relative timing yield improvement graphically as the area of the black region minus the area of the striped region. We next repeat this procedure using the alternate metric $(\mu_S - 3\sigma_S)$ as the cost function during statistical optimization instead of our original metric (μ_S/σ_S).

Table II presents obtained relative timing yield improvements for both the optimization objective functions. We observe our proposed metric (μ_S/σ_S) achieves timing yield improvements of 0.3 on the average, and up to 0.5 with an area overhead of at most 10%. Corresponding average and maximum timing yield improvements using the alternate metric $(\mu_S - 3\sigma_S)$ are found to be 0.27 and 0.49, respectively (for identical area overheads). It is, thus, shown that the proposed approach guides better optimization than that for maximizing the worst case slack, under iso-area constraints. For the design *alu2*, the alternate metric worsens the yield.

We next present a special case of timing yield improvement observed for the MCNC benchmark APEX6. The three PDFs in Fig. 4 denote the slack distributions for the unoptimized circuit (*Init*), circuit following deterministic static timing optimization *(Static)* and circuit following statistical timing optimization (SSTO). The reduced variance of the SSTO slack PDF improves the timing yield (from 0.87 to 0.89) even though it has a smaller mean as compared to *Static* slack PDF. This example illustrates how statistical optimization uses the additional information on variation to achieve larger timing yields, even for iso-area solutions. The proposed algorithm takes less than 480 min for the largest benchmarks on a 400-MHz Sun Ultra 4 machine with 4-GB RAM. The primary reasons for large run times include multiple calls to statistical timing analysis that performs smart pairwise max operations [11]; and an exhaustive search of the best-gate for any node in the inner loop of the algorithm.

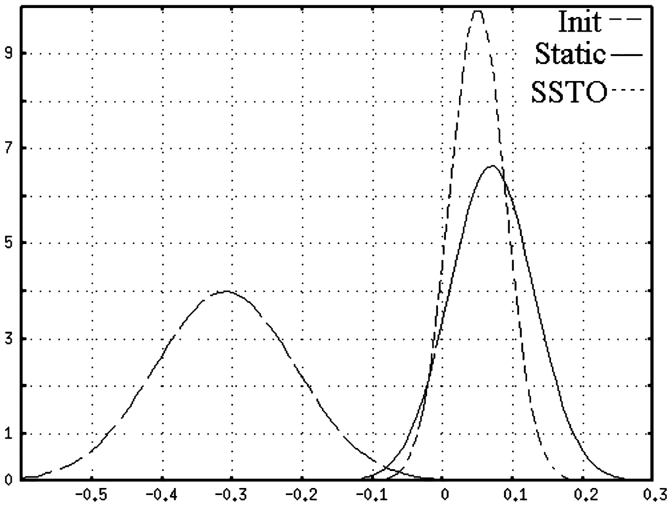

Fig. 4. Pre and post optimization slack PDFs for benchmark APEX6.

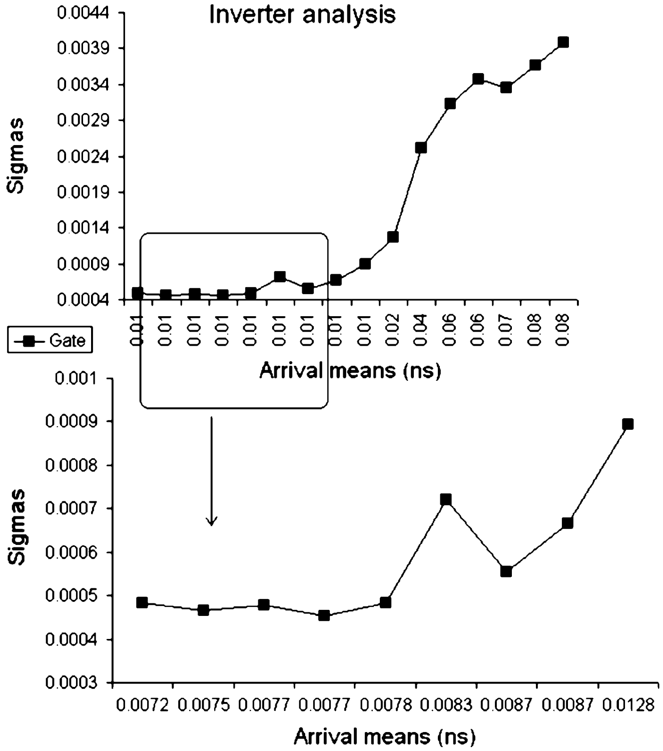

Fig. 5. Arrival means and standard deviations for a class of inverters.

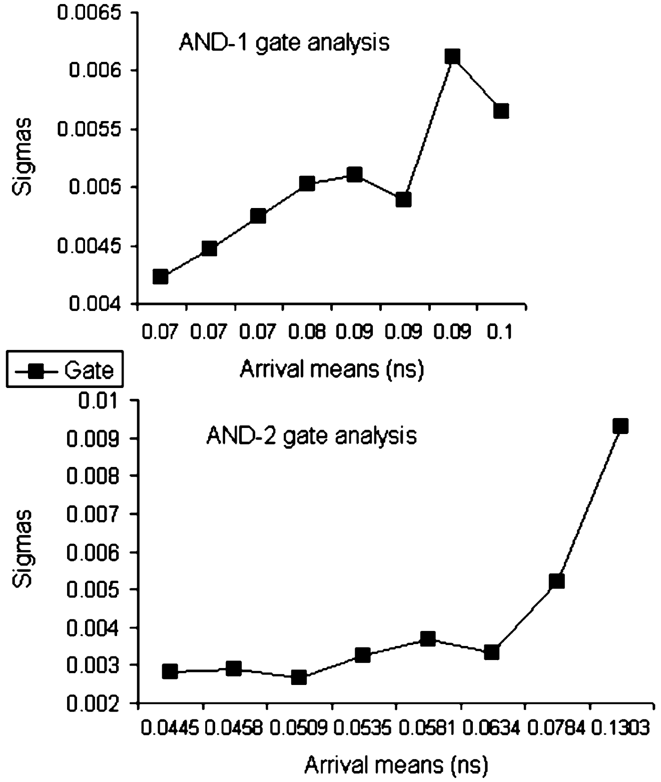

Fig. 6. Arrival means and standard deviations for two classes of AND gates.

VI. Analysis of Statistical Properties of Gate Delays

We perform an analysis of statistical properties of gate delays on different gate classes from our 0.13-μm commercial technology library. We select some nodes arbitrarily from a test circuit; and observe the mean and the standard deviation of the arrival time distribution at each of their outputs, while mapping different gates on them (the different gates belong to the gate-class of the node, for example, NAND or NOR). Fig. 5 presents a plot of the arrival time standard deviation (Sigma) against the arrival time mean for a class of inverters. Dots on the plot represent gates which are sorted on the mean of their output arrival times when mapped to the given node and not in any order of their sizes. Fig. 6 presents similar graphs for two classes of AND gates.

We observe that though most gates of a class make the plots monotonic, there exist exceptions. In some cases during our experiments, we observe that while the deterministic timing driven optimizer resizes a node to a gate with a smaller mean arrival time ignoring the fact that it may have larger variability, the statistical timing driven optimizer selects a gate with a larger mean arrival time, but a significantly lesser variance. Such a choice is found to increase the overall timing yield of the circuit. This behavior provides one reason why statistical timing driven optimization gains an edge over deterministic timing driven optimization.

VII. Conclusion

In this paper, we propose a statistical gate sizing approach to maximize the timing yield of a given circuit under area constraints. Experiments performed in an industrial framework on combinational ISCAS'85 and MCNC benchmarks show timing yield gains of 0.3 on the average, over deterministic timing optimization for at most 10% area penalty. It is further shown that circuits optimized using our metric have larger timing yields than the same optimized using a worst case metric, for iso-area solutions. Finally, we present an insight into statistical properties of gate delays for a commercial 0.13-μm technology library which intuitively provides one reason why statistical timing driven optimization does better than deterministic timing driven optimization.

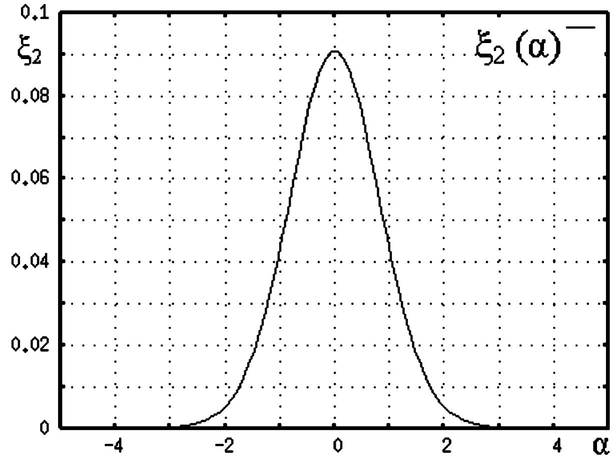

Fig. 7. Plot of ξ_2 against α.

Though this work considers delays as a weighted linear sum of Gaussian random variables, the statistical timing yield improvement approach can be extended to handle nongaussian parameters and nonlinear delay functions as proposed in [14]. However, obtaining a simple metric for timing yield optimization would be a challenging problem.

Appendix

Using notations defined in (6), (7), and (9), we prove that the variance matching method in (10) never involves the computation of the root of a negative quantity. Formally, we prove that $[\xi \triangleq \sigma_{\max}^2 - \sum_{i=1}^{n} c_i^2] \geq 0$.

Proof:

$$\begin{aligned}
\mu_{\max} &= a_0\Phi(\alpha) + b_0\Phi(-\alpha) + \theta\phi(\alpha)\\
\sigma_{\max}^2 &= \left(\sigma_A^2 + a_0^2\right)\Phi(\alpha) + \left(\sigma_B^2 + b_0^2\right)\Phi(-\alpha)\\
&\quad + (a_0 + b_0)\theta\phi(\alpha) - \mu_{\max}^2\\
\sum_{i=1}^{n} c_i^2 &= \sum_{i=1}^{n} \left[a_i\Phi(\alpha) + b_i\Phi(-\alpha)\right]^2
\end{aligned}$$

$$\begin{aligned}
\xi &= \sigma_{\max}^2 - \sum_{i=1}^{n} c_i^2\\
&= \left(\sigma_A^2 + a_0^2\right)\Phi(\alpha) + \left(\sigma_B^2 + b_0^2\right)\Phi(-\alpha)\\
&\quad + (a_0 + b_0)\theta\phi(\alpha) - a_0^2\Phi(\alpha)^2 - b_0^2\Phi(-\alpha)^2\\
&\quad - \theta^2\phi(\alpha)^2 - 2a_0\Phi(\alpha)\theta\phi(\alpha) - 2b_0\Phi(-\alpha)\theta\phi(\alpha)\\
&\quad - 2a_0b_0\Phi(\alpha)\Phi(-\alpha) - \Phi(\alpha)^2\sum_{i=1}^{n} a_i^2\\
&\quad - \Phi(-\alpha)^2\sum_{i=1}^{n} b_i^2 - 2\Phi(\alpha)\Phi(-\alpha)\sum_{i=1}^{n} a_ib_i
\end{aligned}$$

$$
\begin{aligned}
&= (\sigma_A^2 + \sigma_B^2 + a_0^2 + b_0^2)\Phi(\alpha)\Phi(-\alpha) + (a_0 + b_0)\theta\phi(\alpha) \\
&\quad - \theta^2\phi(\alpha)^2 - 2a_0\Phi(\alpha)\theta\phi(\alpha) - 2b_0\Phi(-\alpha)\theta\phi(\alpha) \\
&\quad - 2a_0b_0\Phi(\alpha)\Phi(-\alpha) + a_{n+1}^2\Phi(\alpha)^2 + b_{n+1}^2\Phi(-\alpha)^2 \\
&\quad - 2\Phi(\alpha)\Phi(-\alpha)\sum_{i=1}^{n} a_ib_i \\
&= \Phi(\alpha)\Phi(-\alpha)\left[(\sigma_A^2 + \sigma_B^2 - 2\sum_{i=1}^{n} a_ib_i) + (a_0 - b_0)^2\right] \\
&\quad + \theta\phi(\alpha)[a_0(1 - 2\Phi(\alpha)) + b_0(1 - 2 + 2\Phi(\alpha))] \\
&\quad - \theta^2\phi(\alpha)^2 + a_{n+1}^2\Phi(\alpha)^2 + b_{n+1}^2\Phi(-\alpha)^2.
\end{aligned}
$$

To show $\xi \geq 0$, it is sufficient to show that

$$
\begin{aligned}
\xi_1 &\triangleq \xi - a_{n+1}^2\Phi(\alpha)^2 - b_{n+1}^2\Phi(-\alpha)^2 \geq 0. \\
\xi_1 &= \Phi(\alpha)\Phi(-\alpha)\left[(\sigma_A^2 + \sigma_B^2 - 2\sum_{i=1}^{n} a_ib_i) + (a_0 - b_0)^2\right] \\
&\quad + \theta\phi(\alpha)[a_0(1 - 2\Phi(\alpha)) + b_0(-1 + 2\Phi(\alpha))] - \theta^2\phi(\alpha)^2 \\
&= \Phi(\alpha)\Phi(-\alpha)[\theta^2 + (a_0 - b_0)^2] - \theta^2\phi(\alpha)^2 \\
&\quad + \theta(a_0 - b_0)(1 - 2\Phi(\alpha))\phi(\alpha) \\
&= \theta^2[\Phi(\alpha)\Phi(-\alpha) - \phi(\alpha)^2] + \Phi(\alpha)\Phi(-\alpha)(a_0 - b_0)^2 \\
&\quad + \theta(a_0 - b_0)(1 - 2\Phi(\alpha))\phi(\alpha).
\end{aligned}
$$

If $\theta = 0, \xi_1 = \Phi(\alpha)\Phi(-\alpha)(a_0 - b_0)^2 \geq 0$. For positive θ (since $\theta \geq 0$), it is sufficient to show that

$$
\begin{aligned}
\xi_2 &\triangleq \xi_1/\theta^2 \geq 0. \\
\xi_2 &= \Phi(\alpha)\Phi(-\alpha) - \phi(\alpha)^2 + \frac{a_0 - b_0}{\theta}(1 - 2\Phi(\alpha))\phi(\alpha) \\
&\quad + \Phi(\alpha)\Phi(-\alpha)\frac{(a_0 - b_0)^2}{\theta^2} \\
&= \Phi(\alpha)\Phi(-\alpha) - \phi(\alpha)^2 + \alpha(1 - 2\Phi(\alpha))\phi(\alpha) \\
&\quad + \Phi(\alpha)\Phi(-\alpha)\alpha^2.
\end{aligned}
$$

$\xi_2(\alpha)$ is symmetric and is found to be nonnegative for all real values of α. For values of $|\alpha| \geq 3, \xi_2$ approaches 0 with both $\phi(\alpha)$ and $\Phi(\alpha)$ tending to 0. Fig. 7 shows the plot of ξ_2 as a function of α. ■

REFERENCES

[1] D. Sinha, N. V. Shenoy, and H. Zhou, "Statistical gate sizing for timing yield optimization," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 1037–1041.

[2] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Proc. Int. Conf. Comput.-Aided Des.*, 2003, pp. 900–907.

[3] H. Chang and S. S. Sapatnekar, "Statistical timing analysis under spatial correlations," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 9, pp. 1467–1482, Sep. 2004.

[4] A. Devgan and C. Kashyap, "Block-based static timing analysis with uncertainty," in *Proc. Int. Conf. Comput.-Aided Des.*, 2003, pp. 607–614.

[5] C. Visweswariah, K. Ravindran, K. Kalafala, S. G. Walker, and S. Narayan, "First-order incremental block-based statistical timing analysis," in *Proc. Des. Autom. Conf.*, 2004, pp. 331–336.

[6] J. Le, X. Li, and L. Pileggi, "STAC: Statistical timing with correlation," in *Proc. Des. Autom. Conf.*, 2004, pp. 343–348.

[7] C. E. Clark, "The greatest of a finite set of random variables," *Operations Research*, vol. 9, no. 2, pp. 145–162, Mar.–Apr. 1961.

[8] A. Agarwal, K. Chopra, D. Blaauw, and V. Zolotov, "Circuit optimization using statistical static timing analysis," in *Proc. Des. Autom. Conf.*, 2005, pp. 321–324.

[9] J. Singh, V. Nookala, Z. Luo, and S. Sapatnekar, "Robust gate sizing by geometric programming," in *Proc. Des. Autom. Conf.*, 2005, pp. 315–320.

[10] M. Guthaus, N. Venkateswaran, C. Visweswariah, and V. Zolotov, "Gate sizing using incremental parameterized statistical timing analysis," in *Proc. Int. Conf. Comput.-Aided Des.*, 2005, pp. 1029–1036.

[11] D. Sinha, H. Zhou, and N. V. Shenoy, "Advances in computation of the maximum of a set of random variables," in *Proc. Int. Symp. Quality Electron. Des.*, 2006, pp. 306–311.

[12] O. Coudert, "Gate sizing: A general purpose optimization approach," in *Proc. Eur. Des. Test Conf.*, 1996, pp. 214–218.

[13] O. Coudert, R. Haddad, and S. Manne, "New algorithms for gate sizing: A comparative study," in *Proc. Des. Autom. Conf.*, 1996, pp. 734–739.

[14] H. Chang, V. Zolotov, S. Narayan, and C. Visweswariah, "Parameterized block-based statistical timing analysis with non-Gaussian parameters, non-linear delay functions," in *Proc. Design Autom. Conf.*, 2005, pp. 71–76.

Debjit Sinha (M'06) received the B.Tech. (honors) degree in electrical engineering from the Indian Institute of Technology, Kharagpur, India, in 2001. He is currently pursuing the Ph.D. degree in electrical engineering and computer science at Northwestern University, Evanston, IL.

Following his doctoral degree, he will join IBM Microelectronics, East Fishkill, NY, as an Advisory Engineer, starting in the Fall of 2006. His research interests include algorithms and computer-aided design (CAD) for VLSI circuits, especially related to crosstalk, and statistical timing analysis, and optimization.

Narendra V. Shenoy (M'86–SM'99) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Mumbai, India, in 1988, and the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1993.

He is currently a Synopsys Scientist in the Advanced Technology Group at Synopsys Inc., Mountain View, CA. He has pursued research in the areas of timing analysis, logic synthesis and optimization, retiming, physical synthesis, routing, signal integrity, extraction, and programmable fabrics. His current interests include statistical timing analysis and optimization, and reducing variability in design.

Hai Zhou (M'04–SM'04) received the B.S. and M.S. degrees in computer science and technology from Tsinghua University, Bejing, China, in 1992 and 1994, respectively, and the Ph.D. degree in computer sciences from the University of Texas, Austin, in 1999.

He is currently an Assistant Professor in the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL. Before he joined the faculty of Northwestern University, he was with the Advanced Technology Group, Synopsys Inc., Mountain View, CA. His research interests include VLSI computer-aided design (CAD), algorithm design, and formal methods.

Prof. Zhou has served on the technical program committees of many conferences on VLSI circuits and CAD. He was a recipient of the CAREER Award from the National Science Foundation in 2003.