EVALUATING THE EFFECTIVENESS OF STATISTICAL GATE SIZING FOR POWER OPTIMIZATION

by

Nadathur Satish, Kaushik Ravindran, Matthew Moskewicz,
David Chinnery and Kurt Keutzer

Memorandum No. UCB/ERL M05/28

10 August 2005

EVALUATING THE EFFECTIVENESS OF STATISTICAL GATE SIZING FOR POWER OPTIMIZATION

by

Nadathur Satish, Kaushik Ravindran, Matthew Moskewicz,
David Chinnery and Kurt Keutzer

Memorandum No. UCB/ERL M05/28

10 August 2005

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Evaluating the Effectiveness of Statistical Gate Sizing for Power Optimization

Nadathur Satish   Kaushik Ravindran   Matthew Moskewicz   David Chinnery   Kurt Keutzer
University of California at Berkeley, CA, USA
{nrsatish, kaushikr, moskewcz, chinnery, keutzer}@eecs.berkeley.edu

ABSTRACT

We evaluate the effectiveness of statistical gate sizing to minimize circuit power. We develop reliable posynomial models for delay and power that are accurate to within 5-10% of 130nm library data. We formulate statistical sizing as a geometric program, accounting for randomness in gate delays. For various ISCAS-85 circuits, statistical sizing at a 99.8% target yield provides 25% power reduction compared to a 3σ worst-case deterministic approach. However, this can be replicated by deterministic sizing using a less conservative corner. Statistical sizing, under assumptions of variational independence, is still conservative and further power reductions can be achieved for the same timing target and yield.

1. INTRODUCTION

A standard technique for minimizing the power consumption of digital circuits is to downsize logic gates within the circuit. To first order, a smaller gate dissipates less power. However, smaller gates also have a lower drive strength and hence increase the overall delay. The objective is to size each gate to minimize circuit power while meeting delay constraints. The general optimization problem can be formulated as:

$$\begin{array}{lll} \text{min} & \textit{power} & \\ \text{subject to} & d_p \leq T, \quad \forall p \in P & \\ & s_{i_{lb}} \leq s_i \leq s_{i_{ub}}, \quad \forall i \in G & (1) \end{array}$$

where, G is the set of gates in a combinational circuit, P is the set of paths (from inputs to outputs) in the circuit, s_i is the size variable of gate $i \in G$, $s_{i_{lb}} \geq 0$ and $s_{i_{ub}} \geq 0$ are lower and upper bounds on the gate sizes, and T is the specified timing target.

In this paper, we consider the problem of gate sizing in the presence of variability in the circuit delay elements. As circuits scale to nanometer dimensions, the uncertainty in process parameters and device phenomena also impact delay and power. There can be variability due to manufacturing, due to environmental factors such as Vdd and temperature, and due to device fatigue phenomena such as electromigration and hot electron effects [1]. The consequence of variability is that identically designed circuits exhibit a large spread in delay and power metrics, which severely impacts the parametric yield.

One conservative way to account for variability is to perform deterministic sizing for power optimization at the worst-case values for all random components. While this ensures high yield, it pessimistically estimates power and performance. The other option is to include the parametric variation in gate delay as part of the sizing optimization problem. The objective is to select gate sizes to minimize power for a target yield η. The timing constraint in the optimization problem in Equation (1) is recast as a probabilistic constraint conditioned on yield η: $\Pr(d_p \leq T) \geq \eta, \ \forall p \in P$.

Mani and Orshansky [2] propose an efficient block-based approach for statistical sizing. They derive linear models for delay and power and allow variability in the delay coefficients. The statistical sizing problem is translated into a Robust Linear Program, which is solved optimally as a Second-Order Conic Program (SOCP). The authors report up to 30% power savings compared to a worst-case deterministic sizing approach. However, linear models for gate delay are inaccurate. Single linear fits to the industrial library data we use in our experiments have relative errors on average of between 19% and 30% (cf. Section 4).

A more accurate convex optimization approach is to use Geometric Programming (GP) with posynomial models. Fishburn and Dunlop [3] originally proposed posynomial models for transistor sizing. Boyd, et. al [4] incorporate posynomial delay and power models in a GP formulation for statistical sizing. They argue for the applicability of geometric programming for statistical analysis, but do not focus on results of particular sizing problems for realistic gate libraries.

The Mani-Orshansky [2] block-based statistical sizing approach and the GP modeling due to Boyd [4] serve as a combined starting point for our work. The objective is to evaluate how gate sizing in a statistical setup benefits power optimization compared to a worst-case deterministic sizing formulation. The importance of accounting for variability in delay and power models is clear. Sizing based on worst-case deterministic models is conservative, and this motivated the move to statistical formulations. In order to evaluate the advantages of statistical sizing, we ask the following questions: (a) How are statistical methods different from a worst-case deterministic sizing formulation? (b) How well does statistical sizing reclaim the pessimism inherent in deterministic worst-case approaches?

As a first step, we construct an accurate statistical sizing formulation (Section 3) based on delay and power models for a realistic 130nm library data (Sections 2). We verify the power reductions from statistical sizing compared to deterministic worst-case sizing across various ISCAS-85 benchmark circuits (Section 4). Based on these experiments, we interpret the differences between the statistical and deterministic sizing approaches. In particular, we ask whether incorporating variance in the delay model is akin to deterministic sizing with an appropriately chosen worst case library corner to achieve the desired yield (Section 5). We perform Monte Carlo simulations to compare the yields between the statistical and deterministic sizing approaches. We also analyze how much pessimism is reduced in moving from worst-case deterministic to statistical models (Section 6).

2. DELAY AND POWER MODELS

Gate sizing for combinational circuits minimizes total power subject to constraints on the gate sizes and required timing target (Equation 1). We assume that gates can be continuously sized between the specified bounds. This is a reasonable approximation for libraries with fine granularity of gate sizes, or in a liquid cell methodology. In this section, we describe the gate delay and power models used in our problem formulation. Our models were fitted to a 130nm standard cell library. In section 4 we discuss the accuracy of our models, and compare our delay and power estimates to results obtained from Synopsys Design Compiler.

2.1 Gate Delay Model

For the statistical sizing formulation in [2], the authors adopt a linear model for gate delay. The delay of gate i is given by:

$$d_i = a_i - b_i s_i + c_i \sum_{j \in FO(i)} s_j \qquad (2)$$

where, s_i is the size of gate i, FO is the set of gates that fanout from i and the term $\sum_{j \in FO(i)} s_j$ specifies the fanout load size driven by gate $i \in G$. The parameters (a_i, b_i, c_i) are delay coefficients determined from size and delay values that are obtained empirically by circuit simulation for each gate in the library. We found that the linear delay model in (2) was inaccurate for our library, with relative errors between 19% and 30% (cf. Section 4).

A posynomial model for gate delay is a more accurate way to capture the dependence of gate delay on the gate size and load capacitance [5]. We use posynomial delay models and formulate the optimization problem as a geometric program. The delay of gate i is given by:

$$d_i = a_i + b_i \frac{\sum_{j \in FO(i)} s_j}{s_i} \qquad (3)$$

The parameters a_i and b_i are fit by minimizing the relative least squares error between the posynomial models and the library data. This model reflects two aspects of the dependence between delay and size: (a) sizing up a gate increases its drive strength and hence decreases its delay, and (b) a larger sized gate presents a greater load capacitance to a gate driving it, and hence increases the delay of the driving gate.

One limitation in our gate delay model is that it does not include input slew. For fitting purposes, we assume an input slew of 0.07ns, which is appropriate for a circuit with a relatively tight delay constraint. This assumption is verified later by comparing delay and power for several benchmarks.

Gate delay is specified as a function of the internal size and load capacitance. We incorporate a wire-load model to capture the effect of wire length on the delay. The wire load is a multiple of the number of fan-outs and is an additive component to the fanout load. The wire load for gate i can be expressed as:

$$w_i = w_{i_1} + w_{i_2}|FO(i)| \qquad (4)$$

where, coefficients (w_{i_1}, w_{i_2}) depend on the process technology. We also specify separate delay coefficients for rise and fall timing arcs on each input pin of a gate. An n-input gate has $2n$ equations to describe its delay. For example, the delay equation for the rise timing arc along input j of gate i is the given by:

$$d^r_{i_j} = a^r_{i_j} + b^r_{i_j} \frac{(\sum_{k \in FO(i)} s_k) + w_i}{s_i} \qquad (5)$$

2.2 Gate Power Model

We use a linear model to describe power as a function of size and fanout load. The total power dissipated by a digital circuit is given by:

$$P_{total} = P_{dynamic} + P_{leakage} \qquad (6)$$

Dynamic power consists of two components: switching power and internal power. The switching power is written as:

$$P_{sw} = \frac{1}{2}\frac{1}{T} V_{DD}^2 \sum_{i \in G} \Pr(switch_i)\Big(w_i + \sum_{j \in FO(i)} s_j\Big) \qquad (7)$$

The switching power P_{sw} is directly proportional to the total switched capacitance, which corresponds to the sum of gate sizes and wire loads in the circuit. The term $\Pr(switch_i)$ measures the activity of gate i, i.e. the fraction of the cycles per second when the gate output rises or falls.

Internal power includes the power dissipation for (dis)charging capacitances internal to the gate, and any short circuit (crossbar current) that occurs when there is a conducting path from supply to ground. This is specified for each input pin and rising/falling outputs. We model internal power as a positive linear function of size and load:

$$\begin{aligned}
P_{int} &= \sum_{i \in G} P^r_{int_i} + P^f_{int_i} \\
P^r_{int_i} &= \sum_{j \in inputs(i)} \Pr(rise_j)\Big(m^r_{i_j} + n^r_{i_j} s_i + l^r_{i_j}(w_i + \sum_{k \in FO(i)} s_k)\Big) \\
P^f_{int_i} &= \sum_{j \in inputs(i)} \Pr(fall_j)\Big(m^f_{i_j} + n^f_{i_j} s_i + l^f_{i_j}(w_i + \sum_{k \in FO(i)} s_k)\Big)
\end{aligned} \qquad (8)$$

The coefficients $(m_{i_j}, n_{i_j}, l_{i_j})$ are obtained from a least squares fit. To improve the accuracy of the fit, we compute a piecewise linear function for the rise/fall internal powers. An activity probability $\Pr(rise_j)$ and $\Pr(fall_j)$ is associated with each input pin j of gate i.

The total dynamic power is then given by:

$$P_{dynamic} = \Pr(activity)(P_{sw} + P_{int}) \qquad (9)$$

The term $\Pr(activity)$ denotes the fraction of the second for which the circuit is actively performing computation. Circuit activity is the only source of dynamic power. The other source of power dissipation, when the inputs are held constant, is due to static or leakage power. Leakage power is given by:

$$P_{leakage} = \sum_{i \in G} \sum_{j \in \{0,1\}^{|inputs(i)|}} \Pr(state = j)(m_{i_j} + n_{i_j} s_i) \quad (10)$$

The leakage power is approximately proportional to gate size and is a function of the input state of the gate. Similar to internal power, we obtain coefficients (m_{i_j}, n_{i_j}) by fitting leakage as a piece-wise linear function for each input state. The probabilities of the leakage state and switching activity are obtained from SAIF (switching activity interchange format) files generated by the Synopsys VCS simulator for gate-level Verilog netlists, assuming independent random inputs with equal probabilities of being 0 or 1.

3. GATE SIZING

We base our statistical sizing formulation on the approaches presented by Mani [2] and Boyd [4]. First, we review the formulation of the deterministic gate sizing problem for power optimization. We cast it as a geometric program using the delay and power models presented in Section 2. For simplicity, we do not duplicate constraints for rise and fall timing arcs on each gate input. This formulation is then extended to incorporate variability in the delay model.

3.1 Deterministic Gate Sizing

Consider a combinational logic circuit $C = (G, E, I, O)$, where G is the set of gates, E is the set of nets (edges) between gates, I is the set of inputs and O is the set of outputs. In the absence of variability, the power minimization problem in (1) can be re-expressed as:

$$\begin{array}{lll} & \min & P_{total} \\ \text{subject to} & & \\ s_{i_{lb}} & \leq & s_i \leq s_{i_{ub}}, \quad \forall i \in G \\ d_i & = & a_i + b_i \dfrac{\sum_{j \in FO(i)} s_j}{s_i}, \quad \forall i \in G \\ d_i & \leq & t_i - t_j, \quad \forall j \in FI(i), \quad \forall i \in G \\ t_i & \leq & T, \quad \forall i \in O \\ t_i & = & 0, \quad \forall i \in I \end{array} \quad (11)$$

Variables d_i denote the delay of gate $i \in G$ under the posynomial delay model. $FI(i)$ and $FO(i)$ refer to the fan-ins and fan-outs of a gate i, respectively. The variables t_i denote the maximum signal arrival time at the output of a gate. The arrival times at the circuit inputs I are 0. The timing constraint $t_i \leq T$ is enforced on the arrival times at the primary outputs. We assume output port loads of 3fF. The optimization problem chooses gate sizes s_i to minimize power for the constraint that all circuit paths from input to output are within the target period T. The objective and inequality constraints are posynomials, hence this problem can be solved as a GP.

The general sizing optimization problem in (1) enforces the timing constraint on each circuit path. The updated version of this problem in (11) translates timing constraints on paths to timing constraints on gates, which makes the problem tractable for large circuits. These two problems have the same feasible solution set and optimum objective value. When the delay models are deterministic, the translation from the path-based formulation to the gate-based formulation is exact.

3.2 Statistical Gate Sizing

Statistical variability is introduced in the delay model by allowing randomness in the delay parameters a_i and b_i at each gate. We adhere to popular practice and model gate delay randomness as a Gaussian random variable. The variables a_i and b_i are Gaussian, with expected values $\bar{a}_i$ and $\bar{b}_i$, and standard deviation σ_{a_i} and σ_{b_i}.

The path-based sizing formulation in Equation (1) is easily translated into a statistical problem. A path $p \in P$ is a collection of gates. Then, $d_p = \sum_{i \in p} d_i$ is the delay along path $p \in P$. It is also a Gaussian variable with expected value $\bar{d}_p$ and standard deviation σ_{d_p}. The path-based sizing problem under uncertainty expresses the timing constraint in Equation (1) as: $\Pr(d_p \leq T) \geq \eta$, $\forall p \in P$.

The parameter η corresponds to the timing yield of the circuit. The optimization problem selects gate sizes to meet the timing target T with probability η. Given d_p is normally distributed, the probabilistic constraint can be written as:

$$\begin{array}{ll} \Pr(d_p \leq T) \geq \eta & \\ \Rightarrow & Prob\Big(\dfrac{d_p - \bar{d}_p}{\sigma_{d_p}} \leq \dfrac{T - \bar{d}_p}{\sigma_{d_p}}\Big) \geq \eta \\ \Rightarrow & \dfrac{T - \bar{d}_p}{\sigma_{d_p}} \geq \phi^{-1}(\eta) \\ \Rightarrow & \bar{d}_p + \phi^{-1}(\eta)\sigma_{d_p} \leq T \end{array} \quad (12)$$

where, ϕ is the cumulative probability distribution function (cdf) of the unit normal variable $N(0,1)$ [6]. The term $\phi^{-1}(\eta)$ is the margin coefficient of yield. For a 99.8% yield, $\phi^{-1}(\eta) \approx 3$. The probabilistic inequality constraints are posynomials, hence the problem remains a GP.

However, the path-based formulation is intractable for large circuits. The obvious solution is to convert probabilistic timing constraints on paths to probabilistic timing constraints on gates, just as in the deterministic sizing formulation in Equation (11). The probabilistic timing constraint on gate i is given by: $\Pr(d_i \leq t_i - t_j) \geq \eta$, $\forall j \in FI(i)$.

Unlike the deterministic case, the translation from the path-based to the gate-based formulation is not exact. The gate based formulation makes two implicit assumptions: (a) the gate delay variables (d_i) are independent, and (b) each gate meets its required time with probability η (this η is the target yield over all circuit paths in Equation (12)). The assumption on gate delay independence may be bypassed by describing gate delay as a function of global sources of variation [1], rather than using (a_i, b_i) values that are local to each gate. Regarding the second assumption, the choice of η at each gate is ad hoc, and there is no clear procedure to estimate target yield for each gate. The authors in [2, 4] argue that choosing the yield at each gate to be the target yield η of the circuit is a reasonable approximation. Based on these approximations, and the simplification in (12), the GP formulation for the statistical sizing problem to minimize power can be expressed as:

$$
\begin{array}{lrcl}
\min & & P_{total} & \\
\text{subject to} & & & \\
& s_{i_{lb}} \le & s_i & \le s_{i_{ub}}, \quad \forall i \in G \\
& \hat{d}_i & = & \bar{d}_i + \phi^{-1}(\eta)\sigma_{d_i}, \quad \forall i \in G \\
& \bar{d}_i & = & \bar{a}_i + \bar{b}_i \frac{\sum_{j \in FO(i)} s_j}{s_i}, \quad \forall i \in G \\
& \sigma_{d_i} & = & \sqrt{\sigma_{a_i}^2 s_i^2 + \sigma_{b_i}^2 (\frac{\sum_{j \in FO(i)} s_j}{s_i})^2}, \quad \forall i \in G \\
& t_i & \ge & t_j + \hat{d}_i \quad \forall j \in FI(i), \quad \forall i \in G \\
& t_i & = & 0, \quad \forall i \in I \\
& t_i & \le & T, \quad \forall i \in O
\end{array} \qquad (13)
$$

In the statistical sizing formulation above, the term $\phi^{-1}(\eta)\sigma_{d_i}$ is added to the mean value $\bar{d}_i$ of each gate delay. The additive margin is a function of the gate size, target yield and local variance, and attempts to capture the randomness of the delay parameters. Boyd, et al. call this the *surrogate delay model* [4]. The statistical sizing problem essentially performs deterministic sizing on this surrogate delay model.

The statistical gate sizing formulation in (13) incorporates uncertainty in the circuit delay elements. We do not need to explicitly model the variability in the dynamic and leakage power components in this problem. Even in the presence of variability in the coefficients of the power model, the objective for power minimization will be to minimize the sum of the expected values of power over all gates. This is equivalent to the objective of the sizing problem in (13), where P_{total} is minimized.

4. RESULTS

The objective of our study is to evaluate the benefits of statistical sizing over deterministic sizing using worst-case timing estimates. First, we briefly comment on the fidelity of our models. Following that, we present power optimization results obtained from statistical sizing for a few ISCAS-85 circuits.

4.1 Accuracy of Delay and Power Models

We performed our experiments using posynomial power and delay models fitted to standard cell libraries characterised for an STMicroelectronics 130nm HCMOS9D process. There are five logic gates in our library: inverter, 2-input nand, 2-input nor, 3-input nand, and 3-input nor. Each cell is available in high and low transistor threshold voltages, 0.23V and 0.14V respectively, giving a total of 10 distinct sets of data versus gate size that were fit. In our formulations, we constrain the gate size to be in the interval bounded by the minimum and maximum gates sizes specified in the library for each gate.

We used MATLAB to obtain least square fits (minimizing RMS error) for the delay and power coefficients at each gate input and rise/fall timing arcs. Table 1 presents the relative percentage errors of our estimates. The largest average error for the posynomial fits is 7.4% across all gates (columns 10-11). Compared to this, the linear delay models used by Mani and Orshansky [2] have average relative errors between 19% and 30% (columns 6-7). Even after gate delays are fitted as a maximum of piece-wise linear functions (columns 8-9), the models are still not sufficiently accurate. Their motivation behind using linear functions is to formulate the sizing problem as an SOCP, which is arguably easier to solve compared to a GP. However, gate delays are generally nonlinear functions of size, and fitting them as linear functions substantially increases the error.

We conducted our experiments on five ISCAS-85 benchmark circuits. To validate the accuracy of our models and solver framework, we compared our delay and power estimates for some circuit configurations (obtained for different size and threshold voltage assignments for each gate) against static timing and static power analysis results obtained from Synopsys Design Compiler (Table 2). The solver results are consistently within 5% of the reference values.

Table 2: Comparison of power and delay obtained from our solver against static analysis results from Synopsys Design Compiler (DC) for some circuit configurations (DC reference value / solver's estimate).

Circuit	# Gates	Delay (ps)	Power (μW)	Leakage (μW)
c17	10	92.7 90.6	19.9 21.4	1.1 1.1
c432	259	725.9 693.9	361.8 378.1	21.9 22.8
c499	644	700.7 692.6	772.9 810.0	39.5 40.9
c880	484	696.2 662.6	497.6 524.5	26.3 27.4
c1908	635	995.3 951.8	717.4 747.3	47.6 50.1

4.2 Power Results from Statistical Sizing

First, we present results validating the power savings obtained from statistical sizing for a 99.8% target yield. This corresponds to setting $\phi^{-1}(\eta)$, the margin coefficient of yield, to 3 in the GP in Equation (13). Following the approach in [2], we assume the overall variation in gate delay at 3σ to be around 25% of the mean value. However, it is not clear how the delay coefficients a_i and b_i at each gate map to physicial variability. Given the size range at each gate, we find from experiments that it reasonable to assume a standard deviation of 8% and 10% of the mean values for a_i and b_i. We use the MUSE Generalized Geometric Program solver [7] to solve the GP. We compare these results to the minimum power obtained from worst-case deterministic sizing. In the worst case setting, all the random variables ((a_i, b_i) in the gate delay models) are set to their 3σ values. T_{min} is the minimum delay through the circuit obtained from worst-case deterministic sizing (where the objective it to minimize T_{min} without constraints on power). We use T_{min} as the timing target for the deterministic and statistical sizing problems.

Table 3 presents the power savings obtained from statistical sizing. The T_{min} for each circuit is presented in column 2. Column 3 shows the total power after worst-case deterministic sizing and statistical sizing for a timing target of T_{min}. Statistical sizing reduces power by about 25% on average. We also analyze the minimum power obtained by the two approaches for relaxed timing targets greater than T_{min} (columns 4-7). As expected, the total power decreases as we

Table 1: Size bounds and relative percentage error of the delay and power models.

Cell Name	Size bounds (fF)		Leakage Power (%)	Internal Power (%)	Linear Delay Model (%)		Piece-wise Linear Delay Model (%)		Posynomial Delay Model (%)	
	Low	High			Rise	Fall	Rise	Fall	Rise	Fall
inv	1.0	37.7	8.66	11.99	28.07	29.94	18.86	20.39	6.55	6.35
nand2	2.0	26.0	5.73	6.38	21.69	22.81	15.27	18.41	7.09	2.94
nand3	3.1	42.2	4.84	5.43	21.26	20.27	19.30	13.44	7.44	4.19
nor2	2.2	33.7	4.24	4.62	20.49	22.86	12.05	12.97	4.47	6.56
nor3	3.8	51.1	7.05	3.49	19.55	20.43	10.52	14.08	4.65	7.25
hvt inv	1.0	35.5	8.96	5.36	28.20	29.51	20.17	29.37	6.24	6.30
hvt nand2	2.0	25.3	6.18	4.29	21.69	22.54	13.93	14.97	6.82	2.85
hvt nand3	3.0	40.9	4.72	2.83	20.78	19.06	13.81	14.62	7.28	4.43
hvt nor2	2.0	31.4	4.35	2.50	21.00	24.53	13.63	23.03	4.76	6.37
hvt nor3	3.5	48.1	6.26	2.91	20.08	19.89	13.32	15.30	4.28	7.15

increase the slack in the timing target. However, the power reduction due to statistical sizing is less at larger delay targets. We observed that around a $1.5T_{min}$ target a majority of gates in the circuit are sized at their lower bounds, beyond which no further power minimization is possible due to gate sizing.

Table 3: Minimum power obtained by deterministic ($\sigma = 3$) and statistical approaches ($\eta = 99.8\%$ target yield) for different timing targets (deterministic sizing / statistical sizing).

Circuit	T_{min}	Power (mW)				
		T_{min}	$1.02T_{min}$	$1.05T_{min}$	$1.1T_{min}$	$1.2T_{min}$
c17	0.106	0.024 0.018	0.020 0.016	0.017 0.014	0.013 0.011	0.010 0.009
c432	0.812	0.511 0.373	0.415 0.335	0.345 0.289	0.274 0.239	0.198 0.181
c499	0.787	1.097 0.801	0.928 0.729	0.783 0.641	0.628 0.537	0.458 0.412
c880	0.777	0.521 0.423	0.451 0.391	0.394 0.352	0.334 0.306	0.264 0.249
c1908	1.111	0.925 0.696	0.790 0.634	0.666 0.559	0.540 0.473	0.404 0.370

Table 4: Comparison of execution times between deterministic sizing ($\sigma = 3$) and statistical sizing ($\eta = 99.8\%$ target yield) for the T_{min} timing target.

Circuit	Deterministic Sizing	Statistical Sizing
c17	4 s	10 s
c432	5 m	25 m
c499	17 m	1.5 h
c880	14 m	1 h
c1908	22 m	2.5 h

Though GP formulations provide good accuracy, one drawback is that run times are high for large problem instances. Table 4 shows the execution times for the deterministic and statistical sizing problems. We ran our experiments on an Intel Pentium4 1.8 GHz machine. We observe that the deterministic sizing problems are significantly faster. Both these instances are GP problems: the only difference is that the statistical problem in Equation (13) has additional nonlinear terms due to the variance constraints.

5. YIELD TARGET FOR DETERMINISTIC SIZING

Based on the formulations in [2, 4], we have so far accurately verified the power reductions from statistical sizing over deterministic worst-case sizing. In this section, we evaluate how statistical sizing is different from the deterministic version. The formulation for statistical sizing in Equation (13) implicitly assumes that the variations at each gate are independent. This enables us to express the probabilistic timing constraint for each gate by adding a surrogate delay margin of $\phi^{-1}(\eta)\sigma_{d_i}$ to the mean delay. In the worst case deterministic setting, all random variables (assuming Gaussian distributions) are set to their 3σ values. The statistical sizing problem, for some yield coefficient $\phi^{-1}(\eta)$, essentially solves the deterministic sizing problem on the surrogate delay model. In this sense, the two problems are not very different; statistical sizing, instead of worst-casing the individual (a_i, b_i) random variables at each gate, worst-cases the gate delay random variable.

We then ask whether the statistical version of the problem can be approximated to deterministic sizing for a less conservative process corner. In the deterministic problem, instead of worst-casing all variations at 3σ, we would set all random variables to a $K\sigma$ value that gives a similar delay to assuming the worst case 3σ impact of a_i and b_i. We call K an intermediate margin coefficient of yield for deterministic sizing. The exact value of K would depend on the target yield η and delay data for some library, and must be characterized over a set of circuits.

For our library, we empirically compute the coefficient K at which the minimum power from statistical sizing is equal to the minimum power from deterministic sizing with all the random variables set to the $K\sigma$ values. The values of K obtained for different timing targets for a 99.8% yield across the five ISCAS-85 circuits are tabulated in Table 5.

Table 5: Intermediate coefficient of yield (K) values at which deterministic sizing matches stochastic delay and power estimates for a 99.8% yield.

Circuit	Intermediate Coefficient on Yield (K)				
	T_{min}	$1.02T_{min}$	$1.05T_{min}$	$1.1T_{min}$	$1.2T_{min}$
c17	2.34	2.33	2.30	2.27	2.23
c432	2.40	2.38	2.36	2.33	2.30
c499	2.23	2.22	2.21	2.20	2.20
c880	2.39	2.37	2.34	2.31	2.27
c1908	2.30	2.29	2.27	2.26	2.24

We observe that the K values are remarkably consistent across circuits and multiple timing targets. Hence, we can find a coefficient K for our library at which deterministic sizing (after fixing all random variables to their $K\sigma$ values) closely tracks statistical sizing on the surrogate delay model. For a specific value, $K = 2.39$, we compute minimum power from deterministic sizing. These results for the ISCAS-85 circuits are shown in Table 6 . The optimal power numbers closely match those computed using statistical sizing in Table 3. This shows, at least in the case of our library, that statistical sizing for some target yield can be replaced by deterministic sizing by choosing an appropriate intermediate coefficient K. The advantage of this is that we can then use the standard deterministic sizing approach, which has much smaller runtimes compared to statistical sizing.

Table 6: Minimum power from deterministic sizing at an intermediate coefficient of yield $K = 2.39$ (all random variables are set to $K\sigma$).

Circuit	Power (mW)				
	T_{min}	$1.02T_{min}$	$1.05T_{min}$	$1.1T_{min}$	$1.2T_{min}$
c17	0.018	0.016	0.014	0.012	0.009
c432	0.369	0.333	0.290	0.240	0.182
c499	0.836	0.758	0.663	0.552	0.419
c880	0.419	0.389	0.352	0.307	0.250
c1908	0.711	0.647	0.570	0.481	0.374

6. YIELD EVALUATION

We discussed three approaches to sizing: (i) worst-case deterministic sizing at 3σ values, (ii) statistical sizing for a specified target yield, and (iii) deterministic sizing at an intermediate coefficient of yield K. Worst-case deterministic sizing is overly pessimistic in its timing estimates. The objective in moving to statistical sizing is to reclaim some of this pessimism by accounting for parametric variations. Deterministic sizing at an intermediate yield coefficient also achieves this by reducing the over-estimation of the random variables.

To verify this reduction in pessimism, we perform Monte Carlo simulations for gate sizings from the three approaches. Figure 1 shows Monte Carlo results for the c432 benchmark using gate sizes obtained for $T_{min} = 0.812ns$. The left-most curve corresponds to worst-case deterministic sizing. Even though the circuit was optimized for power under the constraint that circuit delay is less than T_{min}, the worst-case deterministic sizing gives a much tighter clock period. This ensures a 100% yield, but at the expense of increased power (from Table 3, power is $0.511mW$).

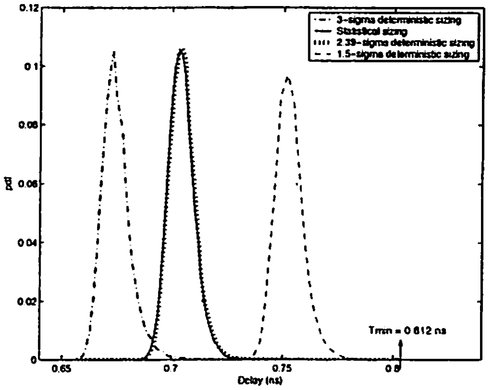

Figure 1: Delay distributions obtained from Monte Carlo on the C432 circuit for the three sizing approaches ($T_{min} = 0.812ns$)

The move to statistical sizing aims to recover some pessimism of the deterministic model. The solid middle curve is the Monte Carlo result on sizes obtained for a 99.8% yield. The dotted curve overlapping this is the result from deterministic sizing at $K = 2.39$. We verify that deterministic sizing at K closely tracks the statistical sizing result. Also, compared to the deterministic worst-case result, these distributions estimate delay less conservatively. This results in a greater power reduction (from Tables 3 and 6, power is $\approx 0.37mW$).

Interestingly, these distributions also have a yield of 100%, and not the expected 99.8% target. This means the circuit delay is still conservatively estimated when statistical sizing is performed. The surrogate delay model only accounts for statistical variations at the gate level. The statistical problem sets each gate delay independently to its 3σ value to target a 99.8% yield. But this does not account for the variance in the delay of a path. Hence path delay is over-estimated, leading to an overall pessimistic timing estimate.

This implies that further power reduction can be achieved for the same timing target and yield. As an example, the right-most curve in Figure 1 shows the delay distribution from deterministic sizing at a lower K value ($K = 1.5$). Even in this case, the yield is 100%, but the power reduces to $0.28mW$.

We observed that in the sizing results from the four approaches, further power reduction is possible at the T_{min} timing target and 99.8% yield, since a majority of gates are not at their minimum sizes. In the ideal case, statistical sizing would select gate sizes to minimize power and provide a timing distribution in which 99.8% of the paths meet T_{min}. However, the conservative nature of the timing estimate, while ensuring 100% yields, limits the achievable power reduction.

7. CONCLUSION

Statistical sizing for power minimization recovers some of the pessimism inherent in deterministic worst-case approaches and achieves a 25% improvement in power. However, the gate-based approaches to statistical sizing still estimate delay conservatively, which limits the achievable power reduction. These methods essentially reduce to deterministic sizing for a less conservative process corner, given by the coefficient of yield K, for a particular library and target yield. In order to overcome the pessimism in these gate-based statistical sizing approaches, we must account for dependent statistical variations along circuit paths to obtain further power savings.

8. REFERENCES

[1] C. Visweswariah, "Death, Taxes and Failing Chips," in *Proceedings of the Design Automation Conference (DAC)*, pp. 343–347, 2003.

[2] M. Mani and M. Orshansky, "A New Statistical Optimization Algorithm for Gate Sizing," in *Proceedings of the IEEE International Conference on Computer Design (ICCD)*, 2004.

[3] J. Fishburn and A. Dunlop, "TILOS: A Posynomial Programming Approach to Transistor Sizing," *IEEE Trans. on CAD*, pp. 326–336, 1985.

[4] S. Boyd, S. J. Kim, D. Patil, and M. Horowitz, "Digital Circuit Optimization via Geometric Programming," tech. rep., Stanford University, January 2005.

[5] M. Ketkar, K. Kasamsetty, and S. Sapatnekar, "Convex Delay Models for Transistor Sizing," in *Proc. of the 2000 Design Automation Conference*, pp. 665–660, 2000.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[7] Barcelona Design, "Muse Generalized GP Solver." http://www.barcelonadesign.com/university/, 2004.