

# Stochastic Logical Effort and Smart Monte Carlo for Timing Yield Estimation and Optimization

Alper Demir

Serdar Tasiran

aldemir@ku.edu.tr

stasiran@ku.edu.tr

Center for Advanced Design Technologies  
Department of Electrical & Electronics Engineering  
Department of Computer Engineering  
Koc University, Istanbul, Turkey

**Abstract**—This paper presents novel techniques for timing yield optimization and for yield estimation in the presence of large statistical process variations in integrated circuits. The techniques are based on our generalization of the logical effort delay model to circuits with stochastic parameter variations. In the spirit of the standard logical effort formalism, the stochastic gate delay model we propose separates the characterization of statistical variability from the gate topology, type, size and loading information. This separation of concerns is very powerful and facilitates the two novel approaches presented in this paper. In the first approach, we perform analytical and/or qualitative reasoning about timing yield and “back of the envelope” timing yield optimization in the same way that the logical effort formalism enables in the absence of timing variations. In the second approach, we improve the accuracy and efficiency of sign-off timing yield estimation based on transistor-level Monte Carlo simulations. We make novel use of importance sampling and other variance reduction methods in conjunction with the stochastic logical effort approximation for this purpose.

**Keywords**—logical effort, statistical variations, timing yield estimation and optimization, inter and intra-die variations, statistical timing analysis, Monte Carlo methods, variance reduction techniques.

## I. INTRODUCTION

Performance variability in integrated circuits due to statistical process variations and environmental fluctuations is becoming more and more significant. Considerable effort has been expended in the EDA community recently to cope with the *statistical timing* problem. Most of this effort has been aimed at generalizing the static timing analyzers to the statistical case.

By contrast, the method of logical effort by Sutherland et. al. [1] “is a way of thinking about delay” in digital circuits. It is an insightful and pragmatic methodology for quickly maximizing the speed of a circuit. In this paper, we develop and use the *stochastic* logical effort formalism

- to generalize the results of the deterministic logical effort formalism [1] to the case with statistical parameters,
- for tractable and back of the envelope optimization of path timing yield (as opposed to nominal path delay) in the spirit of the standard logical effort technique,
- for approximate but very fast and efficient timing yield estimation,
- to guide the generation/selection of sample points in the parameter/probability space in a transistor-level simulation based Monte Carlo method for sign-off timing yield estimation.

We first present a simple and analytical strategy for statistical gate sizing that is founded on the stochastic logical effort formalism, as opposed to a fully numerical statistical gate sizing tool with (incremental) statistical timing analysis in the loop. This strategy has the potential to obtain more optimal results because, as will be detailed later in the paper, it can use a more accurate stochastic gate delay model and a more accurate technique for estimating timing yield. We believe that this is a significant advantage over block-based methods which involve accurate and sophisticated manipulation of statistical models whose accuracy has not been directly justified.

We believe that the greatest value of this approach, like the original logical effort formalism, lies in the insight it provides. Fully numerical optimization techniques with a statistical timing analyzer in the loop “are prone to get stuck in local optima and are unlikely to produce meaningful results unless the user knows approximately what results to expect” [1] for the optimal deterministic sizing tools with a static timing analyzer in the loop. The *statistical gate sizing* problem is a much more difficult (stochastic) nonlinear programming problem than the deterministic gate sizing problem. Thus, when applied to real designs, fully numerical optimization strategies with statistical timing analysis in the loop may face significant difficulties. We believe that

only simple models like the one we are proposing and simple heuristics like the one proposed recently by Boyd and Horowitz in [2] have the potential to become practically useful and also make sense in the statistical timing analysis and optimization arena.

The stochastic logical effort model and the timing yield estimation and optimization methodology we develop in this paper are inherently *path-based*, because they are founded on the standard logical effort formalism [1]. We view this as an advantage since “Synthesis tools make some effort to explore topologies, but still can *not* match experienced designers on *critical paths*.” [1]: We believe that a simple, effective and working path-based timing yield estimation and optimization technique will be more practical than a complicated and non-intuitive block-based approach that relies on models postulated without much justification.

In the second half of this paper, we present methods for estimating circuit timing yield for sign-off. We believe that, given the magnitude of process parameter variations and the non-linear dependency of gate and circuit delay on these variations, the only sufficiently reliable and accurate method for this purpose is detailed, circuit-level simulation. We believe that sufficient accuracy in yield estimation can not be obtained even by applying Monte Carlo simulations at a higher level. Yield estimation techniques not based on Monte Carlo simulation operate by propagating probability density functions across the circuit. To make this process computationally feasible, they are forced to use approximate gate delay models and delay propagation methods that may be too inaccurate when process parameter variations are large. In this case, accurate determination of timing yield must have circuit simulation as its basis as well. The techniques we propose facilitate judicious choice of a set of assignments to process parameters for which full-circuit timing simulations will be performed. Spending computational effort for improving this choice is well justified since full-circuit timing simulations are expensive.

Let  $X$  denote an assignment to the process parameters and let  $f(X)$  denote the value of the joint probability density function for this assignment. In conventional Monte Carlo yield estimation, a number of sample assignments  $X_1, X_2, \dots, X_N$  are generated using the probability density function  $f(X)$ . The overall delay for each  $X_i$  is determined by performing circuit-level timing simulation. An estimator for timing yield is obtained by considering the fraction of samples for which the timing constraint is satisfied.

Because of the computational cost of determining circuit delay for each sample, the number of samples one has to work with is limited. This adversely affects the accuracy of the estimator – the estimator has large variance for small  $N$  which decreases proportionally to  $\sqrt{N}$ . This is a weakness of the conventional Monte Carlo method and has prevented it from finding widespread use for yield estimation.

Numerous techniques for reducing the variance of the estimator exist in Monte Carlo simulation literature (See [3,4] for example). In this study, we concentrate on two of them: Importance sampling and the use of control variates. Approximately speaking, importance sampling biases the choice of samples from the process parameter space more towards areas where the circuit delay violates the timing constraint. In the latter technique, Monte Carlo simulation is used to estimate the difference between actual yield and an approximation.

Both variance reduction techniques require an accurate but inexpensive approximation to circuit yield. We use the stochastic logical effort approach for this purpose. The stochastic logical effort approximation can be used to facilitate other techniques for variance reduction in Monte Carlo estimation, e.g. stratified sampling [5]. In this paper, we (i) demonstrate the use of the stochastic logical effort approximation in Monte Carlo variance reduction techniques, and (ii) propose Monte Carlo simulation in conjunction with variance reduction methods as an accurate yet computationally viable yield estimation approach.

In Section II below, we describe the stochastic gate delay model. Then in Section III, we present results obtained with this model for path timing yield optimization in two cases of practical interest, first with only inter-die variations and then in the presence of both inter-die and intra-die variations. Section IV formalizes the yield estimation problem and gives an overview of the Monte Carlo variance reduction techniques referred to above: Importance sampling and the use of control variates. Also in Section IV, we propose a combination of these two techniques that can potentially further reduce variance for the yield estimation problem. In Section V, we formulate the yield estimation problem as a definite integral and present how we apply the variance reduction methods of Section IV to this problem by using the stochastic logical effort approximation. In Section V-D we propose a set of experiments to test the effectiveness and computational cost of our methods.

## II. STOCHASTIC GATE DELAY MODEL WITH SEPARATION OF CONCERNS

We model the delay of a logic gate using the logical effort formalism [1]

$$d_{abs} = \tau d \quad (1)$$

where  $d_{abs}$  is the absolute delay of a gate measured in seconds,  $\tau$  is the delay of a *reference inverter* (with no parasitic capacitance) driving another inverter, and  $d$  is the delay of the logic gate expressed in units of  $\tau$ .

### A. Capturing statistical variations with a stochastic delay unit

In the logical effort formalism [1], all delays are expressed in terms of  $\tau$  in order to isolate the effects of the particular integrated circuit fabrication process. Thus,  $\tau$  serves as the delay unit that characterizes a given process and its value depends on the process parameters, power supply voltage and temperature. With a simple transistor model, one could derive an analytical expression for  $\tau$ , expressed in terms of transistor lengths and widths, gate oxide thickness, carrier mobilities and some other process parameters. Alternatively, one can *extract* the value of  $\tau$  from suitable test circuits by simulating them in a circuit simulator with a detailed, full transistor model, as discussed in [1]. In the standard logical effort formalism [1], the statistical variations of process and circuit parameters are not considered. Hence, for a given integrated circuit fabrication process, and for given environmental conditions (i.e., power supply voltage and temperature),  $\tau$  is characterized/extracted as a *single number* expressed in picoseconds. In our case, we use  $\tau$  to *capture* and *isolate* the effects of process, circuit and environmental parameters that exhibit statistical variations. Thus,  $\tau$  is not a single number, but it is a *probability distribution*. We propose three alternative techniques for the stochastic characterization of  $\tau$ :

(i) One can derive a simplified analytical expression that relates  $\tau$  to the circuit, process and environmental parameters that exhibit statistical variations, as discussed above. This analytical expression in conjunction with a multi-dimensional probability distribution that characterizes the basic statistical parameters can be used to compute the probability distribution of  $\tau$ . This can be done using a simple and efficient Monte Carlo technique, by sampling the joint distribution of the statistical parameters and by computing the corresponding value of  $\tau$  by simply evaluating the analytical expression, followed by a compilation of a histogram. Even the simplest analytical expression for  $\tau$  will have nonlinear dependence on the statistical parameters. Thus, even if the statistical parameters are jointly Gaussian, the probability distribution of  $\tau$  will not be Gaussian.

(ii) Instead of using an analytical expression as above, one can use a suitable test circuit and a circuit simulator to relate  $\tau$  to the statistical parameters. One can then again use a Monte Carlo technique to compute the probability distribution of  $\tau$ . This characterization will be computationally more expensive than the one above, because the test circuit will need to be simulated many times. However, the size of the circuit will be very small, essentially a CMOS inverter loaded by several others. Moreover, this Monte Carlo stochastic characterization for  $\tau$  will be performed only once and the results will be used many times later when estimating timing yield for a much larger logic circuit composed of inverters and other complex gates. One can also envision that the complex, nonlinear relationship that relates  $\tau$  to the statistical parameters can be represented by a multi-dimensional table or by building a response surface model using the Monte Carlo samples obtained with the circuit simulator.

(iii) Alternatively, one can use fabricated test structures and physically measure  $\tau$  and characterize its probability distribution.

In the standard logical effort formalism,  $\tau$  is characterized as a single number, using a single inverter which serves as the process reference for all of the logic gates on a single die (chip) or functional unit. When we consider statistical parameters which exhibit only inter-die variations, a single inverter can still serve as the statistical process reference for all of the logic gates on a chip. With only inter-die variations, statistical parameters on the chip at all locations are perfectly correlated. Using the stochastic characterization of  $\tau$  for the same reference inverter for all of the logic gates on the die captures this perfect statistical correlation among gates. When we also consider intra-die variations, statistical variations of gates are not perfectly correlated any more. In this case, we introduce an additional gate/location-dependent component  $\tau_r$  and re-express (1) as follows

$$d_{abs} = (\tau + \tau_r) d \quad (2)$$

$\tau$  above captures the effects of perfectly correlated inter-die statistical variations for the gates that reside on the same die.  $\tau_r$ 's for different  $r$  can be considered as either *uncorrelated* or *partially correlated* and they represent the effect of intra-die statistical variations. For two gates  $r_1$  and  $r_2$  that are in proximity with each other on the same die,  $\tau_{r_1}$  and  $\tau_{r_2}$  may be partially correlated. If these gates are far away from each other, then  $\tau_{r_1}$  and  $\tau_{r_2}$  may be considered uncorrelated. The partial correlation among  $\tau_r$ 's may be best represented by expressing them in terms of other (abstract) random quantities which are independent. This can be accomplished through (nonlinear) principal or independent component analysis [6]. The analysis is trivial in the case when  $\tau_r$ 's are jointly Gaussian. Even though it is not as straightforward, there exist techniques that work in the case when  $\tau_r$ 's are not jointly Gaussian. The mean delay of the reference inverter, as well as the effect of inter-die variations, are captured by  $\tau$  in (2), and  $\tau_r$  is an additional, small correction term accounting for the effect of intra-die variations. Thus,  $\tau_r$  is most likely, but not necessarily, zero mean. Moreover, a Gaussian model for  $\tau_r$  may not be too far off. Even when  $\tau$  and hence the total delay  $\tau + \tau_r$  is far from Gaussian, it may be good enough to model  $\tau_r$  as a Gaussian random quantity. Since  $\tau$  captures the effects of inter-die variations, its probabilistic properties such as its mean and standard deviations can be considered as gate and gate location independent. On the other hand, we use  $\tau_r$  to capture the effect of local intra-die variations. As first observed and proposed by Pelgrom in [7], the statistical (squared) variation (i.e., variance) in a statistical parameter (threshold voltage, channel length, delay) of an entity (transistor, gate, cell, block, etc.) is inversely proportional to the total area it occupies. Hence, the probabilistic properties for  $\tau_r$  can be modeled as gate and location dependent to capture this basic fact regarding local, intra-die variations.

### B. Statistical (in)variability of logical effort

We now concentrate on the other factor  $d$  in the gate delay model of (2). In the logical effort formalism,  $d$  is expressed as

$$d = (p + gh) \quad (3)$$

where  $p$  represents the intrinsic (parasitic) delay,  $g$  is the logical effort, and  $h$  is the electrical effort or electrical fanout. Logical effort  $g$  for a logic gate is defined as the (unitless) ratio of its input capacitance to that of an inverter that delivers the same output current. Logical effort  $g$  is a measure of the complexity of the gate, it depends only on its topology and it is independent of the size and the loading of the gate. Parasitic delay  $p$  expresses the intrinsic delay of the gate due to its own internal parasitic capacitance, and it is largely independent of the sizes of the transistors in the gate. Parasitic delay  $p$  is also a unitless quantity, it is expressed in units of  $\tau$ . The electrical effort  $h$  is the ratio of the load capacitance of the logic gate to the capacitance of a particular input [1].

Ideally, the logical effort  $g$  and the unitless parasitic delay  $p$  of a gate would be independent of process and environment parameters, and depend only on the topology of the gate. In reality, the logical effort  $g$  and the parasitic delay  $p$  vary slightly with process parameters and operating conditions. Sutherland et. al in [1] study the process and operating condition sensitivity of  $g$  and  $p$  for NAND and NOR gates with 2, 3 and 4 inputs. They compute  $g$  and  $p$  for processes ranging from a  $2.0\mu$  process to a  $0.35\mu$  one and power supply voltages ranging from 5.0 volts to 2.5 volts. Their results indicate that, over such a wide range of processes and power supply voltages, logical effort  $g$  shows a variation around  $\pm 10\%$  around the mean [1, Table 5.4]. For parasitic delay  $p$ , the variation is around  $\pm 15\%$  around the mean [1, Table 5.5]. Sutherland et. al in [1] also study the process and operating condition variability of  $\tau$  for the same range of processes and power

supply voltages mentioned above. Their results show that  $\tau$  ranges from approximately 25 psecs to 160 psecs, more than a factor of 6 variation. As seen here, almost all of the process and environmental variability shows up in  $\tau$  in (1) and the logical effort  $g$  and the unitless parasitic delay  $p$  exhibit relatively much less variation even with a wide range of processes and power supply voltages. The kind of process and environmental variability we consider in this work is quite different from the setting studied by Sutherland et. al in [1]. The kind of variability considered by them spans across fabrication processes from a  $2.0\mu$  process all the way to a  $0.35\mu$  process, and power supply voltages from 5.0 volts to 2.5 volts. In our case, we concentrate on a single fabrication process and a fixed nominal power supply voltage, and we consider *small statistical* variations in some parameters of the fabrication process, such as channel length and oxide thickness. We therefore do not expect the kind of variation mentioned above in, for example, the reference inverter delay  $\tau$ . We expect at most a  $\pm 35\%$  variability in  $\tau$  and proportionally much less variation in the logical effort  $g$  and the parasitic delay  $p$ . Given the setting and the evidence, one can argue that almost all of the statistical variability will show up in  $\tau$  in (1) and the unitless factor  $d$  in (3) will exhibit *practically insignificant* statistical variability. Thus, we will assume that  $d$  is independent of the statistically varying process parameters and the environmental conditions.

### C. Justification for and implications of the statistical invariability of logical effort

Rabaey et. al in [8] report results that are in support of our claim of the statistical invariability of logical effort. They state that “Electrical and logic effort do not contribute to the delay distribution”. Their observation is inspired and supported by the results they present on the yields of an inverter chain, a NAND chain and a 4-bit adder circuit. The yields they compute for these circuits with Monte Carlo SPICE simulations show little sensitivity to the type and the topology of the circuit. They also state that this will no longer be true if statistical variations in different gates are not perfectly correlated, i.e., with strong intra-die variations. However, they do not present any results in support of this second statement. Next, we demonstrate how the generalized stochastic gate delay model in (2) can be used to explain the reason behind both of the above observations in the light of the discussion we presented in the previous section.

Let us consider a path  $\pi = (g_1, g_2, \dots, g_R)$  with  $R$  gates with  $g_r$  as the  $r$ th gate on the path. When only inter-die variations are present, we have  $\tau_r = 0$  in (2), and hence the total delay for the path can be expressed as follows

$$D_{abs}^\pi = \tau \sum_{r=1}^R d_r = \tau \sum_{r=1}^R (p_r + g_r h_r) = \tau D^\pi \quad (4)$$

which is the product of the two terms,  $\tau$  that captures all statistical variation, and a unitless, deterministic, fixed number  $D^\pi$  that quantifies the *total complexity* of the logic gates on the path. Rabaey et. al in [8] define yield using

$$Y^\pi = \mathbb{P}(D_{abs}^\pi < 1.1 \mathbb{E}[D_{abs}^\pi]) \quad (5)$$

where  $\mathbb{P}(\cdot)$  is the probability measure and  $\mathbb{E}[\cdot]$  is the expectation operator. We would like to draw the attention of the reader to the fact that the delay cut-off value (i.e., delay specification) in the yield definition above is expressed *relative* to the nominal delay  $\mathbb{E}[D_{abs}^\pi]$ , not as an absolute value. If the probability density function (PDF) of  $\tau$  is  $p_\tau(\eta)$  defined for  $0 \leq \eta < +\infty$ , then the PDF of  $D_{abs}^\pi$  in (4) is given by

$$p_{D_{abs}^\pi}(\eta) = \frac{1}{D^\pi} p_\tau\left(\frac{\eta}{D^\pi}\right) \quad (6)$$

because  $D^\pi$  is simply a deterministic constant. Hence, we can compute yield  $Y^\pi$  as follows

$$Y^\pi = \int_0^{1.1 \mathbb{E}[D_{abs}^\pi]} \frac{1}{D^\pi} p_\tau\left(\frac{\eta}{D^\pi}\right) d\eta \quad (7)$$

where the nominal delay  $\mathbb{E}[D_{abs}^\pi]$  is given by  $\mathbb{E}[D_{abs}^\pi] = D^\pi \mathbb{E}[\tau]$ . If we substitute this in (7) and make a change of integration variable using  $u = \frac{\eta}{D^\pi}$ , we obtain

$$Y^\pi = \int_0^{1.1 \mathbb{E}[\tau]} p_\tau(u) du \quad (8)$$

We observe that the yield expression above is *independent* of  $D^\pi$  and depends only on the PDF and the mean of  $\tau$ . Thus, with only inter-die variations considered, the yield of a path defined as in (5) is independent of  $D^\pi$  and hence the types, topologies, sizes and the loading of the logic gates on the path. We arrived at this result due to the following two reasons:

- The cut-off delay in the yield definition in (5) is chosen in *relative* (as a multiple of) to the mean, i.e., nominal, delay.
- The total path delay in (4) can be expressed as the *product* of a deterministic constant  $D^\pi$  and a random variable  $\tau$ .

If we also consider partially uncorrelated intra-die variations, the path delay expression in (4) becomes

$$D_{abs}^\pi = \tau D^\pi + \sum_{r=1}^R \tau_r (p_r + g_r h_r) \quad (9)$$

In this case, we can not express the delay above as the product of a random variable and a deterministic constant, but it can be expressed as a linear combination of random variables. In this case, with a little bit of tinkering, one can conclude that the yield can *not* be independent of  $d_r = p_r + g_r h_r$  in general. Thus, the yield as defined in (5), in general, depends on the types, topologies, sizes and the loading of the logic gates on the path. To put it more precisely, if  $Z = \sum_{k=1}^K a_k X_k$  is a random variable formed as a linear combination of random variables  $X_k, k = 1 \dots K$  with deterministic constants  $a_k, k = 1 \dots K$ , then  $\mathbb{P}(Z < \alpha \mathbb{E}[Z])$  (for some deterministic constant  $\alpha$ ) is independent of  $a_k$ 's if and only if  $a_1 = a_2 = \dots = a_K$ . Hence, in general,  $\mathbb{P}(D_{abs}^\pi < 1.1 \mathbb{E}[D_{abs}^\pi])$  in (5) depends on  $d_r = p_r + g_r h_r$  with  $D_{abs}^\pi$  expressed as a linear combination of the random variables  $\tau$  and  $\tau_r$ 's as in (9). However, if inter-die variations are ignored, i.e. if  $\tau = 0$ , and if all of the gates on the path are identical, i.e. if  $d_1 = d_2 = \dots = d_R$ , then the yield defined by (5) is independent of  $d_r$ , i.e., the type of the gate in the path. In this case, the yield of an inverter chain and a NAND chain will be the same.

As demonstrated above, the simple, analytical, stochastic model we developed for gate delay is very powerful. Using this model, we can not only explain observations made by other authors in the literature, but also analyze and understand new situations.

### D. Summary

We model the delay of a gate, by generalizing the logical effort formalism [1] to the statistical case, using the following equation

$$d_{abs} = (\tau + \tau_r) d = (\tau + \tau_r) (p + gh) \quad (10)$$

where  $\tau$  is a random variable that serves as a *stochastic delay unit*.  $\tau$  characterizes a given fabrication process and its distribution depends on the statistical variability of the process parameters and the environmental conditions such as the power supply voltage and temperature.  $\tau$  captures the effect of inter-die process variations. The same  $\tau$  is shared by all of the gates on a die. Hence,  $\tau$  captures the perfectly correlated portion of the statistical variations, including the nominal delay.  $\tau_r$  above is also a random variable, a zero-mean correction term accounting for intra-die process variability that is not fully but partially correlated among the gates on the same die.  $d = p + gh$  above is a unitless and most importantly a *deterministic* quantity. The information about the topology and the type of the gate, and also its size and load, is captured by  $d$ . The random variables  $\tau$  and  $\tau_r$  above are largely *independent* of the gate type, topology, size and loading, whereas the deterministic quantity  $d = p + gh$  is *independent* of the statistically varying process parameters. Thus, we have a gate delay model with *separation of concerns*. The two *concerns* here are

- Statistical variations in gate delay.
- Gate type, topology, sizing and loading.

With the stochastic gate delay model in (10), we elevate the logical effort formalism from the deterministic, nominal case to the statistical one. Next in Section III, we use the gate delay model in (10) in generalizing some of the results of the logical effort formalism, from the setting of the *minimization of nominal path delay* to the *maximization of path timing yield*.

## III. TIMING YIELD OPTIMIZATION

We now consider the optimization of the delay (deterministic case) and the timing yield (stochastic case) of a path using the logical effort

formalism and the stochastic gate delay model we developed in Section II. Due to the length restrictions on the manuscript, we will be able to present only a summary of the results we have obtained.

We first review the deterministic minimization of path delay using the standard logical effort formalism. Let us now consider a path  $\pi = \langle g_1, g_2, \dots, g_R \rangle$  with  $R$  gates. The total delay of the path can be expressed as in (4). We do not consider statistical variations in this case. Hence,  $\tau$  is a fixed constant.  $h_r$  in (4) is the electrical effort of the  $r$ th gate  $g_r$  on the path. It is the ratio of the load capacitance of  $g_r$  to the capacitance of a particular input.  $h_r$  essentially combines the size information for a gate with the information on the load it is driving. The load of gate  $g_r$  is the input capacitance of gate  $g_{r+1}$ . In path delay minimization, a constraint  $H$  on the ratio of the capacitance at the termination of the path to the input capacitance of the first gate on the path must be specified. This ratio  $H$  is called the *path electrical effort* and is essentially equal to the multiplication of the electrical efforts of all of the gates in the path. Path delay minimization without a constraint on the path electrical effort is not meaningful. By formulating and analytically solving the deterministic path delay minimization problem, one arrives at the following result:

*The path delay is minimized when the product of the logical effort  $g_r$  and the electrical effort  $h_r$  is the same for each logic stage on the path.*

In the terminology of the logical effort formalism, the product of the logical effort and the electrical effort for a gate is called its *effort delay* or *stage effort*. Thus, to minimize the path delay, the stage effort for all of the gates on the path should be made equal to each other.

We now consider inter-die statistical variations only, i.e., the statistical variation of delay is perfectly correlated in all of the gates on the path. In this case, we use (10) with  $\tau_r = 0$  in order to null out the intra-die variations. We use the following definition for the *path timing yield*

$$Y^\pi = \mathbb{P}(D_{abs}^\pi \leq T_c) \quad (11)$$

where  $T_c$  is specified as the cut-off delay. In the presence of inter-die statistical variations, we formulate a *stochastic* optimization problem to maximize the path timing yield (as opposed to minimizing the nominal delay), and by analytically solving it we arrive at the same result we have obtained in the deterministic case:

*The path timing yield is maximized when the product of the logical effort  $g_r$  and the electrical effort  $h_r$ , i.e. the stage effort or the effort delay, is the same for each logic stage/gate in the path.*

We next consider both inter-die and intra-die statistical variations, and use the full stochastic gate delay model in (10). We proceed with a simplification and approximate the PDF of the reference inverter delay  $\tau$  (due to inter-die variations) with a Gaussian PDF, and assume that the PDF of the reference inverter delay  $\tau_r$  due to intra-die variations can also be approximated with a Gaussian PDF.  $\tau_r$ 's (represent corrections for the intra-die statistical variations) above are assumed to be uncorrelated with each other and also with  $\tau$ . These simplifications enable us to reach possibly sub-optimal but much more simpler and practical conclusions on yield maximization in the presence of both inter-die and intra-die statistical variations. We then formulate and solve a *stochastic* optimization problem to maximize the path timing yield in the presence of both inter-die and intra-die statistical variations, and arrive at the following result:

*In the general case when both inter-die and intra-die variations are present, equalizing stage efforts for each logic stage/gate in the path does not maximize the path timing yield. However, the optimal stage efforts and the optimal yield are very close to the ones obtained when stage efforts are chosen equal, and the differences are practically insignificant.*

The above result is not very surprising, because the optimization landscape is relatively flat, as discussed in [1] and [2].

In the gate delay model we used so far, we assumed that the probabilistic properties (i.e., mean, standard deviation and PDF) of  $\tau$  and  $\tau_r$  are gate or location independent. This is justified for  $\tau$  since it models inter-die variations. However, as we discussed in Section II, the variance of  $\tau_r$  can typically be modeled to be inversely proportional to the total area the gate occupies. With this variance model for intra-die variations, the optimal values for  $h_1, h_2, \dots, h_R$  come out in such a way that the differences between the stage efforts (product of  $g_r$  and  $h_r$ ) of different stages are not insignificant anymore. When the path delay PDF with equalized stage efforts is compared with the (stochastically) optimal stage efforts that come from the solution of the stochastic optimization problem, one observes that the (stochastically optimal) mean

path delay is larger than the (deterministically optimal) mean path delay with equalized stage efforts. However, the (stochastically) optimal standard deviation is smaller than the standard deviation with equalized stage efforts, resulting in an overall larger yield. This behavior was also observed by the authors in [2] with their block-based, heuristic stochastic sizing technique. The very simple, almost back of the envelope, timing yield estimation and optimization methodology we are proposing in this paper is able to arrive at the same conclusions.

The rest of the paper is dedicated to presenting our techniques for sign-off timing yield estimation using Monte Carlo transistor-level simulation. We present two Monte Carlo variance reduction methods and a combined method that is novel in this context and explain how we use them for yield estimation.

## IV. OVERVIEW OF MONTE CARLO METHODS

### A. Definitions

In the following analyses, we focus on the variations in delays induced by process parameter variations and assume that the design parameters for the circuit are given by  $S$  and fixed.

A path  $\pi$  in a circuit  $C$  is a sequence of gates  $g_0, g_1, g_2, \dots, g_n$  where  $g_0$ 's inputs are primary inputs of the circuit,  $g_n$ 's output is a primary output of the circuit. Given a circuit and values for its process parameters, a path is said to be *critical* if (i) it is sensitizable, and (ii) its delay is as large as the delays of other sensitizable paths. A path  $\pi$  is said to be *statistically critical* if it is a critical path of  $C$  for some possible assignment to process parameters. We denote by  $\Pi_{crit}$  the set of statistically critical paths. As explained in [9], the set of paths  $\Pi_{crit}$  can be approximated accurately by running a modest number of Monte Carlo simulations on the full circuit and including in  $\Pi_{crit}$  any path that is critical in any of the Monte Carlo runs. We then have the following equation for the delay of a circuit

$$d_C(S, X) = \max_{\pi \in \Pi_{crit}} d_\pi(S, X) \quad (12)$$

where  $d_C(S, X)$  is the delay of the circuit and  $d_\pi(S, X)$  is the delay of path  $\pi$  when the process parameters are given by  $X$ .

### B. Accurate and Efficient Monte-Carlo Evaluation of Integrals

The techniques described in this section involve the estimation of the value of a definite, finite-dimensional integral of the form

$$G = \int_{\Omega} g(X) f(X) dX \quad (13)$$

where  $\Omega$  is a finite domain and  $f(X)$  is a probability density function over  $X$ , i.e.,  $f(X) \geq 0$  for all  $X$  and  $\int_{\Omega} f(X) dX = 1$ . For delay computation,  $X$  is a vector variable that corresponds to the process parameters and  $f(X)$  represents the probability density function of the process parameters. If  $g(X)$  is chosen to be a function that evaluates to 1 when the circuit delay is as desired and 0 otherwise, then the value of the integral  $G$  is the circuit yield.

Monte Carlo estimation for the value of  $G$  is accomplished by drawing a set of samples  $X_1, X_2, \dots, X_n$  from  $f(X)$  (analogous to picking values for the process parameters from their respective probability distributions) and by letting the estimator  $G_N$  be given by

$$G_N = (1/N) \sum_{i=1}^N g(X_i) \quad (14)$$

The variance of  $G_N$  decreases proportionally to  $\sqrt{N}$ . Several techniques exist for improving the accuracy of Monte Carlo evaluation of finite integrals. These techniques reduce the number of Monte Carlo simulations required to estimate the value of an integral accurately, i.e. with small variance. In the following two subsections, we present an overview of two of these techniques based on the exposition in [3]. In Section IV-B.3, we propose a combination of these techniques.

#### B.1 Importance Sampling

Importance sampling improves upon the straightforward approach above by drawing samples for  $X$  from another distribution  $\tilde{f}$  in order to reduce the variance of the estimator  $G_N$ .  $G$  is then written as

$$G = \int_{\Omega} \left( \frac{g(X)f(X)}{\tilde{f}(X)} \right) \tilde{f}(X) dX \quad (15)$$

If  $X_1, X_2, \dots, X_n$  are drawn from  $\tilde{f}$  instead of  $f$ , the new estimator  $\tilde{G}_N$  is expressed as

$$\tilde{G}_N = (1/N) \sum_{i=1}^N \frac{g(X_i)f(X_i)}{\tilde{f}(X_i)} \quad (16)$$

in order to compensate for the different biasing of input samples induced by  $\tilde{f}$ .  $\tilde{f}$  must also satisfy the requirement that, for every  $X_0$  for which  $f(X_0)g(X_0)$  is non-zero,  $\tilde{f}(X_0)$  must also be non-zero for the integral in Eqn. 15 to be well defined.

The ideal choice of  $\tilde{f}$  that minimizes the variance of the estimator  $\tilde{G}_N$  is  $\tilde{f}_{ideal}(X) = \frac{g(X)f(X)}{G}$ .  $\tilde{f}_{ideal}$  cannot be realized in practice since the value of  $G$  is not known a priori. Instead, an  $\tilde{f}$  “similar” to  $(1/G)g(X)f(X)$  is used. This corresponds to picking samples with larger likelihood where the integrand is largest.

## B.2 Control Variates

In this approach, a function  $h(X)$  that “correlates well” with  $g(X)$  is used.  $h$  must be so that the integral

$$H = \int_{\Omega} h(X)f(X)dX \quad (17)$$

can be estimated with very low variance, e.g., is known analytically, and  $\Delta(X) =_{def} g(X) - f(X)$  has much smaller variance than  $g(X)$  itself. Eqn. 13 is then written as

$$G \int_{\Omega} \Delta(X)f(X)dX + \int_{\Omega} h(X)f(X)dX \quad (18)$$

The estimator for  $G$  becomes

$$G_{cm} = H + \frac{1}{N} \sum_{i=1}^N \Delta(X_i)$$

where the samples  $X_i$  are drawn from  $f$ . The reason for the variance reduction obtained is apparent in the equation above.  $H$  can be estimated with 0 or very low variance. If  $g$  is very close to  $f$ , then all  $\Delta(X_i)$  values and thus the contribution to the variance from the second term are very small. We use the correlation technique by using the logical effort approximation to devise a function that approximates circuit delay well.

## B.3 Combining the Two Techniques

One can apply the importance sampling technique for variance reduction when estimating the first term in the integral in Eqn. 18, rewritten here as

$$G_{\Delta} = \int_{\Omega} \Delta(X)f(X)dX \quad (19)$$

The estimator for  $G_{\Delta}$  is given by  $\hat{G} = \frac{1}{N} \sum_{i=1}^N \Delta(X_i)$  where the  $X_i$  are drawn from a probability density function  $\hat{f}$  that approximates

$$\hat{f}_{ideal} = \frac{\Delta(X)f(X)}{G_{\Delta}} \quad (20)$$

This combination of techniques results in further reduced variance for the estimator for  $G$ .

## V. TIMING YIELD ESTIMATION

The problem addressed in this section is formalized as follows. A target delay  $T_c$  is given for the circuit. Given a probability density function  $f(X)$  for the process parameters and given that the design parameters are given by  $S$ , we would like to compute the fraction of circuits that satisfy  $d_C(S, X) \leq T_c$ .

### A. Expressing Timing Yield as an Integral

For a given path  $\pi$ , let us define an “indicator variable”  $I_{\pi}(S, X)$  that evaluates to 1 if the delay along  $\pi$  does not meet the timing constraint, i.e.,  $d_{\pi}(S, X) > T_c$ . We define an indicator variable  $I(S, X)$  for the entire circuit.  $I(S, X) = 1$  if the circuit delay exceeds the target, i.e.,  $d_C(S, X) > T_c$ , 0 otherwise. We have

$$I(S, X) = \bigvee_{\pi \in \Pi_{crit}} I_{\pi}(S, X) \quad (21)$$

$$Loss(S) = 1 - Yield(S) = \int I(S, X)f(X)dX \quad (22)$$

Using this equation, we express the yield computation as a definite integral of the form discussed in the previous section. A straightforward application of Monte Carlo simulation to yield estimation would involve drawing samples  $X_1, X_2, \dots, X_n$  from the process parameter space according to the probability density function  $f$  and using an estimator as described in Eqn. 14.

### B. The Logical Effort Approximation and Variance Reduction

The variance reduction methods discussed in Section IV can benefit from the use of a function that approximates the integrand  $g(X)$  well. For estimating timing yield, we make use of the logical effort approximation to obtain a function that approximates  $I(S, X)$  and has the mathematical properties required by the variance reduction methods.

We define the indicator variable  $I^{LE}(S, X)$  as follows. We apply the method of logical effort to compute an approximate delay for each path in  $\Pi_{crit}$ . We then determine whether any of these approximate delays exceed  $T_c$  and assign  $I^{LE}(S, X)$  to 1 if this is the case, to 0 otherwise. In the following, we describe how this computation is carried out at a given point in the process parameter space given by  $X$ . Given a gate  $g$ , the logical effort approximation to its delay is given by the following equation.

$$d_g^{LE}(S, X) = \mathcal{F}(\{d_{ref}(X), LE(\cdot), \pi, S\}) \quad (23)$$

Here  $d_{ref}(X)$  stands for *reference inverter delay*, i.e., the delay of a minimum-size inverter if the process parameters are given by  $X$ .  $LE(g, \pi, S)$  indicates the logical effort for gate  $g$  computed as a function of the path  $\pi$  that  $g$  lies on and the design parameters  $S$ .  $\mathcal{F}$  is a function that computes the delay for  $g$  as a function of the reference inverter and the logical effort for  $g$ . The power of logical effort is apparent in Eqn. 23. The effects on  $g$ 's delay of process parameters and design parameters are separated. As a result, in different Monte Carlo evaluations, only the new  $d_{ref}(X)$  for the new assignment to  $X$  needs to be re-evaluated.

To characterize  $d_{ref}(X)$ , we run an extensive set of Monte Carlo simulations on a single, minimum-size inverter and store the results in a table. Since this step is performed on a single, minimum-size inverter, it is computationally inexpensive. It is also unavoidable if we want to characterize the dependence of delay on the process parameters.  $I_{\pi}^{LE}(S, X)$ , the logical effort approximation to the delay of a path  $\pi$  is obtained simply by adding the  $d_g^{LE}(S, X)$  values for all gates  $g$  on  $\pi$  and by comparing the total with  $T_c$ . The indicator variable  $I^{LE}(S, X)$  is then given by

$$I^{LE}(S, X) = \bigvee_{\pi \in \Pi_{crit}} I_{\pi}^{LE}(S, X) \quad (24)$$

While applying the logical effort approximation to reduce variance, we will find it useful to have a low-variance estimator for the following integral

$$Loss^{LE} =_{def} \int I^{LE}(S, X)f(X)dX \quad (25)$$

The value of this integral is an approximation to circuit timing loss given by the logical effort approximation. We estimate  $Loss^{LE}$  by Monte Carlo simulations as well. However, these Monte Carlo simulations (we will refer to them as “block-level Monte Carlo simulations” from now on) are not detailed, circuit-level simulations but more similar to inexpensive static timing analysis runs as outlined next. For each gate, we compute its approximate delay  $d_g^{LE}(S, X)$  using Eqn. 23 by performing a simple look-up to the reference inverter delay table and by applying  $\mathcal{F}$ . The circuit delay computation, as outlined above, boils down to one conventional static timing analysis run using  $d_g^{LE}(S, X)$  for each gate. The result obtained by this inexpensive computation is denoted by  $d_C^{LE}(S, X)$  which directly determines the value of  $I^{LE}(S, X)$ . Because of the significantly reduced cost of each such delay computation pass, one can use many more samples from the process parameter space for the purpose of estimating  $Loss^{LE}$  and in this way accomplish low variance.

### B.1 Importance Sampling and Logical Effort

To apply importance sampling to the integral in Eqn. 22, we need a function that approximates  $\frac{I(S, X)f(X)}{Loss(S)}$ . The natural first guess is to use

$$f^{IS} = \frac{I^{LE}(S,X)f(X)}{Loss^{LE}(S)} \quad (26)$$

As is required, this expression is a probability density function. However,  $I^{LE}(S,X)$  can be 0 where  $I(S,X)f(X)$  is non-zero. To overcome this difficulty, we define the indicator variable  $I^{MAR}(S,X)$  as follows.  $I^{MAR}(S,X)$  evaluates to 1 if  $d_C^{LE}(S,X) \geq (1-\epsilon)T_c$ .  $I^{MAR}(S,X)$  adds a ‘‘safety margin’’ parameterized by  $\epsilon$  to  $I^{LE}(S,X)$ . By experimentally determining a proper value of  $\epsilon$ ,  $I^{MAR}(S,X)$  can be guaranteed to be non-zero everywhere  $I(S,X)f(X)$  is non-zero. We then draw samples from the following probability density function.

$$f^{IM} = \frac{I^{MAR}(S,X)f(X)}{G^{IM}} \quad (27)$$

where  $G^{IM} = \int I^{MAR}(S,X)f(X)dX$  and can be determined during the same set of Monte Carlo simulations used for determining  $I^{LE}(S,X)$ . Roughly speaking, the effect of using  $f^{IM}$  for importance sampling is to avoid running full-circuit Monte Carlo simulations where  $I(S,X)$  is guaranteed to be 0 and to increase the likelihood of running full-circuit Monte Carlo simulations where the value of  $I(S,X)f(X)$  is high.

## B.2 Control Variates and Logical Effort

Following the approach in Section IV-B.2, we express Eqn. 22 as

$$Loss(S) = \int (I(S,X) - I^{LE}(S,X))f(X)dX + \int I^{LE}(S,X)f(X)dX \quad (28)$$

The second term is simply  $Loss^{LE}$  and can be estimated inexpensively with low-variance as discussed earlier. Since the block-level Monte Carlo evaluations performed for estimating  $I^{LE}(S,X)$  are computationally much cheaper than full-circuit Monte Carlo simulations, this approach has a distinct advantage over straightforward Monte Carlo evaluation of the  $Loss(S)$  integral. The first term is 0 in most of the  $X$  space and a very large number of samples can be used for the second term. In the control variates method, the first term in Eqn. 28 is estimated using conventional Monte Carlo simulation.

## B.3 Combining Control Variates and Importance Sampling

Let us define  $I^\Delta(S,X) = I(S,X) - I^{LE}(S,X)$ . We now describe how we apply a combination of importance sampling and correlation methods to obtain a reduced-variance estimator for the first term in Eqn. 28, rewritten here as

$$G^\Delta = \int I^\Delta(S,X)f(X)dX$$

We need to pick a probability distribution function  $\tilde{f}$  to use instead of  $f$ .  $\tilde{f}(X)$  must be non-zero everywhere the integrand is non-zero and it is desirable that  $\frac{g(X_i)f(X_i)}{\tilde{f}(X_i)}$  be bounded from above.

Notice that  $I^\Delta(S,X)$  is non-zero if, for given  $X$ ,  $d_C^{LE}(S,X) \leq T_c$  while  $d_C(S,X) > T_c$  or vice versa. Since  $d_C^{LE}(S,X)$  is known to be a good first-order approximation, this is only possible if  $d_C^{LE}(S,X)$  is close to  $T_c$ . This inspires the use of a  $\tilde{f}$  that has the non-zero value  $\kappa f(X)$  if the circuit delay computed using logical effort,  $d_C^{LE}(S,X)$ , is in the interval  $[(1-\epsilon)T_c, (1-\delta)T_c]$  for fixed, small  $\epsilon$  and  $\delta$  and  $\tilde{f} = 0$  otherwise. The value of  $\epsilon$ ,  $\delta$  and  $\kappa$  must be chosen so that  $\int \tilde{f}(X)dX = 1$ . The data required for the computation of  $\kappa$  can be obtained from the set of Monte Carlo simulations required for estimating  $I^{LE}(S,X)$ . The samples  $X_i$  from the process parameter space for which logical effort delay for the circuit was in the interval  $[(1-\epsilon)T_c, (1-\delta)T_c]$  are stored and used for the Monte Carlo estimation of  $G^\Delta$ . Note that this choice of  $\tilde{f}$  simply corresponds to an importance sampling approach where more samples  $X_i$  are chosen from the subset of the process parameter space that results in delay estimates close to  $T_c$ .

## C. Discussion

In essence, importance sampling provides a method for eliminating circuit simulations where the process parameters  $X$  make it uninteresting. The key benefit of the control variates approach is also derived from the fact that the Monte Carlo estimation of  $Loss^{LE}$  can be performed much more cheaply and accurately than  $Loss$ . Since there is an abundance of evidence that the logical effort approach approximates delay (and delay variations) well, we expect the control variates approach to further reduce variance.

## D. Proposed Experiments

To evaluate the degree to which each variance reduction technique improves the accuracy of the yield estimate and amount of computation required by each, we propose a set of experiments. A small circuit  $C$  is chosen and  $X$ ,  $f(X)$  and the timing constraint  $T_c$  are determined. The following experiments are performed:

1. *Conventional Monte Carlo runs*: A very large number of samples  $X_1, X_2, \dots, X_N$  are drawn from  $f(X)$  and the timing yield is estimated. Let us denote this estimator by  $Yield^{Conv}$ .  $N$  is chosen considering that the variance of  $Yield^{Conv}$  decreases by  $\sqrt{N}$ .
2. *Generating low-variance estimators*: For each technique described in Section V, let us refer to the following process as ‘‘one pass of the technique’’: A set of sample assignments  $S_i^{VR}$  to process parameters are generated using the appropriate probability density function and the value of the estimator,  $Yield_i^{VR}$ , is computed using these samples. The size of the set  $S_i^{VR}$  is selected so that the number of circuit-level timing simulations performed as a result is much smaller than  $N$ . To be able to evaluate the variance of the estimator obtained by this technique, we need to perform several passes and obtain a number of samples  $Yield_1^{VR}, Yield_2^{VR}, \dots, Yield_k^{VR}$  of the estimator.
3. *Comparison*: Since the actual timing yield is very close to  $Yield^{Conv}$ , the variance of the estimator can be computed assuming that  $Yield_1^{VR}, Yield_2^{VR}, \dots, Yield_k^{VR}$  are the sample values of the estimator and  $Yield^{Conv}$  is the mean. This variance is compared with the variance that would have been obtained if the same computation time was dedicated to conventional Monte Carlo simulation.

## VI. CONCLUSIONS

Performance variability due to statistical process variations and environmental fluctuations is an important and difficult issue. Statistical modeling, analysis and optimization of digital circuit timing is a challenging, and if a proper approach is not used, a possibly intractable problem. In our opinion, statistical timing analysis techniques that concentrate most of their effort on accurate manipulation of statistical models are not the proper direction to take, because the statistical delay models they use are necessarily approximate and therefore not accurate enough for timing yield analysis and optimization under large parameter variations. For statistical, robust design and optimization purposes, we believe that methodologies and optimization techniques that can produce intuitive and simple design guidelines will prevail. For timing yield estimation accurate enough for sign-off purposes, we conjecture that the only viable techniques are going to be appropriately accelerated Monte Carlo methods, of which we propose several in this paper.

## REFERENCES

- [1] I. Sutherland, B. Sproull, and D. Harris. *Logical Effort: Designing Fast CMOS Circuits*. Morgan Kaufmann, 1999.
- [2] D. Patil, Y. Yun, S.-J. Kim, S. Boyd, and M. Horowitz. A new method for robust design of digital circuits. In *International Symposium on Quality Electronic Design (ISQED)*, 2005.
- [3] Malvin H. Kalos and Paula A. Whitlock. *Monte Carlo Methods, Volume 1, Basics*. Wiley, 1986.
- [4] A. D. Sokal. Monte carlo methods in statistical mechanics: Foundations and new algorithms. In P. Cartier, C. DeWitt-Morette and A. Folacci, editors, *Functional Integration: Basics and Applications (1996 Cargèse summer school)*. Plenum, 1997.
- [5] P. Glasserman, P. Heidelberger, and P. Shahabuddin. Importance sampling and stratification for value-at-risk. In *Proc. 6th Intl. Conference on Computational Finance*, pages 7–24. MIT Press, May 28–31 1999.
- [6] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.
- [7] M. Pelgrom, A.C.J. Duinmaier, and A.P.G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5), October 1989.
- [8] Y. Cao, H. Qin, R. Wang, P. Friedberg, A. Vladimirescu, and J. Rabaey. Yield optimization with energy-delay constraints in low-power digital circuits. In *IEEE Conference on Electron Devices and Solid-State Circuits*, December 2003.
- [9] L. Scheffer. The count of monte carlo. In *ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems (TAU)*, February 2004.