

Stochastic Logical Effort: Designing for Timing Yield on the Back of an Envelope

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—As we move into the nano era in integrated circuit fabrication technologies, the performance variability due to statistical process variations and environmental fluctuations is becoming more and more significant. Considerable effort has been expended in the EDA community during the past several years in trying to cope with the so-called *statistical timing* problem. However, most of this effort has been aimed at generalizing the static timing analyzers to the statistical case in an EDA tool centric manner. In this paper, we take a pragmatic, design centric approach in pursuit of a simple yet powerful stochastic gate delay model that can be used to develop a very efficient timing yield estimation methodology, and that can enable tractable timing yield optimization and eventually lead to simple yet meaningful and useful design guidelines. In doing so, we first develop a generalization of the logical effort delay model for the stochastic case. 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. We then demonstrate why and how the simple stochastic gate delay model that features this *separation of concerns* can be used as a very powerful tool in timing yield estimation and yield optimization. We develop an extremely efficient and simple methodology for optimal gate sizing in order to maximize the timing yield of a path. Using this methodology, we analyze several cases of practical importance and arrive at meaningful and practical conclusions on optimal gate sizing. When only inter-die variations are considered, we show that the sizing that minimizes the nominal path delay also maximizes the path timing yield. However, we also show that this does not hold in general, especially when local, intra-die statistical variations are significant. The simple stochastic gate delay model proposed in this paper can be effectively used to guide the generation and selection of sample points in the parameter/probability space in a transistor-level simulation based Monte Carlo method for timing yield estimation. We discuss in a companion paper how transistor-level Monte Carlo analysis can be accelerated using novel importance sampling and other variance reduction techniques.

Keywords—logical effort, statistical variations, timing yield estimation and optimization, inter and intra-die variations, statistical timing analysis.

I. INTRODUCTION

The method of logical effort by Sutherland et. al. [1], as the authors describe it, “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 timing yield estimation.

In this paper, we are advocating 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. Such a tool may be able to obtain “more optimal” results because it can use a more accurate stochastic gate delay model and a more accurate methodology for estimating timing yield. However, we believe that the greatest value of our approach, like the original logical effort formalism, lies in the insight it provides. Moreover, 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”, an observation originally made by Sutherland et.al. in [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 when compared with the deterministic gate sizing problem, which is already difficult. Thus, when applied to real designs, fully numerical optimization strategies with statistical timing analysis in the loop may face many difficulties and may fall into pitfalls. 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. Here, we are advocating a simple methodology that can produce simple and practical guidelines and, most importantly, insight, as opposed to a methodology which proposes accurate and sophisticated manipulation of inaccurate statistical models with nonlinear delay functions and non-Gaussian parameters that come from nowhere.

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 again cite an observation originally made by Sutherland et.al. in [1]: “Synthesis tools make some effort to explore topologies, but still can *not* match experienced designers on *critical paths*.” We believe that a simple, effective and working path-based timing yield estimation and optimization methodology will be practically much more relevant and useful compared with a complicated, inefficient and possibly non-intuitive block-based approach that relies on unwieldy non-linear, non-Gaussian models that are postulated without much justification.

Fast timing yield estimation based on the statistical logical effort formalism will become the key enabler in an accurate computation of timing yield via a *transistor-level simulation based* Monte Carlo method [3] that is made extremely efficient through the use of novel importance sampling and other variance reduction techniques, as explained in a companion paper [4].

In Section II below, we describe the stochastic gate delay model. Then, we use this model in Section III 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. In Section IV, we provide an outline for how one can use the simple stochastic gate delay model proposed in the paper to guide the generation and selection of sample points in the parameter/probability space in a transistor-level simulation based Monte Carlo method for timing yield estimation.

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, τ 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 τ .

A. Capturing statistical variations with a stochastic delay unit

In the logical effort formalism [1], all delays are expressed in terms of τ in order to *isolate* the effects of the particular integrated circuit

Accepted to ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems (TAU 2006).

The title of the paper was inspired by David Harris’s presentation title *Logical Effort: Designing for Speed on the Back of an Envelope*.

This work was sponsored by the Turkish Academy of Sciences GEBIP program and two TUBITAK Career Awards.

fabrication process. Thus, τ 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 τ , 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 τ 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), τ is characterized/extracted as a *single number* expressed in picoseconds. In our case, we use τ to *capture* and *isolate* the effects of process, circuit and environmental parameters that exhibit statistical variations. Thus, τ is not a single number, but it is a *probability distribution*. We propose three alternative techniques for the stochastic characterization of τ :

- One can derive a simplified analytical expression that relates τ to the circuit, process and environmental parameters that exhibit statistical variations, as mentioned 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 τ . 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 τ by simply evaluating the analytical expression, followed by a compilation of a histogram. Even the simplest analytical expression for τ will have nonlinear dependence on the statistical parameters. Thus, even if the statistical parameters are jointly Gaussian, the probability distribution of τ will not be Gaussian.
- Instead of using an analytical expression as above, one can use a suitable test circuit and a circuit simulator to relate τ to the statistical parameters. One can then again use a Monte Carlo technique to compute the probability distribution of τ . This characterization will be computationally more expensive than the one above, because the test circuit will need to be simulated many times in a circuit simulator. 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 τ 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 τ 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.
- Alternatively, one can use fabricated test structures and physically measure τ and characterize its probability distribution.

In the standard logical effort formalism, τ 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 τ 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 τ_r , and re-express (1) as follows

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

τ above captures the effects of perfectly correlated inter-die statistical variations for the gates that reside on the same die. τ_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, τ_{r_1} and τ_{r_2} may be partially correlated. If these gates are far away from each other, then τ_{r_1} and τ_{r_2} may be considered uncorrelated. The partial correlation among τ_r 's may be best represented by expressing them in terms of other (abstract) random quantities which are independent. This can be accomplished through some sort of (nonlinear) principal or independent component analysis [5]. This is trivial in the case when τ_r 's are jointly Gaussian. Even though it is not as straightforward, there exist techniques that work in the case when τ_r 's are

not jointly Gaussian. We would like to also note that the nominal, mean delay of the reference inverter, as well as the effect of inter-die variations, are captured by τ in (2), and τ_r is an additional, small correction term accounting for the effect of intra-die variations. Thus, τ_r is most likely, but not necessarily, zero mean. Moreover, a Gaussian model for τ_r may not be too far off. Even when τ and hence the total delay $\tau + \tau_r$ is far from Gaussian, it may be good enough to model τ_r as a Gaussian random quantity. Since τ 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 τ_r to capture the effect of local intra-die variations. As first observed and proposed by Pelgrom in [6], 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 τ_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 (2) that models gate delay. 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 τ . 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 2.0μ process to a 0.35μ 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 τ for the same range of processes and power supply voltages mentioned above. Their results show that τ 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 τ 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 2.0μ process all the way to a 0.35μ 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 instance, the reference inverter delay τ . We expect at most a $\pm 35\%$ variability in τ 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 τ 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. Justifications for and implications of the statistical invariability of logical effort

Rabaey et. al in [7] report results that are in support of our claim of the statistical *invariability* of logical effort. They state that ‘‘Electri-

cal 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 $\pi = (\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_R)$ with R gates with \mathbf{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, τ that captures all statistical variation, and a unitless, deterministic, fixed number D^π that quantifies the *total complexity* of the logic gates on the path. Rabaey et. al in [7] 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 τ is $p_\tau(\eta)$ defined for $0 \leq \eta < +\infty$, then the PDF of D_{abs}^π 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^π is simply a deterministic constant. Hence, we can compute yield Y^π 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 $\mathbb{E}[D_{abs}^\pi]$ can be computed using

$$\begin{aligned} \mathbb{E}[D_{abs}^\pi] &= \int_0^{+\infty} \frac{\eta}{D^\pi} p_\tau\left(\frac{\eta}{D^\pi}\right) d\eta \\ &= D^\pi \int_0^{+\infty} u p_\tau(u) du \\ &= D^\pi \mathbb{E}[\tau] \end{aligned} \quad (8)$$

If we substitute the above in (7) and make a change of integration variable using $u = \frac{\eta}{D^\pi}$, we obtain

$$Y^\pi = \int_0^{1.1 \mathbb{E}[\tau]} p_\tau(u) du \quad (9)$$

We observe that the yield expression above is *independent* of D^π and depends only on the PDF and the mean of τ . Thus, with only inter-die variations considered, the yield of a path defined as in (5) is independent of D^π 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^π and a random variable τ .

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

$$\begin{aligned} D_{abs}^\pi &= \sum_{r=1}^R (\tau + \tau_r) d_r \\ &= \tau \sum_{r=1}^R (p_r + g_r h_r) + \sum_{r=1}^R \tau_r (p_r + g_r h_r) \\ &= \tau D^\pi + \sum_{r=1}^R \tau_r (p_r + g_r h_r) \end{aligned} \quad (10)$$

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, one can state the following theorem:

Theorem II.1: 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 α 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}^π expressed as a linear combination of the random variables τ and τ_r 's as in (10). However, based on the theorem above we can conclude the following:

Corollary II.1: 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

$$d_{abs} = (\tau + \tau_r) d = (\tau + \tau_r) (p + gh) \quad (11)$$

where τ is a random variable that serves as a *stochastic delay unit*. τ 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. τ captures the effect of inter-die process variations. The same τ is shared by all of the gates on a die. Hence, τ captures the perfectly correlated portion of the statistical variations, including the nominal delay. τ_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 (or uncorrelated) 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 τ and τ_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.

While statistical variations are captured by gate independent but process dependent τ and τ_r , the gate type, topology, sizing and loading information is captured by $d = p + gh$ which does not exhibit statistical variations. With the stochastic gate delay model in (11), 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 (11) 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. We proceed as follows:

- *Section III-A:* We review the mechanics and the conclusions of the deterministic minimization of path delay using the standard logical effort formalism. This serves as a reference for the stochastic cases we consider next.
- *Section III-B:* We consider the maximization of the timing yield of a path when only *inter-die* statistical variations are present.

• *Section III-C*: We consider the maximization of the timing yield of a path with both *inter-die* and *intra-die* statistical variations.

A. Deterministic path delay minimization

We consider a path $\pi = \langle \mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_R \rangle$ with R gates with \mathbf{g}_r as the r th gate on the path. The total delay of the path can be expressed as follows

$$D_{abs}^\pi = \tau \sum_{r=1}^R d_r = \tau \sum_{r=1}^R (p_r + g_r h_r) \quad (12)$$

We do not consider statistical variations in this case. Hence, τ is a fixed constant. h_r above is the electrical effort of the r th gate \mathbf{g}_r on the path. It is the ratio of the load capacitance of \mathbf{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 \mathbf{g}_r is essentially the input capacitance of gate \mathbf{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

$$H = h_1 h_2 \cdots h_R = \prod_{r=1}^R h_r \quad (13)$$

Path delay minimization without a constraint on the path electrical effort is not meaningful. For instance, if only the capacitance at the termination of the path is specified, then one can increase the sizes of the gates in the path to the point where the delay due to the output load capacitance becomes negligible. However, if the sizes of the gates in the path are made very large, then the delay of some other logic circuit driving the first gate on this path will increase considerably. As such, it is not meaningful to specify only the load capacitance that the path needs to drive. A specification on the ratio of the termination load capacitance to the input capacitance of the first gate is more proper and required.

Given the constraint on the path electrical effort above, the information captured by the electrical efforts is essentially equivalent to the sizing information for the gates. Given the electrical efforts for the gates and the specification on the path electrical effort, one can easily compute the sizes of the transistors in the gates of the path [1]. In minimizing path delay, we choose the electrical efforts of the gates, i.e., h_1, h_2, \dots, h_R , as the optimization variables. The optimization problem can then be formulated as follows:

$$\begin{aligned} \text{minimize} \quad & D_{abs}^\pi = \tau \sum_{r=1}^R (p_r + g_r h_r) \\ \text{subject to} \quad & \prod_{r=1}^R h_r = H \end{aligned} \quad (14)$$

We proceed with the solution of this problem using the method of Lagrange multipliers. The first-order necessary conditions for optimality can be written as

$$\frac{\partial}{\partial h_k} \left[\tau \sum_{r=1}^R (p_r + g_r h_r) + \lambda \prod_{r=1}^R h_r \right] = \tau g_k + \lambda \prod_{\substack{r=1 \\ r \neq k}}^R h_r = 0 \quad (15)$$

$k = 1, 2, \dots, R$

We multiply both sides of the equations above with h_k

$$\tau g_k h_k + \lambda \prod_{r=1}^R h_r = 0 \quad k = 1, 2, \dots, R \quad (16)$$

and use the constraint in (13) to obtain

$$\tau g_k h_k + \lambda H = 0 \quad k = 1, 2, \dots, R \quad (17)$$

Using the above, we can write

$$g_1 h_1 = g_2 h_2 = \cdots = g_R h_R = -\frac{\lambda H}{\tau} \quad (18)$$

which gives us the main conclusion on the solution of the deterministic path delay minimization problem:

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.

B. Timing yield maximization with inter-die variations

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 (11) with $\tau_r = 0$ in order to null out the intra-die variations. The total delay of the path can be expressed as follows

$$D_{abs}^\pi = \tau \sum_{r=1}^R d_r = \tau \sum_{r=1}^R (p_r + g_r h_r) \quad (19)$$

where τ is a random variable, shared by all of the gates. We assume that the PDF of τ is given by $p_\tau(\eta)$ defined for $0 \leq \eta < +\infty$, not necessarily a Gaussian PDF. Please note that, except for τ , all of the other quantities in (19) are deterministic. We use the following definition for the *path timing yield*

$$Y^\pi = \mathbb{P}(D_{abs}^\pi \leq T_c) \quad (20)$$

where T_c is specified as the cut-off delay. In the presence of statistical variations, we formulate a *stochastic* optimization problem to maximize the path timing yield (as opposed to minimizing the nominal delay):

$$\begin{aligned} \text{maximize} \quad & Y^\pi = \mathbb{P}(D_{abs}^\pi \leq T_c) = \mathbb{P}\left(\tau \sum_{r=1}^R (p_r + g_r h_r) \leq T_c\right) \\ \text{subject to} \quad & \prod_{r=1}^R h_r = H \end{aligned} \quad (21)$$

The path timing yield above can be expressed in terms of the PDF $p_\tau(\eta)$ of τ as follows

$$Y^\pi = \int_0^{T_c} \frac{1}{\sum_{r=1}^R (p_r + g_r h_r)} p_\tau\left(\frac{\eta}{\sum_{r=1}^R (p_r + g_r h_r)}\right) d\eta \quad (22)$$

because the PDF of $\alpha\tau$ is given by $\frac{1}{\alpha} p_\tau\left(\frac{\eta}{\alpha}\right)$ when α is a deterministic constant. With a change of the integration variable using

$$u = \frac{\eta}{\sum_{r=1}^R (p_r + g_r h_r)} \quad (23)$$

the above integral turns into

$$Y^\pi = \int_0^{\frac{T_c}{\sum_{r=1}^R (p_r + g_r h_r)}} p_\tau(u) du \quad (24)$$

With the above expression for yield, we can rewrite the optimization formulation in (21) as follows

$$\begin{aligned} \text{maximize} \quad & Y^\pi = \int_0^{\frac{T_c}{\sum_{r=1}^R (p_r + g_r h_r)}} p_\tau(u) du \\ \text{subject to} \quad & \prod_{r=1}^R h_r = H \end{aligned} \quad (25)$$

We proceed with the solution of the above problem again using the method of Lagrange multipliers. The first-order necessary conditions for optimality can be written as

$$\frac{\partial}{\partial h_k} \left[\int_0^{\frac{T_c}{\sum_{r=1}^R (p_r + g_r h_r)}} p_\tau(u) du + \lambda \prod_{r=1}^R h_r \right] = 0 \quad (26)$$

for $k = 1, 2, \dots, R$

and then

$$-\frac{g_k T_c}{\left[\sum_{r=1}^R (p_r + g_r h_r)\right]^2} p_\tau\left(\frac{T_c}{\sum_{r=1}^R (p_r + g_r h_r)}\right) + \lambda \prod_{\substack{r=1 \\ r \neq k}}^R h_r = 0 \quad (27)$$

for $k = 1, 2, \dots, R$

$$\begin{aligned}
Y^\pi &= \int_{-\infty}^{T_c} p_{D_{abs}^\pi}(\eta) d\eta = \int_{-\infty}^{T_c} \left[\int_{-\infty}^{\infty} p_{D_{absinter}^\pi}(\xi) p_{D_{absintra}^\pi}(\eta - \xi) d\xi \right] d\eta \\
&= \int_{-\infty}^{T_c} \left[\int_{-\infty}^{\infty} \frac{1}{\sum_{r=1}^R (p_r + g_r h_r)} p_\tau \left(\frac{\xi}{\sum_{r=1}^R (p_r + g_r h_r)} \right) \frac{1}{\sqrt{\sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2}} p_{N(0,1)} \left(\frac{\eta - \xi}{\sqrt{\sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2}} \right) d\xi \right] d\eta \quad (LE1)
\end{aligned}$$

We multiply both sides of the equations above with h_k , use the constraint in (13) and re-arrange the terms to obtain

$$g_1 h_1 = g_2 h_2 = \dots = g_R h_R = \frac{\lambda H}{T_c} \frac{[\sum_{r=1}^R (p_r + g_r h_r)]^2}{p_\tau \left(\frac{T_c}{\sum_{r=1}^R (p_r + g_r h_r)} \right)} \quad (28)$$

The examination of the equation above takes us to the final conclusion:

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.

This conclusion is exactly the same conclusion we arrived at in Section III-A, when we considered the deterministic problem, i.e., the minimization of nominal path delay.

C. Timing yield maximization with inter-die and intra-die variations

We now consider both inter-die and intra-die statistical variations, and use the full stochastic gate delay model in (11). In this case, from (10), the total path delay can be expressed as follows

$$\begin{aligned}
D_{abs}^\pi &= \sum_{r=1}^R (\tau + \tau_r) d_r \\
&= \tau \sum_{r=1}^R (p_r + g_r h_r) + \sum_{r=1}^R \tau_r (p_r + g_r h_r) \\
&= \tau D^\pi + \sum_{r=1}^R \tau_r (p_r + g_r h_r) \\
&= D_{absinter}^\pi + D_{absintra}^\pi \quad (29)
\end{aligned}$$

where

$$D_{absinter}^\pi = \tau \sum_{r=1}^R (p_r + g_r h_r) \quad , \quad D_{absintra}^\pi = \sum_{r=1}^R \tau_r (p_r + g_r h_r) \quad (30)$$

τ above is a random variable, shared by all of the gates. We assume that the PDF of τ (for the inter-die statistical variations and the nominal behavior) is given by $p_\tau(\eta)$ defined for $0 \leq \eta < +\infty$, not necessarily a Gaussian PDF. τ_r 's (represent corrections for the intra-die statistical variations) above are uncorrelated with each other and also with τ . We also assume that τ_r 's are zero mean Gaussian random variables with variance σ_r^2 . Thus, $D_{absinter}^\pi$ is a random variable with PDF

$$p_{D_{absinter}^\pi}(\eta) = \frac{1}{\sum_{r=1}^R (p_r + g_r h_r)} p_\tau \left(\frac{\eta}{\sum_{r=1}^R (p_r + g_r h_r)} \right) \quad (31)$$

and $D_{absintra}^\pi$ is also a Gaussian¹ random variable with zero mean and variance equal to

$$\sigma_{D_{absintra}^\pi}^2 = \sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2 \quad (32)$$

and hence with a PDF

$$\begin{aligned}
p_{D_{absintra}^\pi}(\eta) &= \\
&= \frac{1}{\sqrt{\sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2}} p_{N(0,1)} \left(\frac{\eta}{\sqrt{\sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2}} \right) \quad (33)
\end{aligned}$$

where $p_{N(0,1)}(\eta)$ is the PDF of a Gaussian random variable with zero mean and unity variance. The PDF of the total delay in (29) is then

¹Even if τ_r 's are not Gaussian, $D_{absintra}^\pi$ may still be approximated with a Gaussian random variable based on the Central Limit Theorem, because it is the linear combination of several τ_r 's.

given as the convolution (denoted by $*$) of $p_{D_{absinter}^\pi}(\eta)$ in (31) and $p_{D_{absintra}^\pi}(\eta)$ in (33) above

$$p_{D_{abs}^\pi}(\eta) = p_{D_{absinter}^\pi}(\eta) * p_{D_{absintra}^\pi}(\eta) \quad (34)$$

because $D_{absinter}^\pi$ and $D_{absintra}^\pi$ are independent. The path timing yield, as defined in (20), can be expressed as in (LE1) at the top of the page. Unfortunately, we can not put the expression for yield in (LE1) in a simpler form, because we have assumed a general PDF $p_\tau(\eta)$ for τ . It is then very unlikely that we will be able to find an analytical or semi-analytical solution to the yield maximization problem using the method of Lagrange multipliers. We were able to solve the yield maximization problem *fully analytically* in Section III-B by assuming a general PDF for τ . We now proceed with a simplification and approximate the PDF of the reference inverter delay τ (due to inter-die variations) with a Gaussian PDF, and assume that the PDF of the reference inverter delay τ_r due to intra-die variations can also be approximated with a Gaussian PDF. This simplification will 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. The more general yield expression in (LE1) can be used for numerical yield maximization and for fast yield estimation.

To proceed further, we assume that the PDF of $p_\tau(\eta)$ of τ is Gaussian with mean μ_o and variance σ_o^2 . In this case, $D_{absinter}^\pi$ becomes a Gaussian random variable with mean and variance

$$\begin{aligned}
\mu_{D_{absinter}^\pi} &= \mu_o \sum_{r=1}^R (p_r + g_r h_r) \\
\sigma_{D_{absinter}^\pi}^2 &= \sigma_o^2 \left[\sum_{r=1}^R (p_r + g_r h_r) \right]^2 \quad (35)
\end{aligned}$$

The total path delay D_{abs}^π is then also Gaussian with mean and variance

$$\begin{aligned}
\mu_{D_{abs}^\pi} &= \mu_o \sum_{r=1}^R (p_r + g_r h_r) \\
\sigma_{D_{abs}^\pi}^2 &= \sigma_o^2 \left[\sum_{r=1}^R (p_r + g_r h_r) \right]^2 + \sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2 \quad (36)
\end{aligned}$$

Then, the yield expression in (LE1) can be simplified as follows

$$\begin{aligned}
Y^\pi &= \int_{-\infty}^{UL} p_{N(0,1)}(\eta) d\eta \\
UL &= \frac{T_c - \mu_o \sum_{r=1}^R (p_r + g_r h_r)}{\sqrt{\sigma_o^2 \left[\sum_{r=1}^R (p_r + g_r h_r) \right]^2 + \sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2}} \quad (37)
\end{aligned}$$

The path timing yield optimization problem can be formulated as follows

$$\begin{aligned}
&\text{maximize } Y^\pi = \int_{-\infty}^{UL} p_{N(0,1)}(\eta) d\eta \\
&\text{subject to } \prod_{r=1}^R h_r = H \quad (38)
\end{aligned}$$

where UL given in (37) is a function of all of the optimization variables h_1, h_2, \dots, h_R . We proceed with the solution of the above problem again using the method of Lagrange multipliers. The first-order necessary conditions for optimality can be written as

$$\frac{\partial}{\partial h_k} \left[\int_{-\infty}^{UL} p_{N(0,1)}(\eta) d\eta + \lambda \prod_{r=1}^R h_r \right] = 0 \quad \text{for } k = 1, 2, \dots, R \quad (39)$$

and then

$$h_k \frac{\partial [\text{UL}]}{\partial h_k} p_{N(0,1)}(\text{UL}) + \lambda \prod_{r=1}^R h_r = 0 \quad \text{for } k = 1, 2, \dots, R \quad (40)$$

After we define

$$\begin{aligned} \text{NUL} &= T_c - \mu_0 \sum_{r=1}^R (p_r + g_r h_r) \\ \text{DUL} &= \sqrt{\sigma_0^2 \left[\sum_{r=1}^R (p_r + g_r h_r) \right]^2 + \sum_{r=1}^R \sigma_r^2 (p_r + g_r h_r)^2} \end{aligned} \quad (41)$$

we proceed further and obtain (LE2) at the top of the next page. By noting that UL, NUL and DUL in (LE2) are k -independent, and after an examination of the equations in (LE2) along with the constraint equation in (13), we arrive at the following conclusion on the optimal choice for the electrical efforts h_1, h_2, \dots, h_R : If

- $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_R^2 = \sigma^2$, i.e., the variance of the part of the reference inverter delay due to intra-die variations is the same for all of the gates on the path,
- $p_1 = p_2 = \dots = p_R$, i.e., the normalized parasitic delays of the gates in the path are equal to each other, which essentially means that we have a chain of the same kind of gate in the path,

then 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.

In the general case when both inter-die and intra-die variations are present, and when at least one of the two conditions above is violated, the optimal values for the electrical efforts h_1, h_2, \dots, h_R can not be chosen using the simple rule above, but they can be computed by solving the nonlinear system of equations in (LE2) along with the constraint equation in (13). We have written a simple MATLAB script for the numerical solution of these equations using *fsolve*. We provide an initial guess for *fsolve* by computing the electrical efforts h_1, h_2, \dots, h_R so that the stage effort or the effort delay (product of g_r and h_r) is the same for each logic stage/gate in the path. In all of the examples we have tried, *fsolve* is able to find the optimal solution that satisfies (LE2) and (13) in just a few iterations in a fraction of a second. In all of these examples, the optimal solution was very close to the initial guess and the difference between the optimal yield value and the yield value at the initial guess was practically insignificant. This is due to the fact that 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 τ and τ_r are gate or location independent. This is justified for τ since it models inter-die variations. However, as we discussed in Section II, the variance of τ_r can typically be modeled to be inversely proportional to the total area the gate occupies. We can easily incorporate this effect in our model by expressing the variance of τ_r as follows:

$$\sigma_r^2 = \frac{\sigma^2}{NI_r \prod_{q=1}^{r-1} h_q} \quad (42)$$

where NI_r is the number of inputs for gate g_r , σ^2 is the intra-die variation delay variance for a minimum sized inverter, and $\prod_{q=1}^{r-1} h_q$ is the normalized (with the per input capacitance of the first gate in the path) per input capacitance for gate g_r . The effect of this variance model can be easily incorporated in the first-order necessary conditions for optimality in (LE2). We have modified our simple MATLAB script mentioned above to include this variance model. The optimal values for h_1, h_2, \dots, h_R can still be computed in a fraction of a second, in a few iterations. With this new 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 produce the same behavior.

IV. FAST AND EFFICIENT TIMING YIELD ESTIMATION AND IMPORTANT SAMPLE POINT GENERATION

We provide an outline here for how one can make use of the stochastic logical effort model described in Section II to guide the generation and selection of sample points in the parameter/probability space in a transistor-level simulation based Monte Carlo method for timing yield estimation. One can also use it for very fast and efficient estimation of the timing yield.

We first obtain stochastic characterizations for τ (for inter-die variations) and τ_r (for intra-die variations). As explained in detail in Section II, these stochastic characterizations may be in the form of a PDF, a table look-up or response surface model built from extensive transistor-level Monte Carlo simulations or from measurements of a simple test circuit where an inverter is loaded with several others. This characterization needs to be done only once for every fabrication process. The table look-up or response surface model mentioned above most likely expresses τ and τ_r in terms of basic statistical process parameters which are independent from each other (or made independent through some sort of principal or independent component analysis [5]). The stochastic characterization for τ_r will include information on how its stochastic characteristics such as its PDF or variance changes with location on the chip or gate size. This stochastic characterization will possibly also include information on the correlation between τ_r 's at different locations on the chip. Please recall that the same τ is shared by all of the gates on a particular die since it represents inter-die variations. We then combine the stochastic characterizations for τ and τ_r with the topology, type, size and loading information for the gates on a collection of paths in a digital circuit.

Armed with the above, we can answer the following question in a very efficient manner: Given a point in the probability space, i.e. an assignment to the basic statistical process parameters, does the maximum of the delays of the paths exceed a specification on total delay? To answer this question, we first use the stochastic characterizations of τ and τ_r 's to compute their values for the given point in the probability space. We then compute the delays of all of the paths under consideration using these value assignments for τ and τ_r and the logical effort model. We finally compute the maximum of the path delays and compare it with the specification.

Being able to answer the question above in a very efficient manner enables us to generate "important" sample points in the probability space in a transistor-level Monte Carlo analysis for timing yield computation accelerated through the use of novel importance sampling techniques. The stochastic logical effort model is used here to test the "importance" of the sample point before an expensive transistor-level simulation is performed, as explained in more detail in [4].

One can accumulate the answers given to the above question at all of the tested sample points generated in the probability space and use them to compute a quick estimate for the timing yield of the collection of paths under consideration. Once an adequate number of "important" sample points are generated, a transistor-level simulation is performed at every one of them to obtain a much more accurate estimate of the timing yield.

V. CONCLUSIONS

There is no doubt that the performance variability due to statistical process variations and environmental fluctuations is an important and very difficult to cope with issue that is posing a challenge to the IC design and EDA community. Statistical modeling, analysis and optimization of digital circuit timing is a challenging, and if not approached wisely, a possibly intractable problem. It is practically almost impossible to gather detailed and accurate variability data and develop precise statistical models for parameter and performance variability. We think that trying to develop statistical timing analysis techniques that can very accurately manipulate grossly inaccurate statistical data and imprecise stochastic models is a futile effort. We believe that only simple methodologies and optimization techniques that do not rely on detailed and precise models of statistical variability and that can produce simple design guidelines will prevail in the statistical, robust design arena. We also believe that the only meaningful and practical techniques for accurate timing yield estimation are going to be the appropriately accelerated Monte Carlo methods.

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] 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.

$$\frac{-\mu_o g_k h_k \text{DUL} - g_k h_k \left\{ \sigma_o^2 \left[\sum_{r=1}^R (p_r + g_r h_r) \right] + \sigma_k^2 (p_k + g_k h_k) \right\} \text{UL}}{\text{DUL}^2} p_{N(0,1)}(\text{UL}) + \lambda H = 0 \quad \text{for } k = 1, 2, \dots, R \quad (\text{LE2})$$

- [4] S. Tasiran and A. Demir. Smart monte carlo for yield estimation. In *ACM/IEEE International Workshop on Timing Issues in the Specification and Synthesis of Digital Systems (TAU)*, February 2006.
- [5] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.
- [6] M. Pelgrom, A.C.J. Duinmaijer, and A.P.G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5), October 1989.
- [7] 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.